DA-DPO：用于减少多模态大语言模型幻觉的高效、难度感知偏好优化

Longtian Qiu *qiult@shanghaitech.edu.cn*
上海科技大学，中国上海

Shan Ning *ningshan2022@shanghaitech.edu.cn*
上海科技大学，上海，中国；临港实验室，上海，中国

张楚瑜 *zhangchy2@shanghaitech.edu.cn*
上海科技大学，上海，中国

孙嘉轩 *sunjx2022@shanghaitech.edu.cn*
中国上海，上海科技大学

Xuming He *hexm@shanghaitech.edu.cn*
上海科技大学，中国上海，上海智能视觉与成像工程技术研究中心

已在 OpenReview 上评审： *https://openreview.net/forum?id=M52CgPcgGx*

arXiv:2601.00623v1 [cs.AI] 2 Jan 2026

摘要

直接偏好优化（DPO）在缓解多模态大语言模型（MLLMs）的幻觉方面展现出巨大潜力。然而，现有的多模态直接偏好优化方法常因偏好数据中的难度不均衡而出现过拟合。我们的分析表明，多模态大语言模型往往过度关注容易区分的偏好对，这阻碍了对幻觉进行细粒度抑制，并降低了整体性能。为解决这一问题，我们提出难度感知直接偏好优化（DA-DPO），这是一种旨在平衡学习过程的高效框架。DA-DPO包含两个主要组件：(1)难度估计利用具有互补生成目标和对比目标的预训练视觉语言模型，通过分布感知投票策略整合其输出，从而无需额外训练即可生成鲁棒的难度分数；(2) 难度感知训练依据估计的难度对偏好对重新加权，降低简单样本的权重，同时强调较难的样本，以缓解过拟合。该框架通过优先处理具有挑战性的样本，实现更有效的偏好优化，且无需新增数据或额外的微调阶段。大量实验表明，DA-DPO能够持续提升多模态偏好优化，在标准基准测试中增强对幻觉的鲁棒性并改善泛化能力，同时保持较高的计算效率。项目页面见 https://artanic30.github.io/project_pages/DA-DPO。

1 引言

近年来，多模态大语言模型（MLLMs）取得了长足进展（Liu 等，2023a；OpenAI，2023年；Li 等，2024），显著提升了视觉语言任务的表现，例如图像描述（Lin 等人，2014）和视觉问答（Agrawal 等人，2015；Mathew 等人，2021年；Marino 等人，2019年）。通过将强大的大语言模型与最先进的视觉模型相结合，多模态大语言模型增强了多模态理解与推理能力。然而，多模态大语言模型面临的一个长期挑战是，它们倾向于生成

无法可靠地以视觉输入为依据的回答，这往往会导致“幻觉”，即描述中包含不存在或不准确的视觉细节。这一局限性影响了多模态大语言模型的可靠性，成为其应用于要求高度事实准确性的场景时的一大障碍。

近期的研究转向将直接偏好优化（DPO）（Rafailov 等，2024）作为缓解多模态大语言模型幻觉的一种有前景的方法。DPO 鼓励模型使其输出与偏好数据保持一致，这些数据倾向于忠实的回答并减少幻觉。至关重要的是，DPO 的效果取决于成对偏好数据的质量。为解决这一问题，早期方法（Sun 等，2023b；Yu 等，2024a）依赖人工标注，但此类数据的收集既耗费人力，也难以扩展。近期，若干研究（Pi 等，2024；Li 等，2023d；Zhou 等，2024c；Yang 等，2025）提出了构建多模态偏好数据的自动化策略。这些方法利用训练好的模型大规模生成成对偏好数据，显著提升了数据对多样化场景的覆盖率，从而增强模型减少幻觉的能力。

尽管原始 DPO 方法在减少幻觉方面效果显著，但在现有成对偏好数据上训练，往往会导致通用多模态能力明显退化，如图1a所示。我们将这一局限归因于训练数据中简单样本与困难样本之间的不平衡，如图1b所示。简单样本通常包含容易区分的忠实回答与幻觉回答，而困难样本则需要更细致的推理才能加以区分。这种不平衡会造成训练偏差，使模型过拟合简单案例，却无法从更具挑战性的样本中学习。我们在第3节中对此现象进行了详细的实证分析，结果表明，模型虽然能迅速适应简单样本，却难以泛化到困难样本，最终限制了基于偏好的对齐效果。

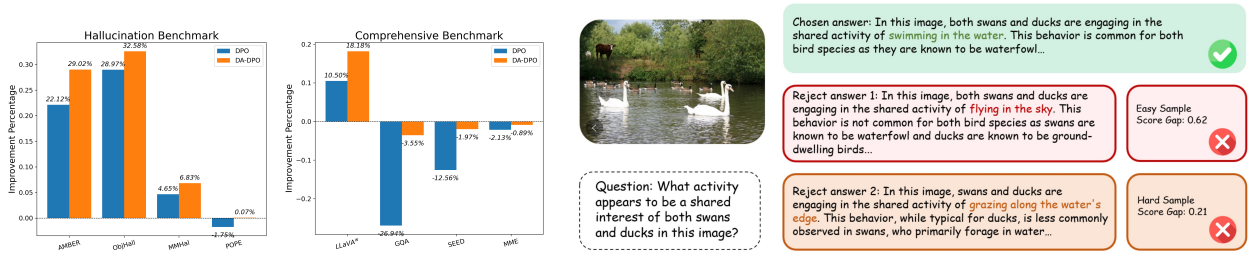
为了解决这一局限，我们提出了一种难度感知训练框架，在偏好优化过程中动态平衡简单样本与困难样本的贡献。实施这一策略的一项关键挑战是缺乏用于估计样本难度的显式监督。为此，我们提出一种轻量级的免训练策略：汇集多个采用不同范式训练的预训练视觉语言模型（VLM）的信号，从而获得稳健的难度分数，无需显式训练专门模型即可估计成对偏好数据的难度。随后，我们利用这些难度分数对偏好数据重新加权，从而有效开展难度感知训练，在强调困难样本的同时防止对简单样本过拟合。

具体而言，我们提出**难**度**感**知**直**接**偏**好**优**化（DA-DPO）框架，该框架包含两个步骤：难度估计 和难度感知训练。第一步使用多个视觉语言模型评估每对偏好样本的难度。具体来说，我们利用对比式视觉语言模型（例如 CLIP（Radford 等，2021））和生成式视觉语言模型（例如 LLaVA（Liu 等，2023b）），从互补的角度估计难度。通过分布感知投票策略汇总这些模型的输出，其中每个视觉语言模型的权重根据其在训练数据上观测到的分类可靠性自适应确定。在这些分数的基础上，第二步通过动态调整 DPO 中每个样本的优化强度来进行难度感知训练。具体而言，难度分数会调整学习得到的策略与初始策略之间允许的偏离程度。这一机制能加强对困难样本的学习，同时限制对简单样本进行不必要的偏移。

我们在三个规模和能力各异的热门多模态大语言模型上进行了实验。为进行全面比较，我们报告并分析了两组基准测试上的性能：幻觉基准（Wang 等，2023a；Rohrbach 等，2018；Sun 等，2023b；Li 等，2023e）和通用多模态大语言模型基准（Hudson & Manning，2019；Liu 等，2023b；Fu 等，2023；Li 等，2023a）。实验结果展现了我们方法的有效性。

我们的主要贡献总结如下：

- 我们对多模态偏好优化训练进行了分析，并通过实验证明了过拟合问题的存在，该问题可能导致性能欠佳。
- 我们提出了一种高效且经济的框架，利用视觉语言模型（VLM）在无需额外训练的情况下估计样本难度，并通过难度感知训练利用该估计改进偏好建模。



(a) (b)

图 1：**(1a) 直接偏好优化和 DA-DPO 的性能比较**。我们展示了直接偏好优化和 DA-DPO 相较于未经偏好优化的 LLaVA v1.5 7B 所带来的性能提升。幻觉表示在 4 个幻觉基准上的性能，综合表示在 4 个综合多模态大语言模型基准上的性能。具体内容见实验部分。**(1b) 简单与困难的成对样本：**“简单样本”因偏好回答与非偏好回答之间差异明显而具有较大的分数差距；而“困难样本”的差异较小，因此对学习更有价值。

- 我们在幻觉和综合基准测试上评估了我们的方法，实验结果表明，该方法能够以经济高效的方式显著提升各种多模态大语言模型的性能。

2 预备知识

本节简要概述从基于人类反馈的强化学习（RLHF）到直接偏好优化（DPO）的流程。

基于人类反馈的强化学习基于人类反馈的强化学习（RLHF）是一种广泛采用的框架，用于使大语言模型与人类价值观和意图对齐。标准方法（Bai 等，2022a；Ouyang 等，2022）首先训练奖励模型，然后优化 KL正则化奖励目标，以平衡偏好对齐与输出多样性。该优化过程可表示为：

$$\max_{\pi_\theta} \mathbb{E}_{x\sim D,\, y\sim\pi_\theta(x)}[r_\phi(x,y)] - \beta\,\mathrm{KL}[\pi_\theta(y|x)\,\|\,\pi_{\mathrm{ref}}(y|x)], \tag{1}$$

其中， π_{ref} 是参考策略（通常为监督微调模型）， β 用于控制奖励最大化与保持接近 π_{ref}之间的权衡。该目标通常通过近端策略优化（Ouyang 等，2022）进行优化。

成对偏好优化 尽管上述基于人类反馈的强化学习取得了成功，但近端策略优化很难优化。为提升近端策略优化的效率，直接偏好优化（Rafailov 等，2024）利用最优策略对奖励函数重新参数化：

$$r(x,y) = \beta\log\left(\frac{\pi_\theta(y\mid x)}{\pi_{\mathrm{ref}}(y\mid x)}\right) + \beta\log Z(x), \tag{2}$$

其中， $Z(x)$表示用于确保正确归一化的配分函数。超参数 β与公式（1）中的 KL 权重类似，用于控制权衡：较大的值会促使 π_θ 更接近参考策略，从而保持泛化能力和鲁棒性；较小的值则会更加注重偏好对齐，但有过拟合的风险。

基于这一奖励公式，我们可以将其直接整合到 Bradley-Terry 模型中，该模型以概率方式处理成对偏好。这样一来，我们无需训练单独的奖励模型，即可优化偏好目标。优化目标如下：

$$\min_{\pi_\theta} -\mathbb{E}_{x,y_c,y_r}[\log\sigma(r(x,y_c) - r(x,y_r))], \tag{3}$$

其中 $r(x,\ y)$ 可以是由 π_θ 参数化的任意奖励函数，例如式（2）中定义的函数； y_c 和y_r 分别表示成对偏好数据中被选中和被拒绝的回答。

3 多模态偏好优化分析

本节将系统探究多模态偏好优化中普遍存在的过拟合挑战。通过实证分析，我们证明模型倾向于对较简单的训练样本过拟合，同时逐渐降低从较难样本中有效学习的能力。这一现象在 DPO（Rafailov 等，2024）等成对训练范式中尤为明显。这种过拟合行为最终会损害模型在多样化现实场景中的表现。我们通过训练动态和奖励趋势分析得出的定量证据，进一步证实了这些发现。

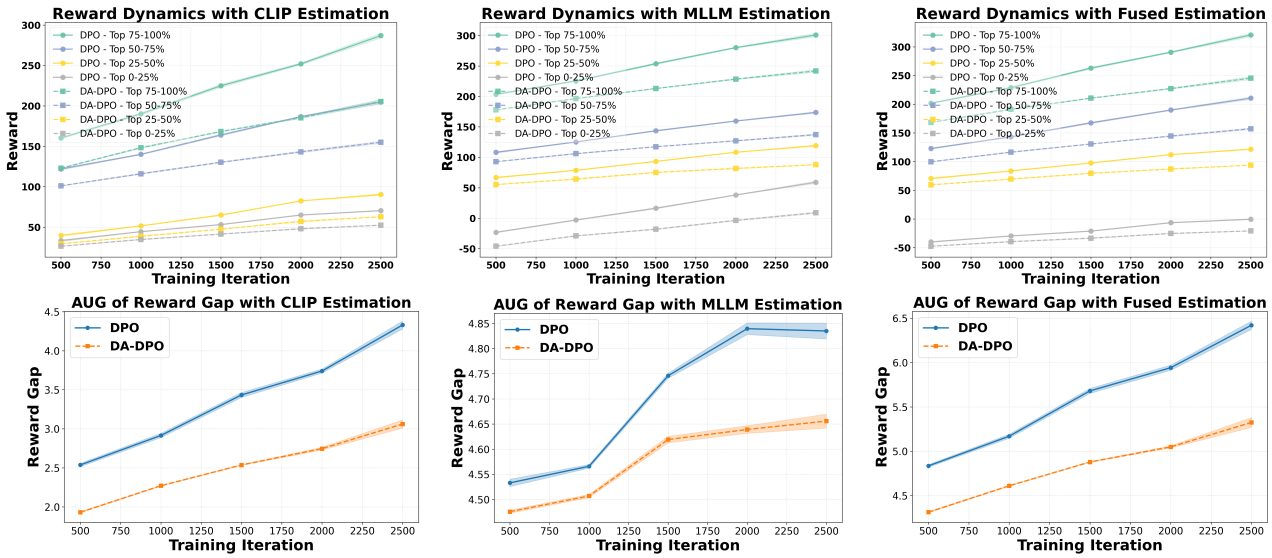


图2：**最简单与最困难样本之间的奖励动态和间隙下面积（AUG）**。我们展示了在 BPO 数据集上分别使用直接偏好优化和 DA-DPO 训练的 LLaVA-v1.5-7B 在留出验证集上的奖励轨迹。第一行展示了不同估计难度的数据分桶的奖励如何随训练迭代变化。难度通过三种不同的代理指标进行估计，详见第4.1节。这里，最简单的样本（图例中的前 75–100%）指被选中和被拒绝的回答之间差距最大的样本。第二行报告了<sty l e id='5'>间隙曲线下面积（AUG），用于量化训练过程中最简单与最困难样本之间累积的奖励差距，可作为奖励差异动态的简洁指标。图中的阴影区域表示三个独立随机种子之间的标准差；不过，由于训练随机性引起的方差较低，这一影响并不明显。

分析设置 为了系统地分析偏好优化中的过拟合，我们构建了一个受控的评估设置。我们按照 90% 与 10% 的比例，将数据集划分为训练集和留出验证集。DPO 和 DA-DPO 模型均在占 90% 的训练部分上进行训练，并在训练迭代过程中定期评估它们在验证集上的奖励表现。由于没有关于样本难度的预言机标注，我们使用三种不同的代理指标估算每个验证样本的难度，具体如第4.1节所述。根据各代理指标得出的难度排序，我们将验证样本划分为四个大小相等的分桶，从最简单到最困难依次排列。通过这一设置，我们可以考察模型对不同难度样本的奖励如何变化，从而深入了解偏好优化过程中的过拟合行为。为评估结果的统计可靠性，我们使用三个不同的随机种子重复实验，并报告相应的标准差。

奖励动态分析 我们从两个互补的角度分析奖励动态。第一个角度考察不同难度样本的奖励在整个训练过程中的变化情况。根据三种不同的代理指标（详见第4.1节），验证样本

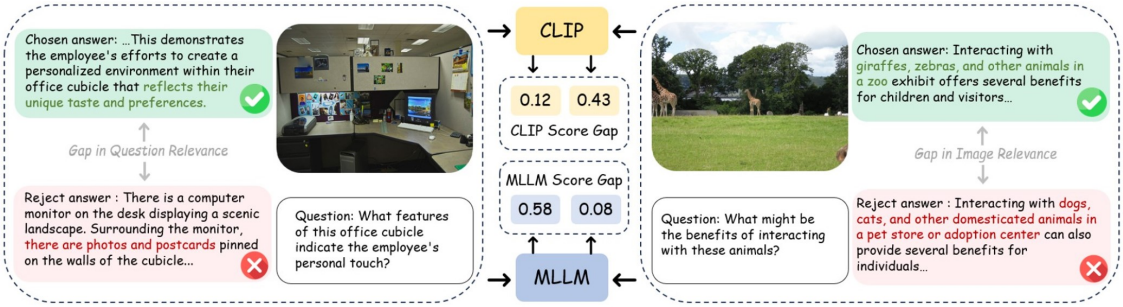


图3：**对比式和生成式视觉语言模型的示意图**。如右侧所示，CLIP 能够捕捉图像与答案之间的不对齐，因此，当被选答案和被拒答案在图像相关性上存在差异时，CLIP 分数差较大。然而，如左侧所示，多模态大语言模型更擅长捕捉问题与答案之间的逻辑联系，因此，当被选答案和被拒答案在问题相关性上存在差异时，多模态大语言模型分数差较大。

会按照难度进行排序，并划分为多个分桶，从最容易到最困难依次排列。如图2第一行所示，我们观察到，对于直接偏好优化（DPO）和我们提出的DA-DPO，较困难分桶的奖励始终低于较容易分桶的奖励。这一趋势表明，模型拟合困难样本的能力有限，而这些样本通常需要细粒度理解能力。此外，我们观察到，在*DA-DPO*中，较容易分桶的奖励增长速度相比朴素*DPO*更慢。这种较慢的增长源于DA-DPO的自适应加权机制，该机制根据对每个训练实例难度的估计来调整其重要性。通过这种方式，DA-DPO有效减少了对较容易样本的过度优化，从而缓解了标准DPO训练中经常出现的过拟合倾向。

为进一步量化这一现象，我们引入第二个视角：最容易样本与最难样本之间的间隙曲线下面积（AUG）。计算 AUG 时，首先在训练迭代过程中的每个评估点，求最容易分桶的奖励与最难分桶的奖励之差，然后对整个训练轨迹中的这一差值进行积分。这提供了一个累积指标，用于衡量训练过程中容易样本与困难样本之间的奖励差异。AUG 越大，表明优化对容易样本的偏向越强。如图2第二行所示，DA-DPO 的 AUG 始终小于朴素 DPO，表明 DA-DPO 能够对不同难度的样本进行更加均衡的优化。这些分析共同凸显了多模态偏好优化中的过拟合现象，以及 DA-DPO 缓解这一问题的有效性。

4 方法

本节介绍难度感知直接偏好优化框架，该框架以经济高效的方式解决标准多模态直接偏好优化训练中的过拟合问题。在第4.1节中，我们首先讨论偏好数据的估计，其中使用 CLIP 和多模态大语言模型评估偏好数据的难度。在第4.2节中，我们将说明如何利用预训练视觉语言模型对数据难度的估计，为难度感知DPO训练提供信息并加以引导。

4.1 数据难度估计

评估偏好数据难度的关键挑战在于缺乏显式监督。为此，我们提出了一种轻量、免训练策略，利用预训练的对比式和生成式视觉语言模型，从互补的角度估计样本难度，如图3所示。为整合多个模型的估计结果，我们采用分布感知投票策略，其中每个模型的贡献与其在训练集上的偏好分类准确率成正比。由此，无需额外训练模型，即可为每个样本得出稳健的难度分数。具体细节将在后续章节中介绍。

对比估计 CLIP 通过对比训练目标在网络规模图像描述上进行训练，并且已被证明具备有关图像与文本相关性的泛化知识。我们利用 CLIP 评估成对 DPO 数据的难度。具体而言，我们首先计算选定响应 y_c 和拒绝响应 y_r 的 CLIP 文本嵌入（分别记为 f_c 和 f_r），以及 DPO 数据中图像 m 的 CLIP 图像嵌入（记为 v_m）。CLIP 分数 c_c 和 c_r 分别表示两个响应的图文相关性，计算方式如下：

$$c_c = \mathrm{CosSim}(f_c, v_m), \quad c_r = \mathrm{CosSim}(f_r, v_m). \tag{4}$$

接下来，我们引入 CLIP 分数差 c_g，用来反映 DPO 样本的难度；分数差越大，表明被选中和被拒绝的回答在图文相关性方面越容易区分。形式上，CLIP 分数差 c_g 定义如下：

$$c_g = c_c - c_r. \tag{5}$$

为了将这一指标与其他基于视觉语言模型的估计相结合，我们通过数据集级高斯投影将 c_g 归一化到统一尺度。给定一个包含 N 个样本的数据集，归一化分数的计算方式如下：

$$\hat{c}_g = \Phi(\frac{c_g - \mu_{c_g}}{\sigma_{c_g}}), \tag{6}$$

其中，μ_{c_g} 和σ_{c_g} 分别是 DPO 数据集中所有 CLIP 分数差 c_g 的均值和方差。归一化后的 CLIP 分数差为 $\hat{c}_g \in [0, 1]$，而 $\Phi(\cdot)$表示标准高斯分布的累积分布函数。

生成式估计 近期的多模态大语言模型（MLLMs）基于大语言模型（LLMs）构建，通过将视觉编码器连接到大语言模型，并采用语言建模目标，学习与视觉世界交互。为简化表述，我们将其称为生成式视觉语言模型。尽管这类模型容易产生幻觉，但它们从与 CLIP 互补的另一个角度，为评估 DPO 数据难度提供了有用信息。我们提取多模态大语言模型对选定响应 y_c 和拒绝响应 y_r的分数，分别记为 $m_c \in \mathbb{R}$ 和 $m_r \in \mathbb{R}$，其定义如下：

$$m_c = \sum_{t=1}^{T} \log\left(\mathrm{P}(y_c^t | m, y_c^1, \ldots, y_c^{t-1}; \pi_{\mathrm{ref}})\right), \tag{7}$$

$$m_r = \sum_{t=1}^{T} \log\left(\mathrm{P}(y_r^t | m, y_r^1, \ldots, y_r^{t-1}; \pi_{\mathrm{ref}})\right), \tag{8}$$

其中，y_c^t 和 y_r^t 是被选中和被拒绝的回答中的 t^{th} 标记，T 是序列长度。为评估成对 DPO 数据的难度，按如下方式计算多模态大语言模型分数差 m_g：

$$m_g = m_c - m_r. \tag{9}$$

出于与 CLIP 分数归一化类似的原因，我们采用高斯归一化来获得归一化的多模态大语言模型分数差 $\hat{m}_g \in [0, 1]$。具体过程定义如下：

$$\hat{m}_g = \Phi(\frac{m_g - \mu_{m_g}}{\sigma_{m_g}}), \tag{10}$$

其中，μ_{m_g}，σ_{m_g}是 DPO 数据集中所有多模态大语言模型分数差 m_g的均值和方差。

分布感知投票融合为此，我们从两个角度评估成对 DPO 数据，得到两个难度分数。如图 4 所示，两个视觉语言模型在偏好分类上的表现各不相同：CLIP 在与图像描述相关的偏好数据上表现出色，而多模态大语言模型在与视觉问答相关的

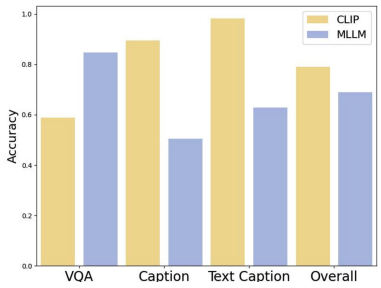


图 4：**偏好分类对比**。偏好分类用于评估预训练视觉语言模型对被选答案给出的奖励分数是否高于被拒答案。我们报告了三个子类别的分类准确率，以及在 BPO训练数据集上的总体表现。

数据上表现更好。我们提出一种数据驱动的投票策略，根据偏好分类结果自适应地组合难度分数。具体而言，我们使用 CLIP 和多模态大语言模型的分类准确率（分别记为 cls_c 和 cls_m），来确定每个 DPO 样本中 $beta$ 的权重，如下所述:

$$\hat{\beta} = (\frac{cls_c}{cls_c + cls_m}\hat{c}_g + \frac{cls_m}{cls_c + cls_m}\hat{m}_g) + 1, \tag{11}$$

我们在此添加一个常数项 1，以提高训练过程的稳定性。如果没有这一项，当 $\hat{c}_g$ 和 $\hat{m}_g$ 都接近零时，所得的 β 也会趋近于零，可能导致训练崩溃。

4.2 难度感知训练

估计偏好对的难度后，我们得到一个能够反映成对 DPO 数据难度的稳健分数。借鉴先前的工作（Wu 等，2024），我们通过自适应校准式（1）中的 β 来进行难度感知DPO训练。与标准DPO训练相比，这种方法能够调整每个训练样本的权重，从而减少对简单样本的过拟合。形式上，所提出的难度感知DPO目标函数定义如下:

$$\mathcal{L}_{\text{DA-DPO}} = -\mathbb{E}_{(x,y_c,y_r,\hat{\beta})\sim D}\left[\log\sigma\left(\hat{\beta}\, r(x,y_c) - \hat{\beta}\, r(x,y_r)\right)\right], \tag{12}$$

其中，$\hat{\beta}$ 是难度感知缩放因子，$r(x, y)$表示式（2）中定义的奖励函数，该函数本身包含温度系数 β，y_c 和 y_r 分别是被选中和被拒绝的回答，而 σ（·）是Sigmoid函数。

5 实验

5.1 实现细节

训练设置 为验证所提方法的有效性，我们采用BPO（Pi 等，2024）中的成对偏好数据集。该数据集包含18万条成对偏好数据，其中负面回答由图像弱化提示和大语言模型错误注入生成。在训练参数方面，我们训练模型1个周期，并将 β 设为0.2。对于LLaVA V1.5，我们遵循先前工作（Pi 等，2024），采用LoRA（Hu等人，2021）训练，秩设为32，LoRA alpha设为256。学习率设为2e-6。对于LLaVA-OneVision，我们使用官方推荐的训练脚本进行全量微调，学习率为5e-7。LLaVA V1.5 7B模型的训练约需7小时，LLaVA-OneVision 7B则约需22小时。在分析部分，我们保留分数排名最高和最低的20%，作为高置信度的基于CLIP的预测结果。

视觉语言模型的选择 在视觉语言模型的选择上，我们采用 CLIP（Radford 等，2021）ViTL/14@336。对于每个数据样本中的优选和负面文本回答，我们对尽可能长的文本进行编码，

表 1：**幻觉与综合基准测试结果**。 对于每一列， ↑ 符号表示数值越高越好，而 ↓ 符号表示数值越低越好。 $AMB.Gen.$ 和 $AMB.\ Dis.$ 分别指 AMBER 基准测试的生成和判别部分。 $*$ 符号表示 BPO（Pi 等，2024）官方模型权重的评估结果。我们说明，‡ 侧重于利用 CLIP 知识构建偏好数据，而我们则利用 CLIP 解决标准DPO训练中的过拟合问题。mDPO†通过解决对语言偏好的过度优先问题，提供了一种改进多模态偏好优化的替代方法；这与 DA-DPO 方法正交。

	AMB.^{G}				AMB.^{D}	ObjHal			MMHal		POPE	LLaVA^W	GQA	Seed^I	MME^P	MME^C
	C_s ↓	覆盖率 ↑	幻觉率 ↓	认知 ↓	F1 ↑	C_s ↓	C_i ↓	分数 ↑	幻觉率 ↓	F1 ↑		分数 ↑	准确率↑	准确率↑	分数↑	分数↑
仅供参考（不可直接比较）																
CogVLM	5.6	57.2	23.6	1.3	-	-	-	-	-	-	-	-	-	-	-	-
mPLUG-Owl2	10.6	52.0	39.9	4.5	-	-	-	-	-	-	56.1	-	-	-	-	-
InstructBLIP	8.8	52.2	38.2	4.4	-	-	-	-	-	-	58.2	-	58.8	1212.8	-	
Qwen-VL	5.5	49.4	23.6	1.9	-	36.0	21.3	2.89	0.43	-	-	59.3	-	1487.6	360.7	
GPT-4V	4.6	67.1	30.7	2.6	-	13.6	7.3	3.49	0.28	-	-	-	-	-	-	
LLaVA-v1.5-7B	7.8	51.0	36.4	4.2	74.7	54.7	15.9	2.19	0.57	85.8	63.8	62.0	66.1	1510.7	307.5	
+ HA-DPO	7.2	33.6	19.7	2.6	-	39.9	19.9	1.97	0.60	-	-	-	-	-	-	
+ CLIP-DPO‡	3.7	47.8	16.6	1.3	-	-	-	-	-	85.8	-	-	-	1468.7	-	
+ mDPO†	4.4	52.4	24.5	2.4	-	35.7	9.8	2.39	0.54	-	-	-	-	-	-	
主要结果																
LLaVA-v1.5-7B	7.8	51.0	36.4	4.2	74.7	54.7	15.9	2.19	0.57	85.8	63.8	**62.0**	**66.1**	**1510.7**	307.5	
+ 直接偏好优化*	5.5	**58.4**	35.7	**2.0**	83.9	43.3	10.0	**2.24**	0.53	84.3	70.5	45.3	57.8	1409.4	315.0	
+ DA-DPO	**4.3**		**28.0**	2.1	**85.6**	**3**		2.22	**0.50**	**85.9**	**75.4**	59.8				
LLaVA-v1.5-13B	7.1	52.4	33.9	3.8	**88.6**	56.3	15.8	2.22	0.57	**86.9**	71.2	**63.0**	**67.5**	**1496.7**	**308.9**	
+ 直接偏好优化	6.6	**61.9**	45.7	2.5	85.5	39.0	8.8	2.05	0.55	86.0	74.1	60.9	53.5	1450.6	276.4	
+ DA-DPO	**5.1**		**32.2**	**1.9**	87.3				**0.48**	85.6	**74.5**	61.5				
LLaVA-OV-7B	8.4	**74.8**	70.4	9.8	90.6	41.3	8.1	2.76	0.37	**87.2**	**80.0**	**59.6**	**75.8**	**1580.4**	**406.4**	
+ 直接偏好优化	7.7	66.5	59.0	4.3	**91.9**	34.3	8.6	2.61	0.38	86.1	72.9	55.0	73.7	1545.4	364.6	
+ DA-DPO	**7.0**		**51.8**	**3.7**	91.7				**0.30**	86.0	73.7	59.2				

从开头截取，并截断其余文本。对于多模态大语言模型，我们使用 LLaVA v1.5 7B（Liu 等，2023a）根据图像和问题计算回答的概率。我们还在第 5.5 节中对这些模型的选择进行了消融研究。

5.2 评估基准

为了全面评估偏好优化对多模态大语言模型的影响，我们选取了两类基准测试。幻觉基准用于衡量模型减少事实性错误的能力，这是多模态偏好对齐的首要目标。综合基准则评估模型的一般多模态能力，确保减少幻觉的同时不会牺牲总体性能。

幻觉评估。遵循以往工作（Pi 等，2024；Wang 等，2024a；Ouali 等，2024），我们在多个幻觉基准上对 DA-DPO 进行全面评估，包括 AMBER（Wang 等，2023a）、MMHalBench（Sun 等，2023b）、Object HalBench（Rohrbach 等，2018）和 POPE（Li 等，2023e）。1）AMBER 提供了一个多维框架，适用于评估生成任务和判别任务。2）MMHalBench 是一个问答基准，包含八种问题类型和 12 个物体主题。我们使用 GPT-4，并遵循官方评估脚本。3）Object HalBench 是用于评估物体幻觉的标准基准，我们遵循（Wang 等，2024a）中的设置。

表2：**不同视觉语言模型估计的影响**。我们报告了采用不同策略训练的模型在幻觉基准上的表现，例如 AMBER 基准测试和综合基准测试 SeedBench。蓝色行表示采用的策略。

	VLMs		AMB.^{G}				AMB.^{D}	Seed^I
	CLIP	MLLM	$C_s \downarrow$	覆盖率 ↑	幻觉率 ↓	认知 ↓	F1 ↑	分数 ↑
DPO	×	×	5.5	58.4	35.7	2.0	83.9	57.8
	✓	×	4.7	**58.0**	31.9	2.0	85.2	63.8
Ours	×	✓	4.6	57.6	29.9	2.3	85.3	64.3
	✓	✓	**4.3**	57.4	**28.0**	**2.1**	**85.6**	**64.8**

表3：**使用 VLFeedback 和 LLaVA-RLHF 数据集训练的 LLaVA v1.5 7B 的结果**。VLFeedback 是通过收集多个视觉语言模型的回答并使用 GPT4-V 过滤结果而自动构建的数据集。LLaVA-RLHF 是一个由人工标注的数据集。

	AMB.^{G}				AMB.^{D}	ObjHal		MMHal		POPE	LLaVA^W	GQA	Seed^I	MME^P	MME^C
	$C_s \downarrow$	覆盖率 ↑	幻觉率 ↓	认知 ↓	F1 ↑	$C_s \downarrow$	$C_i \downarrow$	分数 ↑	幻觉率 ↓	F1 ↑	分数 ↑	准确率↑	准确率↑	分数↑	分数↑
LLaVA-v1.5-7B	7.8	51.0	36.4	4.2	74.7	54.7	15.9	2.19	0.57	85.8	63.8	62.0	66.1	1510.7	307.5
VLFeedback															
+ 直接偏好优化	6.5	**55.1**	34.5	**2.3**	**84.6**	49.0	13.0	2.19	0.65	84.6	72.1	59.8	63.5	1368.5	294.2
+ DA-DPO	**5.6**	53.0	**29.7**	2.7	**85.8**	**48.3**			**0.53**	84.2	**73.0**	**60.3**			
LLaVA-RLHF															
+ 直接偏好优化	7.3	**50.8**	33.7	3.9	**86.6**	58.3	16.9	2.10	**0.57**	84.4	68.7	60.7	64.5	1415.9	**343.9**
+ DA-DPO	**5.8**	50.7	**27.6**	**3.0**	**86.6**	**48.0**			0.59	**84.5**	**71.3**	**60.9**			

4）POPE 使用基于轮询的查询来评估模型的幻觉。我们报告 POPE 中三类问题的平均 F1分数。

综合评估。为了评估多模态大语言模型的有用性，我们采用以下基准：1）LLaVA-Bench（Liu 等，2023b），这是一个现实世界基准，包含 60 项任务，用于评估 LLaVA 的视觉指令遵循和问答能力。我们使用官方脚本，通过 GPT-4 计算分数。2）Seedbench（Li 等，2023a），包含 1.4 万个多项选择视觉问答样本，用于评估多模态大语言模型的综合能力。3）MME（Fu 等，2023）通过是非题评估感知和认知能力。4）GQA（Hudson & Manning，2019年）采用开放式答案生成形式，评估现实世界中的视觉推理和组合式问答能力。

5.3 基线

所提出的难度感知直接偏好优化框架旨在以经济高效的方式引入预训练视觉语言模型，从而改进成对偏好优化。我们主要在三个多模态大语言模型上，将 DA-DPO 与标准直接偏好优化（Rafailov 等，2024）进行比较：LLaVA V1.5 7B/13B（Liu 等，2023a）和 LLaVA-OneVision 7B（Li 等，2024）。此外，我们还提供其他多模态大语言模型的结果以供参考，例如 GPT4-V（OpenAI，2023年）、CogVLM（Wang 等，2023b）、mPLUG-Owl2（Ye 等，2023）、InstructBLIP（Dai 等，2023）、Qwen-VL（Bai 等，2023）、HA-DPO（Zhao 等，2023）、CLIP-DPO（Ouali 等，2024）和 mDPO（Wang 等，2024a）。

5.4 结果分析

幻觉基准 为展示所提出方法在减少幻觉方面的有效性，我们在表 1 中给出了各种幻觉基准上的评估结果。我们观察到，与 DPO 相比，DA-DPO 提升了模型在大多数基准测试上的性能。例如，使用 LLaVA v1.5 7B 进行训练时，AMBER 生成任务中的幻觉率 从 35.7 降至 28.0。此外，性能

表 4：**视觉语言模型的选择**。其中$ViTL$ 表示 CLIP ViTL/14@336，$EVA\ 8B$ 表示 EVA-CLIP 8B。$LLaVA\ 7B$ 表示 LLaVA v1.57B，$OV\ 7B$ 表示 LLaVA-Onevision 7B。

模型选择		AMB.^{G}				AMB.^{D}	随机种子^I
CLIP	MLLM	C_s ↓	覆盖率 ↑	幻觉率 ↓	认知 ↓	F1 ↑	分数 ↑
ViTL	LLaVA 7B	**4.3**	57.4	28.0	2.1	85.6	64.8
EVA 8B	LLaVA 7B	4.9	57.1	31.8	2.6	83.4	56.2
ViTL	OV 7B	4.6	**57.7**	31.4	2.0	85.7	65.2
EVA 8B	OV 7B	**4.3**	55.6	**25.5**	2.0	**86.0**	**65.5**

表 5：**DA-DPO 与直接筛选基线的比较（移除 10%、25%、50% 的简单样本）**。带有 ↑ 的指标表示数值越高越好；带有 ↓ 的指标表示数值越低越好。

模型	AMBER^G				AMBER^D	POPE	MMHal		ObjHall	LLaVA^W	GQA	Seed	MME	
	C_s ↓	$Cov.$ ↑	$Hal.$ ↓	$Cog.$ ↓	F1↑	F1↑	分数↑	幻觉率↓	C_s ↓	C_i ↓	分数↑	准确率↑	准确率↑	P↑ C↑
DPO	5.5	58.4	35.7	**2.0**	83.9	84.3	2.23	0.53	43.3	10.0	70.48	45.3	57.7	1409.4 315.0
过滤 10%	5.6	**63.2**	39.6	2.6	85.9	84.9	2.22	0.51	39.0	9.3	72.49	53.4	50.3	1365.1 300.0
过滤 25%	5.3	61.3	37.8	2.3	85.8	84.3	**2.29**	**0.50**	38.3	9.0	67.04	53.3	51.1	1318.6 **330.3**
过滤 50%	5.3	62.2	38.3	2.0	**86.0**	85.5	2.03	0.56	42.7	10.2	71.21	**56.3**	52.2	1359.7 306.7
DA-DPO	**4.3**	57.4	**28.0**	2.1	85.6	**85.9**	2.22	**0.50**	**37.7**		**75.38**	**59.86**		

三个多模态大语言模型在对象幻觉基准上的性能均显著提升，这表明 DA-DPO 大幅减少了图像描述中的幻觉。

综合基准 我们在五个综合基准测试上评估所提出的方法。表1所示结果表明，尽管偏好优化能够减少幻觉，但通用能力的表现会受到影响。然而，与标准直接偏好优化相比，DA-DPO 在大多数基准测试上减轻了这种性能下降。此外，DA-DPO 显著提升了对话能力：使用 LLaVA v1.5 7B 时，其在 LLaVA-Bench 上的表现达到 75.4，而直接偏好优化为 70.5。

5.5 消融研究

难度感知训练的影响 为更好地理解所提出的方法，我们进行了消融研究，以评估各项设计的有效性。如表2所示，我们以标准的直接偏好优化（DPO）（Rafailov 等，2024）为起点。随后，我们加入来自 CLIP 的奖励分数，以控制训练过程中的样本权重，并展示仅使用多模态大语言模型奖励分数的直接偏好优化结果。最后，我们采用自适应融合策略结合两种奖励分数。结果表明，使用 CLIP 或多模态大语言模型的奖励分数均能提升幻觉率 和$CHAIR_s$，而结合两者可取得最佳表现，这凸显了各项框架设计的必要性。

偏好数据集的影响为证明所提出的方法能够缓解不同类型多模态成对偏好数据中的数据偏差，我们使用了 VLFeedBack（Li 等，2023d）。该数据集包含由能力水平各异的多模态大语言模型生成、并由 GPT4-V（OpenAI，2023年）评分的 8 万条回答。与 BPO（Pi 等，2024年）相比，该数据集采用了不同的自动生成方法，涵盖了多模态偏好数据的主流生成流程。我们遵循 mDPO（Wang 等，2024a）的设置，选取了 1 万个偏好数据样本。此外，我们还在 LLaVA-RLHF（Sun 等，2023b）上进行训练，这是使用最广泛的人工标注多模态偏好数据集，包含 1 万个样本。如表3所示，DA-DPO 在大多数基准测试中优于直接偏好优化（DPO），这表明过拟合是多模态偏好数据中的普遍问题，而 DA-DPO 能够以较低成本有效缓解这一问题。

视觉语言模型的选择 我们对视觉语言模型的选择进行了消融研究。这些模型用于估计成对偏好数据的难度，并在训练过程中指导难度感知偏好优化

表 6：**使用基于 CLIP 和基于多模态大语言模型的代理指标估计的不同难度分桶上的性能**。我们使用原始 DPO 对 LLaVA-1.5-7B 进行微调，其中每个分桶中的训练数据均根据估计奖励差距从 BPO 数据集中采样。前 *0–25%*分桶对应奖励差距最小的样本。

	AMB.^{G}				AMB.^{D}	ObjHal		POPE
分桶	C_s↓	覆盖率↑	幻觉率↓	认知↓	F1↑	C_s↓	C_i↓	F1↑
CLIP 估计								
前 0–25%	5.3	62.0	37.6	2.2	**86.8**	**41.3**	10.8	85.6
前 25–50%	**5.1**	**62.7**	**37.5**	**2.0**	86.2	43.0	**9.7**	**85.8**
前 50–75%	5.2	61.1	37.9	2.4	85.7	51.0	10.8	85.8
前 75–100%	5.4	55.0	38.9	2.5	85.6	52.7	12.9	85.7
MLLM 估计								
前 0–25%	4.9	**59.1**	35.4	2.5	85.6	46.7	11.8	85.7
前 25–50%	**4.6**	53.9	**32.3**	**1.9**	**86.6**	**42.3**	**9.0**	85.8
前 50–75%	5.4	55.5	30.5	2.0	84.6	48.3	11.4	**86.5**
前 75–100%	5.5	54.3	32.4	2.2	84.8	46.3	10.8	86.3

在所提出的框架中。我们使用两种参数规模不同的 CLIP 模型进行实验：CLIP ViT- L/14@336（Radford 等，2021）和 EVA-CLIP 8B（Sun 等人，2023a）。对于多模态大语言模型，我们使用 LLaVA v1.5 7B（Liu 等，2023a）和 LLaVA-OneVision 7B（Li 等，2024）。如表 4 所示，我们发现，不同能力的视觉语言模型取得的性能相近，这归因于公式（10）和公式（6）中的高斯归一化。该归一化确保视觉语言模型仅对偏好数据进行排序，使我们提出的框架能够稳健应对视觉语言模型选择的变化。

与直接过滤基线的比较 为进一步验证我们的样本重新加权策略的有效性，我们将 DA-DPO 与一种直接从训练数据中剔除简单样本 的基线方法进行比较。具体而言，我们根据难度指标，从 BPO 数据集中剔除 10%、25% 和 50% 的简单样本，并使用 DPO 算法训练模型。表 5 总结了 DA-DPO 和直接筛选基线在多种基准测试上的表现。我们发现，DA-DPO 在大多数指标上始终优于基于过滤的方法，尤其是在 POPE、MMHal 和物体幻觉等幻觉相关基准测试上。值得注意的是，直接过滤基线的表现会随过滤比例不同而波动，剔除更多简单样本并未带来持续的提升。

我们认为，直接过滤效果有限，源于其固有的权衡：虽然它剔除了真正无信息量的样本，但也丢弃了一部分有价值的训练数据，从而降低了偏好信息的多样性和覆盖范围。相比之下，DA-DPO 通过降低简单样本的权重，而非进行硬过滤，解决了这一问题；这种做法既保留了数据多样性，又突出了信息丰富的样本。因此，DA-DPO 能够在难度和标注风格各异的基准测试上保持鲁棒性。

估计奖励差距与模型幻觉为考察奖励差距与幻觉行为之间的关系，我们设计了一项受控验证实验。具体而言，我们使用基于 CLIP 和基于多模态大语言模型的代理模型来估计训练样本的奖励差距，并根据估计差距范围将数据均分为四个分桶。随后，我们在每个分桶上独立训练一个原始 DPO 模型，以分析样本难度如何影响偏好优化。如表 6 所示，幻觉减少幅度最大的是处于 25–50% 分桶中的样本，而随着奖励差距增大，幻觉减少幅度逐渐降低。有趣的是，奖励差距最小的分桶（0–25%）并未带来最显著的幻觉改善。我们认为，这是因为数据构建过程中，该子集中可能包含噪声样本或容易引起混淆的样本，例如“拒绝”的答案在语义上实际上正确的情况，这可能会干扰有效的模型优化。

6 相关工作

6.1 视觉语言模型

视觉语言模型（VLM）（Radford 等，2021；Jia 等，2021；Li 等，2022）极大地推动了多模态理解的发展，并在广泛的下游任务中取得了优异表现（Guo 等，2022；Ning 等，2023）。CLIP（Radford 等，2021）等对比式视觉语言模型通过大规模对比目标对齐图像与文本，并在识别任务上展现出出色的零样本迁移能力（Nukrai 等，2022；Qiu 等，2024；Liao 等，2022；Ning 等，2023；Yao 等，2022；Peebles 与 Xie，2022）。随着大语言模型的兴起，后续研究将大语言模型与视觉编码器相结合，以增强图像条件文本生成能力（Li 等，2023b；Dai 等，2023；Liu 等，2023b；Zhu 等，2023；Chen 等，2023；Lin 等，2023），从而实现指令遵循和开放式推理。近期的研究工作（Liu 等，2023a；Li 等，2024；Gao 等，2024；Bai 等，2025）进一步利用高分辨率输入、多图像上下文乃至视频序列来提升视觉能力。尽管取得了这些进展，视觉语言模型在将视觉证据与文本回应对齐时仍容易产生幻觉，因此需要采用基于偏好的优化方法来提升忠实度。

6.2 大语言模型中的偏好对齐

对齐问题（Leike 等，2018）旨在确保智能体行为符合人类意图。早期方法采用基于人类反馈的强化学习（RLHF）（Bai 等，2022a;b；Glaese 等，2022；Nakano 等，2021；Ouyang 等，2022；Scheurer 等，2023；Stiennon 等，2020；Wu 等，2021；Ziegler 等，2019），通过近端策略优化（PPO）（Schulman 等，2017）等策略优化方法来最大化人工标注奖励。近来，直接偏好优化（DPO）（Rafailov 等，2024）将对齐重新表述为针对离线偏好数据的直接优化问题，从而无需强化学习。Gibbs-DPO（Xiong 等，2023）等扩展方法支持在线偏好优化，而 β-DPO（Wu 等，2024）则通过引入基于奖励模型的动态校准，解决了对温度参数 β 的敏感性问题。然而，这类方法在多模态领域的效果较差，因为奖励模型容易受到奖励投机的影响（Sun 等，2023b）。其他研究探索了替代策略，例如自我奖励机制或数据重加权（Wang 等，2024b；Zhou 等，2024a），以缓解偏好数据中过度自信的标注和分布偏差等问题。

6.3 多模态模型中的偏好对齐

近期的研究已将偏好对齐从纯语言模型扩展到多模态领域。一项主要工作方向是构建多模态偏好数据集。（Sun 等，2023b；Yu 等，2024a）收集人工标注，而其他研究则借助 GPT-4V 等强大的多模态模型（Li et al., 2023c；Yu et al., 2024b；Zhou et al., 2024d；Yang 等，2025）生成偏好信号。然而，人工标注和大模型推理的成本都高得难以承受，限制了其可扩展性。为解决这一问题，其他方法（Deng et al., 2024；Pi 等，2024）探索了用于偏好数据生成的自动化或自训练方法。这些研究通过损坏图像或误导性提示合成不受偏好响应，使模型能够在没有外部监督的情况下学习偏好。另一项研究方向（Ouali 等，2024）利用 CLIP 为多种候选响应打分，并根据图文相似度对受偏好与不受偏好的输出进行排序。从训练目标的角度来看，大多数多模态对齐方法采用标准的 DPO目标函数（Li et al., 2023c；Zhao et al., 2023；Zhou et al., 2024b），在配对数据上优化偏好。其他方法则采用强化学习（Sun 等，2023b）或对比学习（Sarkar et al., 2024；Jiang et al., 2024）来改善对齐。为进一步减少对纯语言信号的过拟合，（Ouali 等，2024）扩展了 DPO，同时优化文本和视觉偏好。尽管取得了这些进展，多模态 DPO 仍容易过拟合，尤其是在不平衡偏好数据上训练时。本文提出一种通用的难度感知训练框架，明确考虑样本难度，从而缓解过拟合并提升多模态偏好优化的鲁棒性。

7 结论

本文对多模态偏好优化中的过拟合问题进行了实证分析，该问题往往源于不平衡数据分布。为解决这一问题，我们提出 DA-DPO，这是一种兼具成本效益的框架，由难度估计和难度感知训练组成。我们的方法利用预训练的对比式和生成式视觉语言模型，以无需训练的方式估计样本难度，并根据这些估计自适应地重新加权数据——提高较难样本的权重，同时防止对较容易样本过拟合。在幻觉基准和通用基准上的实验表明，这一范式能够有效改善多模态偏好优化。

局限性 尽管取得了这些令人鼓舞的结果，我们的框架依赖于这样一个假设：预训练视觉语言模型能够可靠地评估偏好数据。虽然自适应投票策略在现有数据集上展现出鲁棒性，但对于与这些视觉语言模型的预训练目标存在显著差异的领域，其泛化能力仍不确定。未来的工作可以探索整合领域自适应或自我改进机制，以进一步增强鲁棒性。

8 致谢

本研究得到国家自然科学基金委员会（NSFC）项目62350610269、上海市以人为本人工智能前沿科学研究基地以及教育部智能感知与人机协同重点实验室（上海科技大学）的资助。本研究还得到上海科技大学高性能计算平台的支持。

参考文献

Aishwarya Agrawal、Jiasen Lu、Stanislaw Antol、Margaret Mitchell、C. Lawrence Zitnick、Devi Parikh 和 Dhruv Batra。视觉问答：视觉问答。国际计算机视觉杂志，123：4–31，2015年。

Jinze Bai、Shuai Bai、Shusheng Yang、Shijie Wang、Sinan Tan、Peng Wang、Junyang Lin、Chang Zhou 和 Jingren Zhou。Qwen-VL：一款能力多样的前沿大型视觉语言模型。*arXiv*，abs/2308.12966，2023年。

Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang、Humen Zhong、Yuanzhi Zhu、Mingkun Yang、Zhaohai Li、Jianqiang Wan、Pengfei Wang、Wei Ding、Zheren Fu、Yiheng Xu、Jiabo Ye、Xi Zhang、Tianbao Xie、Zesen Cheng、Hang Zhang、Zhibo Yang、Haiyang Xu 和 Junyang Lin。Qwen2.5-VL 技术报告。*arXiv* 预印本 *arXiv:2502.13923*，2025年。

Yuntao Bai、Andy Jones、Kamal Ndousse、Amanda Askell、Anna Chen、Nova DasSarma、Dawn Drain、Stanislav Fort、Deep Ganguli、Tom Henighan 等。通过人类反馈强化学习训练有帮助且无害的助手。*arXiv* 预印本 *arXiv:2204.05862*，2022 年 a。

Yuntao Bai、Saurav Kadavath、Sandipan Kundu、Amanda Askell、Jackson Kernion、Andy Jones、Anna Chen、Anna Goldie、Azalia Mirhoseini、Cameron McKinnon 等。宪法式人工智能：来自人工智能反馈的无害性。*arXiv* 预印本 *arXiv:2212.08073*，2022 年 b。

Jun Chen、Deyao Zhu1、Xiaoqian Shen1、Xiang Li、Zechun Liu2、Pengchuan Zhang、Raghuraman Krishnamoor-thi2、Vikas Chandra2、Yunyang Xiong 和 Mohamed Elhoseiny。MiniGPT-v2：作为视觉-语言多任务学习统一接口的大语言模型。*arXiv* 预印本 *arXiv:2310.09478*，2023年。

Wenliang Dai、Junnan Li、Dongxu Li、Anthony Meng Huat Tiong、Junqi Zhao、Weisheng Wang、Boyang Albert Li、Pascale Fung 和 Steven C. H. Hoi。InstructBLIP：通过指令微调迈向通用视觉语言模型。*arXiv*，abs/2305.06500，2023年。

Yihe Deng、Pan Lu、Fan Yin、Ziniu Hu、Sheng Shen、James Zou、Kai-Wei Chang 和 Wei Wang。通过在图像理解上进行自训练增强大视觉语言模型。*arXiv*预印本 *arXiv:2405.19716*，2024年。

Chaoyou Fu、Peixian Chen、Yunhang Shen、Yulei Qin、Mengdan Zhang、Xu Lin、Zhenyu Qiu、Wei Lin、Jinrui Yang、Xiawu Zheng 等。MME：多模态大语言模型的综合评测基准。*arXiv*预印本 *arXiv:2306.13394*，2023年。Peng Gao、Renrui Zhang、Chris Liu、Longtian Qiu、Siyuan Huang、Weifeng Lin、Shitian Zhao、Shijie Geng、Ziyi Lin、Peng Jin、Kaipeng Zhang、Wenqi Shao、Chao Xu、Conghui He、Junjun He、Hao Shao、Pan Lu、Hongsheng Li 和 Yu Qiao。Sphinx-x：为多模态大语言模型系列扩展数据和参数。*ArXiv*，abs/2402.05935，2024年。Amelia Glaese、Nat McAleese、Maja Trębacz、John Aslanides、Vlad Firoiu、Timo Ewalds、Maribeth Rauh、Laura Weidinger、Martin Chadwick、Phoebe Thacker 等。通过有针对性的人类判断改进对话智能体的对齐。*arXiv*预印本 *arXiv:2209.14375*，2022年。Ziyu Guo、Renrui Zhang、Longtian Qiu、Xianzheng Ma、Xupeng Miao、Xuming He 和 Bin Cui。Calip：利用无参数注意力实现 CLIP 的零样本增强。*arXiv*预印本 *arXiv:2209.14169*，2022年。Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。LoRA：大语言模型的低秩适配。*arXiv*预印本 *arXiv:2106.09685*，2021年。Drew A. Hudson 和 Christopher D. Manning。GQA：面向真实世界视觉推理和组合式问答的新数据集。*2019*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第6693–6702页，2019年。Chao Jia、Yinfei Yang、Ye Xia、Yi-Ting Chen、Zarana Parekh、Hieu Pham、Quoc V. Le、Yun-Hsuan Sung、Zhen Li 和 Tom Duerig。利用噪声文本监督扩展视觉和视觉语言表征学习。*ArXiv*，abs/2102.05918，2021年。

Chaoya Jiang、Haiyang Xu、Mengfan Dong、Jiaxing Chen、Wei Ye、Ming Yan、Qinghao Ye、Ji Zhang、Fei Huang 和 Shikun Zhang。《用于多模态大语言模型的幻觉增强对比学习》。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 27036–27046 页，2024年。

Jan Leike、David Krueger、Tom Everitt、Miljan Martic、Vishal Maini 和 Shane Legg。《通过奖励建模实现可扩展智能体对齐：一项研究方向》。*arXiv* 预印本 *arXiv:1811.07871*，2018年。

Bo Li、Yuanhan Zhang、Dong Guo、Renrui Zhang、Feng Li、Hao Zhang、Kaichen Zhang、Yanwei Li、Ziwei Liu 和 Chunyuan Li。《LLaVA-OneVision：轻松实现视觉任务迁移》。*arXiv*，abs/2408.03326，2024年。

Bohao Li、Rui Wang、Guangzhi Wang、Yuying Ge、Yixiao Ge 和 Ying Shan。Seed-bench：使用生成式理解对多模态大语言模型进行基准测试。*arXiv*，abs/2307.16125，2023a。

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。*arXiv* 预印本 *arXiv:2301.12597*，2023b。Lei Li、Zhihui Xie、Mukai Li、Shunian Chen、Peiyi Wang、Liang Chen、Yazheng Yang、Benyou Wang 和 Lingpeng Kong。Silkie：大型视觉语言模型的偏好蒸馏。*arXiv* 预印本 *arXiv:2312.10665*，2023c。Lei Li、Zhihui Xie、Mukai Li、Shunian Chen、Peiyi Wang、Liang Chen、Yazheng Yang、Benyou Wang 和 Lingpeng Kong。Silkie：大型视觉语言模型的偏好蒸馏。*arXiv*，abs/2312.10665，2023d。URL `https://api.semanticscholar.org/CorpusID:266348439`。Liunian Harold Li、Pengchuan Zhang、Haotian Zhang、Jianwei Yang、Chunyuan Li、Yiwu Zhong、Lijuan Wang、Lu Yuan、Lei Zhang、Jenq-Neng Hwang 等。基于视觉 grounding 的语言-图像预训练。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第10965–10975页，2022年。

Yifan Li、Yifan Du、Kun Zhou、Jinpeng Wang、Wayne Xin Zhao 和 Ji-Rong Wen。评估大型视觉语言模型中的物体幻觉。*arXiv* 预印本 *arXiv:2305.10355*，2023年e。

Yue Liao、Aixi Zhang、Miao Lu、Yongliang Wang、Xiaobo Li 和 Si Liu。Gen-vlkt：简化关联并增强人-物交互检测的交互理解。*2022*年 *IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第 20091–20100 页，2022年。

Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár 和 C Lawrence Zitnick。Microsoft COCO：上下文中的常见物体。载于计算机视觉—*ECCV 2014*：第 *13* 届欧洲会议，苏黎世，瑞士，*2014* 年 *9* 月 *6* 日至 *12* 日，论文集，第五部分 *13*，第 740–755 页。施普林格，2014 年。

Ziyi Lin、Chris Liu、Renrui Zhang、Peng Gao、Longtian Qiu、Han Xiao、Han Qiu、Chen Lin、Wenqi S hao、Keqin Chen、Jiaming Han、Siyuan Huang、Yichi Zhang、Xuming He、Hongsheng Li 和 Yu Jiao Qiao。Sphinx：面向多模态大语言模型的权重、任务与视觉嵌入的联合混合。*arXiv*，abs/2311.07575，2023 年。

Haotian Liu、Chunyuan Li、Yuheng Li 和 Yong Jae Lee。通过视觉指令微调改进基线。*arXiv*，abs/2310.03744，2023a。

Haotian Liu、Chunyuan Li、Qingyang Wu 和 Yong Jae Lee。视觉指令微调。*arXiv* 预印本 *arXiv:2304.08485*，2023b。

Kenneth Marino、Mohammad Rastegari、Ali Farhadi 和 Roozbeh Mottaghi。OK-VQA：一个需要外部知识的视觉问答基准。*2019*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第3190–3199页，2019年。

Minesh Mathew、Dimosthenis Karatzas 和 CV Jawahar。DocVQA：一个用于文档图像视觉问答的数据集。在*IEEE/CVF*冬季计算机视觉应用会议论文集中，第2200–2209页，2021年。

Reiichiro Nakano、Jacob Hilton、Suchir Balaji、Jeff Wu、Long Ouyang、Christina Kim、Christopher Hesse、Shantanu Jain、Vineet Kosaraju、William Saunders 等。WebGPT：基于人类反馈的浏览器辅助问答。*arXiv* 预印本 *arXiv:2112.09332*，2021年。Sha Ning、Longtian Qiu、Yongfei Liu 和 Xuming He。HoiCLIP：利用视觉语言模型进行高效的人-物交互检测知识迁移。*2023 IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第 23507–23517 页，2023年。David Nukrai、Ron Mokady 和 Amir Globerson。使用注入噪声的 CLIP 进行纯文本图像描述训练。*ArXiv*，abs/2211.00575，2022年。OpenAI。Vision - OpenAI API。 `https://platform.openai.com/docs/guides/vision`，2023年。

Yassine Ouali、Adrian Bulat、Brais Martinez 和 Georgios Tzimiropoulos。CLIP-DPO：将视觉语言模型作为偏好来源，以修正大型视觉语言模型中的幻觉。*arXiv* 预印本 *arXiv:2408.10433*，2024年。

Long Ouyang、Jeffrey Wu、Xu Jiang、Diogo Almeida、Carroll Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray 等。通过人类反馈训练语言模型遵循指令。神经信息处理系统进展，35:27730–27744，2022年。William S. Peebles 和 Saining Xie。使用 Transformer 构建可扩展的扩散模型。*2023 IEEE/CVF* 国际计算机视觉会议（*ICCV*），第 4172–4182 页，2022年。Renjie Pi、Tianyang Han、Wei Xiong、Jipeng Zhang、Runtao Liu、Rui Pan 和 Tong Zhang。通过自举偏好优化增强多模态大语言模型。*arXiv*，abs/2403.08730，2024年。URL `https://api.semanticscholar.org/CorpusID:268379605`。

LongtianQiu、Shan Ning 和 Xuming He。通过纯文本训练挖掘细粒度图文对齐，以实现零样本图像描述。*arXiv*，abs/2401.02347，2024年。

Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移视觉模型。载于国际机器学习大会，第 8748–8763 页。机器学习研究会论文集，2021年。Rafael Rafailov、Archit Sharma、Eric Mitchell、Christopher D Manning、Stefano Ermon 和 Chelsea Finn。直接偏好优化：你的语言模型其实是一个奖励模型。神经信息处理系统进展，36，2024年。Anna Rohrbach、Lisa Anne Hendricks、Kaylee Burns、Trevor Darrell 和 Kate Saenko。图像描述中的物体幻觉。载于自然语言处理实证方法会议，2018年。URL https://api.semanticscholar.org/CorpusID:52176506。Pritam Sarkar、Sayna Ebrahimi、Ali Etemad、Ahmad Beirami、Sercan Ö Arık 和 Tomas Pfister。通过数据增强对比调优缓解物体幻觉。*arXiv* 预印本 *arXiv:2405.18654*，2024年。Jérémy Scheurer、Jon Ander Campos、Tomasz Korbak、Jun Shern Chan、Angelica Chen、Kyunghyun Cho 和 Ethan Perez。大规模使用语言反馈训练语言模型。*arXiv* 预印本 *arXiv:2303.16755*，2023年。John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Radford 和 Oleg Klimov。近端策略优化算法。*arXiv* 预印本 *arXiv:1707.06347*，2017年。Nisan Stiennon、Long Ouyang、Jeffrey Wu、Daniel Ziegler、Ryan Lowe、Chelsea Voss、Alec Radford、Dario Amodei 和 Paul F Christiano。通过人类反馈学习生成摘要。神经信息处理系统进展，33：3008–3021，2020年。Quan Sun、Yuxin Fang、Ledell Yu Wu、Xinlong Wang 和 Yue Cao。EVA-CLIP：大规模改进 CLIP 的训练技术。*arXiv*，abs/2303.15389，2023a。Zhiqing Sun、Sheng Shen、Shengcao Cao、Haotian Liu、Chunyuan Li、Yikang Shen、Chuang Gan、Liang-Yan Gui、Yu-Xiong Wang、Yiming Yang 等。通过事实增强型基于人类反馈的强化学习对齐大型多模态模型。*arXiv* 预印本 *arXiv:2309.14525*，2023b。Fei Wang、Wenxuan Zhou、James Y. Huang、Nan Xu、Sheng Zhang、Hoifung Poon 和 Muhao Chen。mDPO：面向多模态大语言模型的条件偏好优化。*arXiv*，abs/2406.11839，2024a。URL https://api.semanticscholar.org/CorpusID:270560448。Junyang Wang、Yuhang Wang、Guohai Xu、Jing Zhang、Yukai Gu、Haitao Jia、Ming Yan、Ji Zhang 和 Jitao Sang。一种无需大语言模型的多维基准，用于评估多模态大语言模型的幻觉。*arXiv* 预印本 *arXiv:2311.07397*，2023a。Weihan Wang、Qingsong Lv、Wenmeng Yu、Wenyi Hong、Ji Qi、Yan Wang、Junhui Ji、Zhuoyi Yang、Lei Zhao、Xixuan Song、Jiazheng Xu、Bin Xu、Juanzi Li、Yuxiao Dong、Ming Ding 和 Jie Tang。CogVLM：面向预训练语言模型的视觉专家。*arXiv*，abs/2311.03079，2023b。URL https://api.semanticscholar.org/CorpusID:265034288。Zhaoyang Wang、Weilei He、Zhiyuan Liang、Xuchao Zhang、Chetan Bansal、Ying Wei、Weitong Zhang 和 Huaxiu Yao。CREAM：一致性正则化自奖励语言模型。*arXiv*，abs/2410.12735，2024b。URL https://api.semanticscholar.org/CorpusID:273375180。Jeff Wu、Long Ouyang、Daniel M Ziegler、Nisan Stiennon、Ryan Lowe、Jan Leike 和 Paul Christiano。通过人类反馈递归地总结书籍。*arXiv* 预印本 *arXiv:2109.10862*，2021年。Junkang Wu、Yuexiang Xie、Zhengyi Yang、Jiancan Wu、Jinyang Gao、Bolin Ding、Xiang Wang 和 Xi-angnan He。 β-DPO：采用动态 β的直接偏好优化。神经信息处理系统进展，37：129944–129966，2024年。

Wei Xiong、Hanze Dong、Chenlu Ye、Han Zhong、Nan Jiang 和 Tong Zhang。通过人类反馈进行吉布斯采样：一种可证明的基于人类反馈的强化学习 KL 约束框架。*arXiv* 预印本 *arXiv:2312.11456*，2023年。

Zhihe Yang、Xufang Luo、Dongqi Han、Yunjian Xu 和 Dongsheng Li。通过直接偏好优化减少大型视觉语言模型中的幻觉：同策略数据是关键。*2025*年 *IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第 10610–10620 页，2025 年。

Lewei Yao、Jianhua Han、Youpeng Wen、Xiaodan Liang、Dan Xu、W. Zhang、Zhenguo Li、Chunjing Xu 和 Hang Xu。DetCLIP：用于开放世界检测的词典增强视觉概念并行预训练。*arXiv*，abs/2209.09407，2022 年。

Qinghao Ye、Haiyang Xu、Jiabo Ye、Mingshi Yan、Anwen Hu、Haowei Liu、Qi Qian、Ji Zhang、Fei Huang 和 Jingren Zhou。mPLUG-Owl2：通过模态协作革新多模态大语言模型。*2024*年 *IEEE/CVF* 计算机视觉与模式识别会议（*CVPR*），第13040–13051页，2023年。

Tianyu Yu、Yuan Yao、Haoye Zhang、Taiwen He、Yifeng Han、Ganqu Cui、Jinyi Hu、Zhiyuan Liu、Hai-Tao Zheng、Maosong Sun 等。RLHF-V：通过细粒度纠正性人类反馈实现行为对齐，构建可信赖的多模态大语言模型。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第13807–13816页，2024a。

Tianyu Yu、Haoye Zhang、Yuan Yao、Yunkai Dang、Da Chen、Xiaoman Lu、Ganqu Cui、Taiwen He、Zhiyuan Liu、Tat-Seng Chua 等。Rlaif-v：通过开源人工智能反馈提升多模态大语言模型的可信度，使其超越 GPT-4V。*arXiv* 预印本 *arXiv:2405.17220*，2024b。

Zhiyuan Zhao、Bin Wang、Linke Ouyang、Xiaoyi Dong、Jiaqi Wang 和 Conghui He。超越幻觉：通过幻觉感知直接偏好优化增强大型视觉语言模型。*arXiv* 预印本 *arXiv:2311.16839*，2023年。

Wenxuan Zhou、Ravi Agrawal、Shujian Zhang、Sathish Indurthi、Sanqiang Zhao、Kaiqiang Song、Silei Xu 和 Chenguang Zhu。Wpo：通过加权偏好优化增强基于人类反馈的强化学习。在自然语言处理实证方法会议，2024a。URL `https://api.semanticscholar.org/CorpusID`:270559089。

Wenxuan Zhou、Ravi Agrawal、Shujian Zhang、Sathish Reddy Indurthi、Sanqiang Zhao、Kaiqiang Song、Silei Xu 和 Chenguang Zhu。WPO：通过加权偏好优化增强基于人类反馈的强化学习。*arXiv* 预印本 *arXiv:2406.11827*，2024 b。

Yiyang Zhou、Chenhang Cui、Rafael Rafailov、Chelsea Finn 和 Huaxiu Yao。通过偏好微调对齐视觉大语言模型中的模态。$ArXiv$，abs/2402.11411，2024c。

Yiyang Zhou、Chenhang Cui、Rafael Rafailov、Chelsea Finn 和 Huaxiu Yao。通过偏好微调对齐视觉大语言模型中的模态。*arXiv* 预印本 *arXiv:2402.11411*，2024d。

Deyao Zhu、Jun Chen、Xiaoqian Shen、Xiang Li 和 Mohamed Elhoseiny。MiniGPT-4：利用先进的大语言模型增强视觉-语言理解。*arXiv* 预印本 *arXiv:2304.10592*，2023年。

Daniel M Ziegler、Nisan Stiennon、Jeffrey Wu、Tom B. Brown、Alec Radford、Dario Amodei、Paul Christiano 和 Geoffrey Irving。《基于人类偏好对语言模型进行微调》。*arXiv* 预印本 *arXiv:1909.08593*，2019年。

表7：**融合比例消融实验**。 公式 11 中，融合比例 表示$\hat{c}_g$ 和 $\hat{m}_g$ 的权重。蓝色行表示自适应分数融合的权重。

融合比例		AMB.^{G}				AMB.^{D}	Seed^I
CLIP	多模态大语言模型	C_s ↓	覆盖率 ↑	幻觉率 ↓	认知 ↓	F1 ↑	分数 ↑
20%	80%	4.9	57.7	30.9	**2.0**	85.4	63.2
40%	60%	**4.3**	57.5	28.3	**2.0**	85.6	64.0
53%	47%	**4.3**	57.4	**28.0**	2.1	**85.6**	**64.8**
60%	40%	4.5	57.6	28.4	2.1	85.3	63.8
80%	20%	4.6	**58.3**	28.6	**2.0**	85.5	62.8

表8：**采用不同 β 时 DA-DPO 和直接偏好优化的性能**。我们给出了在 BPO 数据集上训练的直接偏好优化和难度感知直接偏好优化的结果。我们用蓝色标出报告的超参数。对于 DA-DPO，我们选择了性能最佳的 β，其值为 0.2。对于直接偏好优化，我们遵循先前工作的报告，β等于 0.1。

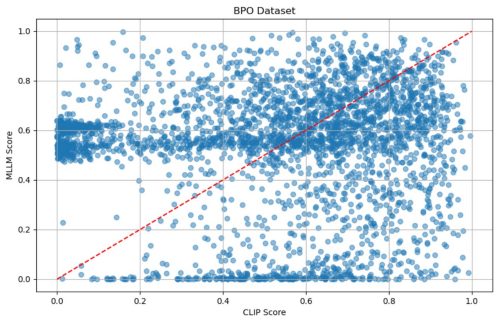
	AMB.^{G}				AMB.^{D}	ObjHal		POPE	GQA	Seed^I	MME^P	MME^C	
	C_s ↓	覆盖率 ↑	幻觉率 ↓	认知 ↓	F1 ↑	C_s ↓	C_i ↓	F1 ↑	分数 ↑	准确率 ↑	分数 ↑	分数 ↑	
DA-DPO													
$\beta = 0.1$	4.6	**59.0**	32.5	**2.0**	85.4	40.4	10.2	85.6	58.6	64.2	1398.2	312.8	
$\beta = 0.2$		57.4	28.0	2.1	85.6	39.7	9.9	**85.9**	**59.8**	**64.8**	1406.6	**323.2**	
$\beta = 0.3$	4.7	53.6	**25.8**	2.4	**86.3**	**38.6**	**9.5**	84.8	56.4	58.4	**1414.2**	315.6	
$\beta = 0.4$	4.9	54.2	28.4	2.6	86.0	41.6	10.4	85.2	58.4	62.2	1378.2	306.9	
DPO													
$\beta = 0.1$		58.4	35.7	**2.0**	83.9	**43.3**		84.3	45.3	**57.7**			
$\beta = 0.2$	**4.8**	**58.5**	30.0	**2.0**	83.4	44.2	10.2	83.8	58.5	47.9	1369.1	299.6	
$\beta = 0.3$	4.9	56.6	**29.4**	**2.0**	84.4	45.3	10.8	84.3	56.8	54.7	1387.1	310.7	
$\beta = 0.4$	4.9	56.5	30.8	**2.0**	**84.5**	45.3	12.1	**85.2**	**58.9**	55.4	1402.4	312.3	

A 自适应分数融合的影响

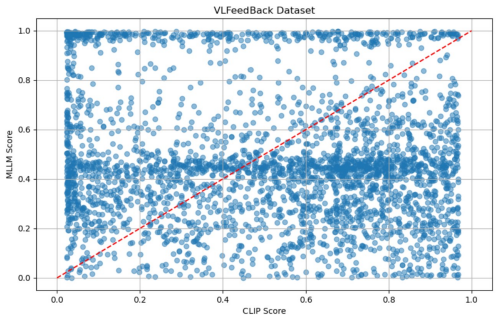
所提出的框架整合了两类预训练视觉语言模型，以评估成对偏好数据的难度。我们设计了一种自适应分数融合机制，用于确定每个奖励模型的权重，无需从数据角度进行超参数选择。如表7所示，我们展示了多种固定融合分数的结果。我们发现，自适应分数融合取得了最佳性能。

B β 的敏感性

我们提出了一种难度感知偏好优化策略，旨在缓解对齐训练期间对简单样本过拟合的问题。为此，我们动态校准了 β，具体说明见正文。然而，我们的方法会影响 DPO目标函数中 β 的尺度。如表8所示，我们对 DPO 与 DA-DPO 之间的 β 进行了消融实验。结果表明，对于 DPO 和 DA-DPO，改变β 会导致它们在不同基准测试上的表现各有优劣。尽管如此，与原始 DPO 相比，所提出的 DA-DPO 始终取得更好的总体表现，这说明我们的难度感知校准能有效提升鲁棒性，同时减少幻觉。



(a) BPO



(b) VLFeedBack

图5：**预训练视觉语言模型之间相关性的可视化**。我们展示了三个数据集 BPO 和 VLFeedBack 中同一偏好数据样本对应的归一化多模态大语言模型分数和 CLIP 分数。横轴为 CLIP 分数，纵轴为多模态大语言模型分数。红线表示多模态大语言模型和 CLIP 的预测结果相同。

表9：BPO 和 VLFeedback 数据集上融合分数的分类准确率。

数据集	奖励模型类型	VQA	图像描述	文本视觉问答	总体
BPO	CLIP	0.5877	0.8947	0.9827	**0.7732**
BPO	MLLM	0.8478	0.5053	0.6282	**0.6415**
VLFeedback	CLIP	0.5827	0.6352	–	**0.5897**
VLFeedback	MLLM	0.8104	0.5193	–	**0.7715**

C 预训练视觉语言模型相关性

在我们提出的难度感知直接偏好优化框架中，我们利用 CLIP 和多模态大语言模型从不同角度评估偏好数据的难度。为验证这一说法，我们给出了两个视觉语言模型之间的难度相关分数。如图5所示，两个模型给出的分数没有显著的正相关，其相关系数较低，便是明证。这表明两个模型捕捉了响应评估的不同方面，并且它们的评分模式在不同数据集上并不总是一致。

D 融合分数分类准确率

为全面理解分数融合过程，我们在正文基础上补充了详细的数值结果和评估指标。分数融合在数据集层面进行，其中 BPO 和 VLFeedback 数据集采用不同的融合分数。

评估指标。 我们根据以下标准，使用成对偏好数据计算分类准确率：对于每个样本，视觉语言模型会分别为被选答案和被拒答案分配分数。如果被选答案的分数高于被拒答案，则认为该样本分类正确。

结果。 表9报告了两个数据集上各类别（适用时包括视觉问答、图像描述和文本视觉问答）的分类准确率以及总体准确率。我们评估了两种类型的视觉语言模型：视觉编码器（CLIP）和多模态大语言模型（MLLM）。具体而言，我们使用 EVA-CLIP 8B 作为 CLIP 模型，使用 LLaVA-1.5 7B 作为 MLLM。

根据总体准确率，我们将基于 CLIP 的分类分数（`cls_c`）和基于多模态大语言模型的分类分数（`cls_m`）分别设定如下：对于 BPO 数据集，分别为 $\texttt{cls_c} = 0.7732$ 和 $\texttt{cls_m} = 0.6415$；对于 VLFeedback数据集，分别为 $\texttt{cls_c} = 0.5897$ 和 $\texttt{cls_m} = 0.7715$。这些结果表明，CLIP 和基于多模态大语言模型的奖励估计在不同类别上具有互补优势。

表 10：**奖励模型与代理指标的计算量与时间估计**。浮点运算次数以科学计数法表示，实际耗时小时数根据 NVIDIA A100 GPU 估算。奖励模型训练的数据设为 180k。

Task	浮点运算次数	实际耗时小时数（A100）
传统方法		
奖励模型训练	2.6×10^{19}	241.9
奖励模型估计	5.1×10^{18}	40
DA-DPO		
CLIP 估计	1.4×10^{16}	0.7
多模态大语言模型估计	5.1×10^{18}	40

表 11：**数据难度归一化的消融实验**。我们报告了在 BPO 数据集上使用 LLaVA-v1.5 7B 训练的不同归一化策略的性能。

	AMB.^{G}				AMB.^{D}	ObjHal		POPE	GQA	Seed^I	MME^P	MME^C	
	$C_s \downarrow$	覆盖率 ↑	幻觉率 ↓	认知 ↓	F1 ↑	$C_s \downarrow$	$C_i \downarrow$	F1 ↑	分数 ↑	准确率 ↑	分数 ↑	分数 ↑	
DPO	5.5	**58.4**	35.7	2.0	83.9	43.3	10.0	85.6	45.3	57.7	**1409.4**	315.0	
基于排名的方法	4.8	57.2	31.3	**1.9**	85.7	42.7	10.7	**86.1**	59.5	62.1	1404.8	302.5	
长度控制	4.7	57.0	29.3	**1.9**	**85.9**	**37.7**	**9.6**	85.9	**60.0**	64.1	1402.6	316.0	
高斯归一化 **4.3**		57.4	**28.0**	2.1	85.6	39.7		85.9	59.8	**64.8**	1406.6	**323.2**	

E 效率比较

为验证我们提出的框架在效率方面的优势，我们比较了传统奖励模型训练与评估和基于估计的方法在浮点运算次数和实际运行时间方面的计算成本。如表 10 所示，训练一个完整的奖励模型所需的实际运行时间超过基于多模态大语言模型的估计的 6 倍，约为基于 CLIP 的估计的 345 倍。这些结果清楚地表明，与传统方法相比，我们的方法效率显著更高。

F 难度估计归一化消融实验

为了评估不同归一化策略对数据难度估计的影响，我们采用两种替代方法进行了实验。第一种策略 基于排名的方法，通过对数据进行排名并将数值线性映射到 $[0, 1]$ 范围内来归一化分数，而不是使用高斯归一化。第二种策略 长度控制，应用于多模态大语言模型生成式估计过程中：将公式 7 和 8 中对数概率的总和除以响应词元长度，以减轻长度偏差。如表 11 所示，三种归一化策略的表现均优于原始 DPO。我们还发现，基于排名的方法略逊于另外两种策略，这表明保留估计分数的原始分布有助于有效进行难度估计。

G 更多模型变体的实验

为验证所提出方法在 LLaVA 系列之外的有效性，我们在 Qwen2.5-VL 3B 上进行了一项小规模实验。我们使用 BPO 数据集中的 5 万个样本子集，分别训练基线方法直接偏好优化和所提出的 DA-DPO。结果如表 12所示。我们发现，与直接偏好优化相比，DA-DPO 能更有效地降低幻觉指标，同时保持在综合基准上的表现。这表明 DA-DPO 是一种通用框架，并不局限于 LLaVA 系列模型。

表 12：**各方法在幻觉和通用基准测试上的比较**。结果报告基于 Qwen2.5-VL3B，以及使用直接偏好优化和 DA-DPO 训练的模型。所提出的 DA-DPO 持续降低幻觉指标，同时保持或提升通用能力。

	AMB.^{G}			AMB.^{D}	ObjHal	POPE		GQA	Seed^I	MME^P	MME^C	
	$C_s \downarrow$	覆盖率 ↑	幻觉率 ↓	认知 ↓	F1 ↑	$C_s \downarrow$	$C_i \downarrow$	F1 ↑	分数 ↑	准确率 ↑	分数 ↑	分数 ↑
Qwen2.5-VL 3B	7.5	**68.0**	48.8	5.0	**89.6**	30.7	7.6	**88.6**	**62.5**	75.9	**1607.9**	**625.7**
DPO	6.6	62.5	47.7	3.4	87.3	30.0	8.4	78.5	55.5	75.9	1558.5	616.4
DA-DPO	**6.5**	57.6	**42.3**	**2.6**	88.3	**25.0**	'	80.6	56.4	**76.0**	1561.5	623.9

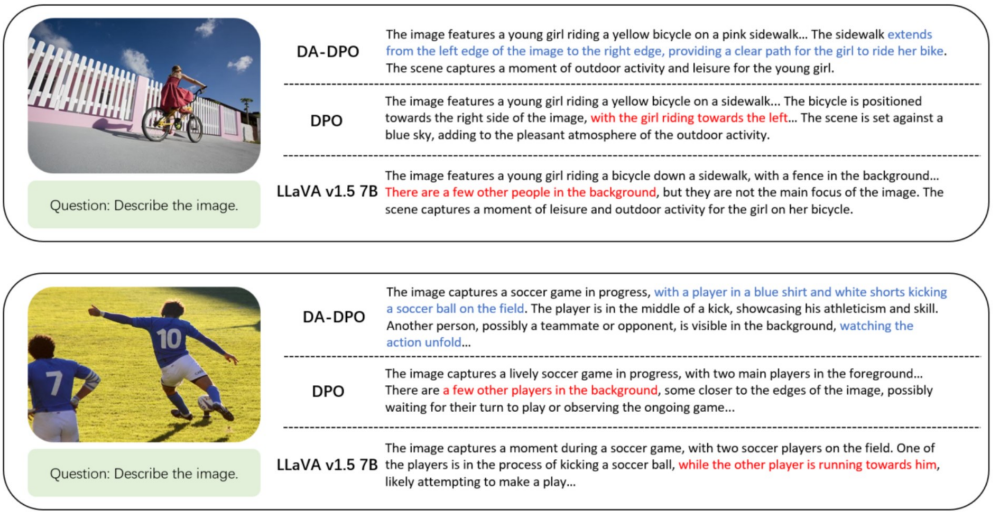


图6：**预测结果可视化**。我们展示了所提方法与基线模型的输出，以突出我们方法的特点。结果来自 AMBER 的生成组件，该组件旨在检测图像描述中的幻觉。

可视化 为了更好地了解我们方法的性能，我们在图6中展示了模型输出的可视化结果，并将其与直接偏好优化和参考模型的结果进行比较。这些可视化结果凸显了我们的方法在更有效地减少幻觉方面的能力。通过考察这些输出可以看出，我们的方法与预期响应更加一致，在场景理解准确率和生成内容的连贯性方面表现更佳。