arXiv:2611.32316v1 [cs.CV] 24 Nov 2026

ASTERISM：多模态大语言模型真的会查看所有图像吗?

Chanseul Cho Dongwoo Kim NooriBae SeungjaeMoon Seongsu Lee YoungminRo†机器智能实验室,
首尔市立大学{chanseul2001, shanghai110930}@gmail.com{2bee5, msj0243, ssulee0206, youngmin.ro}
@uos.ac.kr

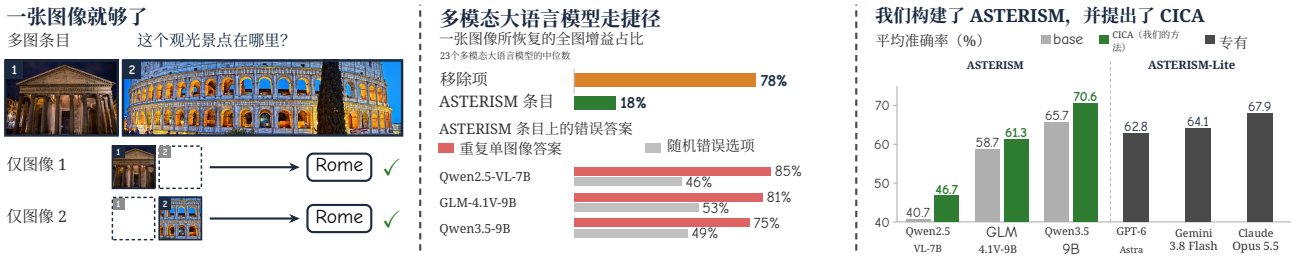


图1. **许多多图像条目会奖励多模态大语言模型采用单图捷径，而 CICA 会将功劳转移给那些需要多张图像的词元。**多图像基准中的这个条目仅凭任意一张照片就能作答（左）。将其他图像置灰后，在审计移除的条目上，一张图像就能恢复大部分全图增益（即所有图像相较于仅文本所带来的准确率提升）；但在 ASTERISM 条目上，这种情况很少见，而且 ASTERISM 条目中的大多数错误答案都会重复单图像答案（中）。CICA 提高了每个基础模型的 ASTERISM 得分；在困难子集 ASTERISM-Lite 上，得分最高的专有多模态大语言模型达到 68%（右）。

摘要

多图像基准旨在测试多模态大语言模型（*MLLM*）是否会综合不同图像中的证据，但其中许多条目仅凭一张图像就能作答。在来自 *19* 个公开基准测试的约 *3.3* 万个条目中，我们的审计和两阶段单图过滤器发现，其中 *24.2%* 的条目仅凭一张图像或问题本身就能确定答案。对于这些条目，多模态大语言模型仅凭一张图像就能获得相对于仅依据文本作答时、所有图像所带来准确率提升的中位数 *78%*；因此，这些条目会奖励仅凭一张图像作答的行为，我们称之为单图捷径。即使在单张图像不足以作答的条目上，多模态大语言模型也会采取这种捷径。在这类条目中，它们的大多数错误答案都重复了仅凭一张图像时给出的答案，重复率远高于随机机会。我们从这 *19* 个基准测试中保留那些需要多张图像才能作答的条目，构建了 *ASTERISM*；据我们所知，这是经过审计的最大多图像基准。其 *1.87* 万个条目（其中 *2,000* 个经过修正或重写）采用统一格式、统一评分器，并涵盖七类任务。我们还发布了 *ASTERISM-Train*，其中包含 *4.9* 万个训练条目，以及 *ASTERISM-Lite*——一个困难子集，其中每个条目都经过人工逐一核验。在 *ASTERISM-Lite* 上，表现最强的专有多模态大语言模型得分为 *68%*。如今，可验证奖励强化学习（*RLVR*）已广泛应用于

多模态大语言模型的后训练中，奖励机制同样会奖励这种捷径，因为从一张图像中读出正确答案，与综合多张图像得出答案获得的奖励一样多。与此一致，现有的 *RLVR* 目标对 *ASTERISM* 得分提升甚微。为了在训练中不再奖励这种捷径，正如 *ASTERISM* 在评估中所做的那样，

中，我们提出跨图像信用分配（CICA），这是可验证奖励强化学习的一种扩展，能够将信用分配转移到那些需要多张图像才能处理的词元上。在三个开源多模态大语言模型上，CICA 使 *ASTERISM* 得分提高了 *2.6* 至 *6.0* 分，增幅是使用相同数据训练的通用强化学习目标和视觉专用强化学习目标各自增益的两倍以上。我们的数据集、代码和模型将公开发布于 https://github.com/chanseul01/ASTERISM。

1. 引言

人们通过联系世界的多个视角来理解世界。每张照片只展示一台摄像机在某个瞬间所看到的内容，因此我们会从多个侧面观察一个房间，以了解其布局；比较相隔数月拍摄的患者扫描图像，以查看疾病如何发展；翻阅一份长篇报告，以找到能够回答问题的页面。在这些情况下，单张图像都不够用。多模态大语言模型（MLLMs）如今可以将交错图像作为此类任务的输入 [7, 49, 67], , 然而能够很好理解单张

†通讯作者。

图像的 MLLM 仍难以关联跨图像的线索 [18, 22,40, 41, 121]。多图像基准 [19, 49, 52, 68, 70, 76, 116, 125, 129, 151, 155, 162, 167, 168] 用于衡量这项能力，并指导 MLLM 的构建与训练方式 [18]。然而，只有在回答多图条目时必须使用其中不止一张图像，将这些图像中的证据结合起来，或从中找到正确的图像，该条目才真正检验了这项能力。我们要 выяс定有多少基准测试条目满足这一条件，以及 MLLM 是否会根据不止一张图像作答。

许多基准测试条目不符合这一条件，而即使在符合条件的条目中，多模态大语言模型也会依赖一张图像。我们将根据多图条目中的一张图像作答称为单图捷径，这是一种捷径学习 [32]。它是语言先验在多图像场景中的对应形式：在语言先验下，视觉问答模型仅凭问题作答 [2, 37]，而多模态大语言模型则会幻觉出文本所暗示的内容 [38, 63]。与语言先验不同，它利用视觉证据，因此移除所有图像 [17, 122]并不能揭示这种捷径，而逐张展示图像则可以。在 19 个公开基准的 32,688 项中，我们通过两个多模态大语言模型、人工裁定者和两阶段单图过滤器进行审计，发现其中 24.2% 的条目仅凭一张图像或问题本身就能作答；在八个基准中，超过四分之一的条目也是如此，其中包括 MuirBench、Q-Bench2 和 MIRB（第2 节和表 S2）。[1] 这些条目会奖励走捷径的行为。在仅显示一张图像、其余图像灰显的情况下，对于审计中移除的条目，23 个开源多模态大语言模型能够获得其中位数为 78% 的全图增益（即所有图像相较于仅有文本所带来的准确率提升）；而对于审计中保留的条目，这一比例为 18%（第2.2节和图1）。

即使面对需要多张图像的项目，多模态大语言模型也会采取捷径。在此类项目中，Qwen2.5-VL-7B、Qwen3.5-9B 和 GLM-4.1V- 9B-Thinking 的错误答案中，有 75–85% 会重复模型仅根据单张图像给出的答案，而随机机会下这一比例约为一半（图 1）。

在这 19 个基准测试中，我们保留需要多张图像的条目，并以此构建 ASTERISM（第 3节）。这些条目包括视觉感知、多模态推理、检索增强问答和文档基准测试中的多图像条目 [31, 42, 44, 60, 87]。除了利用捷径的条目外，审计还发现了标签或评分错误、含糊不清或无法回答的问题，以及重复项。总计有 32.8% 的条目并未按其设计考查多图推理。ASTERISM 修正了 569 个条目的标签、答案格式或文本，并新增了 1,406 个改写条目，其中每个选项都将来自不同图像的证据配对；人工审核员会检查每个修正后的标签和每次改写。各来源的格式、提示和评分器各不相同，对同一技能的称呼也不一致，因此 ASTERISM 将所有条目统一为一种格式、一个评分器和七类任务（第 3.2节）。采用可验证奖励的强化学习（RL）（RLVR）用于后训练时，也会奖励利用捷径的行为。通过组相对策略优化（GRPO） [114] 和 DAPO 等算法 [157], , RLVR 提升了大语言模型在数学和代码方面的表现 [39]，也提升了多模态大语言模型在视觉感知和推理方面的表现 [46, 83]。近期的多模态大语言模型在后训练中采用了 RLVR [11, 43, 104, 147]。它的奖励只检查最终答案，并且响应中的每个词元都获得相同的优势，因此更新过程无法指出哪些词元赢得了奖励。对于多图条目，只从一张图像中读出正确答案，与整合多张图像得出答案获得的功劳一样多；因此，奖励不会推动已经采用捷径的模型转而整合多张图像（第 2.2节）。与此一致，现有的 RLVR目标对多图准确率的提升微乎其微。22 个公开的 Qwen2.5-VL-7B 后训练变体中，大多数使用强化学习进行训练，它们在 ASTERISM 上的提升最多为 1.6 个百分点（表 S9和 S10）；此外，我们训练的任何骨干网络采用五种 RLVR目标（两种通用目标和三种专为视觉输入设计的目标），效果提升都很小（表 4）。

因此，许多多图像项目和可验证奖励强化学习都会奖励这种捷径。ASTERISM 会从评估中移除奖励该捷径的项目；在训练中，跨图像信用分配（CICA）则将可验证奖励强化学习的信用转移给那些需要多张图像的词元。审计会针对每个项目询问：单张图像是否已经足以给出答案；CICA 则会针对响应中的每个词元提出同样的问题。CICA 比较策略在所有图像下对该词元的自身似然，以及在最佳单图像下的似然；最佳单图像是指使该词元最有可能出现的那一张图像。CICA 保留可验证奖励强化学习及其可验证奖励，仅改变该奖励如何分配给词元（第 4节）。其他可验证奖励强化学习方法会根据词元的视觉激活、视觉定位或对关键视频帧的注意力来调整词元权重或进行屏蔽 [20, 24, 138]，或者在缺少关键视觉线索时推动策略偏离其预测 [158]。多图像强化学习方法会加入图像对比、拼图或跨图视觉定位等任务 [18, 143, 169]（相关工作见补充材料 A）。CICA 不增加任务、参数或采样轨迹。在依赖图像的词元中，它会提高那些无法由最佳单图像解释的词元的权重。

我们提出了三项贡献。

•我们发现，许多多图像条目会让单图捷径带来奖励，而多模态大语言模型也会采用这一捷径。仅凭一张图像或问题本身，就足以确定 19 个公开基准中的 24.2% 的条目；即使对于需要多张图像才能作答的条目，31 个开源多模态大语言模型中的每一个，其错误答案重复单图像答案的频率也高于匹配随机基准水平（第2节）。

• 我们发布了 ASTERISM，据我们所知，这是规模最大的经审查多图像基准测试，其中包含 18.7k 个需要不止一张图像的条目，其中 2.0k 个经过了修正或重写；我们还在该基准测试上评估了 58 个开源多模态大语言模型（见第3节和第5.1节）。ASTERISM-Lite 是由 400 个较难的 ASTERISM 条目组成的子集，我们对这些条目逐一进行了人工验证。在 ASTERISM-Lite 上，最强的专有多模态大语言模型得分为

[1]编号带有 S 的图形、表格和页面位于补充材料中。

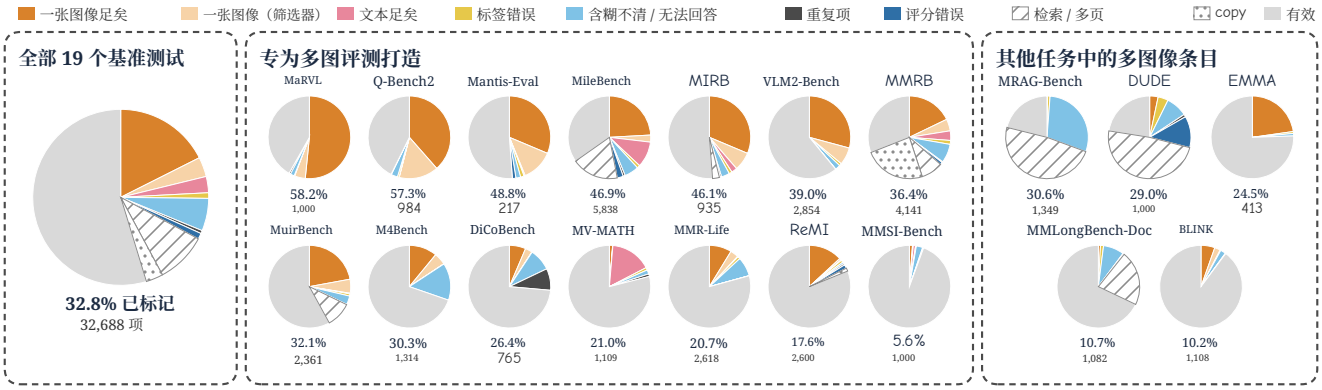


图2. **仅凭一张图像或问题本身，就能回答许多多图像条目。** 各源基准中条目的审计结果和单图筛选器判定。每张图表下方的百分比表示标记项所占的比例，即仅凭一张图像或问题就能回答的条目，或存在缺陷的条目。斜线部分表示仅凭一张图像或问题就能回答的检索条目和多页文档条目，点状部分表示其他来源条目的副本。即使移除了这两类条目，计算该百分比时也不会将其计入。

68%，而开源多模态大语言模型的排名与其在 ASTERISM 上的排名高度一致（见第3.3 节和表3及S4）。我们还发布了 ASTERISM-Train，其中包含基于公共数据集和图像构建的 49k 个带有可验证答案的训练条目，并移除了 ASTERISM 图像的近重复项（见第3.4节）。

- 我们提出了 CICA，这是可验证奖励强化学习（RLVR）的扩展，它将功劳从一张图像就能解释的响应词元转移到需要多张图像才能解释的词元。我们展示了它能使三个骨干模型的得分提高 2.6 至 6.0 分，超过了在相同数据上训练的五种强化学习目标；而且，与它们相比，当所有图像都可见时，CICA 带来的提升比仅显示一张图像时更大。尽管它只使用多图像条目进行训练，却能在十个单图像基准测试上保持或提高骨干模型的得分（第4 节和5节）。

2. 许多多图像条目奖励单图捷径

2.1. 一张图像或问题本身就能回答五分之一以上的条目

在 19 个公开基准的 32,688 项中，一张图像足以解决 21.2%，仅凭问题又可解决另外 3.0%；这些统计计入了审计和单图筛选器的两个阶段（补充材料 B.1和B.2），因此这些条目奖励单图或纯文本作答。这些基准大多是为多图像评估而构建的，其余则在其他任务中包含多图像条目，例如检索增强回答或文档理解（图 2）。

审计前，筛选器的第一阶段会移除这样 545 个项目：来自四个模型系列的四个开源多模态大语言模型 [1, 23, 56, 144] 在选项的每一种排列下，都能根据项目中的某一张图像答对 [80]（补充材料B.1）。来自不同模型系列的两个多模态大语言模型 Claude Fable 5 [5] 和 GPT-5.6-Sol [99]，分别独立审计剩余的每个项目。每个多模态大语言模型回答两次：第一次仅依据问题和选项，随后结合带编号的图像，并将两次答案分别与真实标注进行比较。随后，它会指出（如果存在）能够单独提供足够信息的那张图像，并按照固定优先级分配八种判定结果中的一种（补充材料 B.1）。人工裁定者负责解决分歧，人工审核员会检查每个标签修正以及部分分层样本中的判定结果（第3.1节）。

这些条目在不同来源中的占比各不相同（表 S2）。仅凭一张图像或问题本身，就能解决 MaRVL 中 55.8% 的条目（其陈述需依据一对图像进行核验），以及 MuirBench、Q-Bench2 和 MIRB 中 28–54% 的条目；而在 MV-MATH 的多图数学题中，仅凭问题文本就能解决 16.1%。对于检索问题和多页文档问题，找到正确图像或页面本身就是任务，因此即便审计将其移除，也不计入上述占比（图 2中的斜线阴影部分；规则见补充材料 B.4），所以这些占比是保守估计。除这些捷径条目外，审计还发现了一些含糊不清或无法回答的条目（6.0%）、带有标签或评分错误的条目（2.1%），以及与其他条目重复的条目（0.5%）。总计而言，32.8% 的条目并未如其设定那样测试多图推理。

2.2. 多模态大语言模型采用单图捷径

我们通过向多模态大语言模型隐藏图像，来检验它们是否会走这条捷径。我们保持提示、图像的数量、排序和尺寸不变，并将隐藏的图像置为均匀灰色，因此输入保留了条目的布局，只失去了隐藏图像的内容。我们在三种条件下测试每个模型。在仅文本条件下，隐藏所有图像；在一张图像条件下，仅显示一张图像，并对所选图像求平均准确率；在所有图像条件下，不隐藏任何图像。全图增益是所有图像相较于仅文本所带来的准确率提升，而恢复比率

$$\rho = \frac{\mathrm{Acc}_{\mathrm{one}} - \mathrm{Acc}_{\mathrm{text}}}{\mathrm{Acc}_{\mathrm{all}} - \mathrm{Acc}_{\mathrm{text}}} \tag{1}$$

是单张图像所恢复的全图增益占比。接近 1 的比值意味着一张图像带来了约

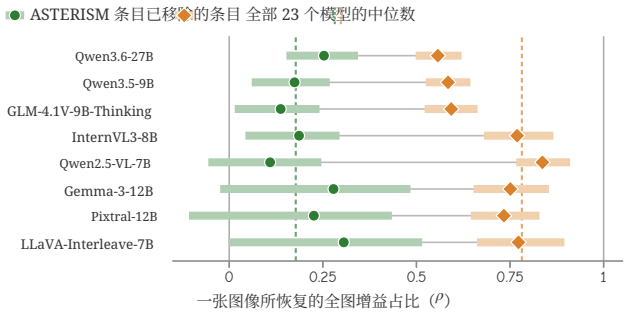


图3. **在审计移除的条目上，一张图像可恢复大部分全图增益；而在 ASTERISM 条目上，恢复效果很小。**恢复比率 ρ 针对第 5.2 节中训练的三个骨干模型，在多项选择 ASTERISM 条目和已移除条目上的表现，以及各系列中 23 个模型里准确率最高的模型。色带表示 95% 置信区间，虚线标出全部 23 个模型的中位数。

所有图像准确率的效果，而接近 0 的比值意味着相较于仅使用文本，一张图像几乎没有带来额外收益。我们测试了 Qwen、InternVL、LLaVA、Gemma、Pixtral 和 GLM 系列的 23 个开源多模态大语言模型，其中包括针对多图像任务训练的模型。测试使用了 1,000 个 ASTERISM 项目和 1,000 个已移除项目中的多项选择题；这些项目因一张图像或问题本身即可回答而被审计移除。每个项目包含两到六张图像。

在已移除项上，一张图像即可恢复大部分全图增益。 在 23 个模型中，中位数 ρ 在已移除项上为 0.78，在 ASTERISM 条目上为 0.18，差值的 95% 区间为 [0.49, 0.74]（按条目和模型计算），且每个模型在已移除项上的恢复幅度都更大（图 3和 S18）。将其他条目的图像替换灰色图像后，这一差距依然存在：已移除项的中位数比率为 0.65，ASTERISM 条目的为 0.14，而且每个模型在已移除项上的恢复幅度仍然更大（表 S13）。因此，已移除项会奖励这种捷径，而 ASTERISM 条目对其奖励很少。

错误答案遵循一张图像。 即使对于需要不止一张图像的项目，大多数错误答案也会重复模型根据一张图像给出的答案。对于本测试中的每个多项选择 ASTERISM 项目，我们在每次仅显示一张图像时收集模型的答案，即其单图像答案，并检查在显示所有图像时给出的错误答案是否重复其中某个答案。对于 Qwen2.5-VL-7B、Qwen3.5-9B 和 GLM-4.1V-9B-Thinking，75–85% 的错误答案会重复一个单图答案，而随机抽取错误选项时这一比例为 46–53%（表 1）。这一随机基准水平是对各项目 $|S_i \cap W_i|/|W_i|$的平均值，其中 S_i 为单图答案集合， W_i 为项目 i的错误选项。在有两个选项的项目中，按照构造方式，这两个比例相同，从而缩小了这一差距；而在有三个或更多选项的项目中，差距扩大到 30.0–41.0 个百分点（补充材料 H.2）。对于每个模型，在那些不同于纯文本答案的错误答案中，这一差距仍然存在，其中该答案被排除在 W_i之外。

表1. **即使在 ASTERISM 条目上，错误答案也会沿用单图像答案。**在与图 3 相同的多项选择题上，匹配度表示每个模型错误选项中重复单图像答案的比例；随机机会（灰色）则表示随机抽取错误选项时的相同比例。右侧部分排除了与纯文本答案相同的错误答案，而 Δ 表示此部分中的匹配度减去随机机会。

	所有错误答案	不含纯文本答案	
项目	**匹配** / 随机机会	**匹配** / 随机机会	Δ
Qwen2.5-VL-7B			
ASTERISM	**84.6** / 46.2	**68.7** / 46.0	**+22.7**
已移除	**81.5** / 47.6	**77.6** / 49.0	**+28.6**
GLM-4.1V-9B-Thinking			
ASTERISM	**81.1** / 53.4	**69.3** / 51.7	**+17.6**
已移除	**80.2** / 51.1	**73.8** / 49.4	**+24.4**
Qwen3.5-9B			
ASTERISM	**75.4** / 48.8	**65.7** / 48.3	**+17.4**
已移除	**81.0** / 51.2	**73.6** / 44.5	**+29.1**

第二个匹配随机基准水平使用另一个项目的单图答案，该项目具有相同的正确字母以及相同数量的选项和图像。这三个模型的错误答案在 47–53% 的情况下与这些答案匹配，而在此次比较中，它们自己的单图答案所对应的比例为 75–84%。对于 31 个开源多模态大语言模型，错误答案同样更频繁地重复模型自身的单图答案，而不是这一匹配水平，并且差异的 95% 自助法区间位于零以上（表 S14）。

因此，许多多图像条目都会奖励多模态大语言模型采用的单图像捷径。我们因此基于需要一张以上图像的条目构建 ASTERISM（第 3 节），并提出 CICA，通过将可验证奖励强化学习的功劳转向那些无法由一张图像解释的响应词元（第4 节）。

3. ASTERISM：一个包含需要多张图像的条目的基准测试

ASTERISM从19个来源中各自仅保留需要多张图像的条目，使用同一个评分器为每个条目评分，并将这些条目划分为七类任务。所有条目均采用统一格式：按顺序排列图像，提示遵循来源的官方指示（若有），并注明答案格式，因此一次运行即可评估全部19个来源，并报告每个任务类别的得分。在选项数量相同的条目中，每个字母作为正确答案的频率大致相同。图 4 展示了这些条目从来源到基准测试的筛选过程，补充材料B.2 列出了各来源的计数，表 S1则将 ASTERISM 的条目核查与其他多图像基准的条目核查进行了比较。

3.1. 筛选、改写与人工审核

审计判定决定如何处理每个条目。有效条目（审计未发现缺陷）将保留其问题、图像和答案。标签错误的条目将更正答案，评分错误的条目将采用接受所要求形式的答案格式。可通过文本回答、单张图像足以作答、含糊不清、

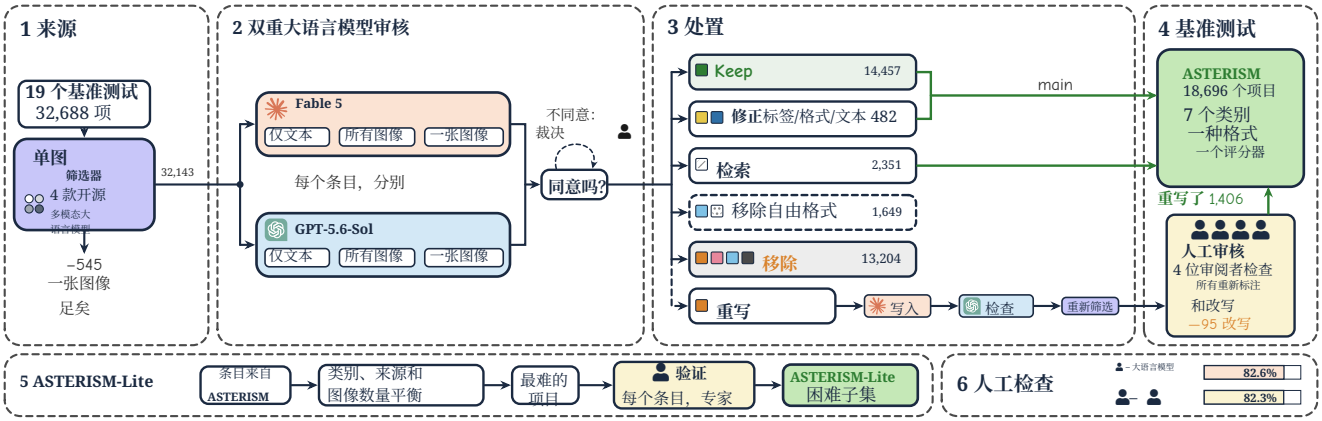


图4。 **构建 ASTERISM 和 ASTERISM-Lite**。 每个审计员框中的三个标签分别表示其两种回答（仅依据文本，以及结合所有图像）和足以作答的单张图像（如有）。每个经审计的条目只计数一次，并根据其最终处置归入从保留到移除的相应类别；87 个更正后的检索项计入检索。保留、修复和检索共包含 ASTERISM 的 17,290 个源条目，而 1,406 个改写条目则是新增条目。移除自由格式还包括其他来源条目的副本、模糊题项，以及其改写版本所取代的源条目（补充材料 B.2）。人物图标表示人工处理步骤。色块颜色遵循图 2；绿色表示保留。

无法回答和重复的条目将被移除。只有当图像支持一个可验证的答案，且转换后的条目通过独立审核时，MileBench 的自由形式条目才会作为多项选择题纳入（补充材料B.2）。

部分条目（主要是带旗帜标记的条目）会被重写，使其图像集仍可用作多图像测试（图S6）。Claude Fable 5 会基于相同图像提出新问题和选项，使每个选项都结合来自不同图像的证据；GPT-5.6-Sol 则检查是否恰有一个选项正确。重写后的条目保留其源条目的操作类别（补充材料 B.4）。这 1,406 个改写条目均通过单图筛选器和人工审核。

人工审核员会检查审计所纠正的每个标签以及每项改写是否有正确答案，并确认每项改写都需要多张图像。在看到标签之前，审核员能正确回答 94% 的改写，并判断其中 98 % 需要多张图像。未通过审核或自动检查的改写会被移除。两位审核员还按判定分层抽取了 1,000 个条目作为样本进行标注。他们的答案与审计判定有 82.6% 一致，彼此对 82.3% 的条目意见一致，因此审计与一位审核员的一致程度，和两位审核员彼此的一致程度相同（补充材料B.3）。

3.2. 七类任务与一个评分器

19 个来源对相同技能的称呼各不相同，因此每个条目会根据其答案所需的跨图像操作，或依据适用于某些子集的规则，归入七类任务之一（补充材料 B.4）。这些类别包括聚合（跨图像计数或概括）、比较/变化、推理/验证、空间对应、时间/动作、文本/网络和检索；检索要求找到包含证据的图像或页面（图 5）。条目还带有横跨这些类别的标签，例如 11 种跨图像

操作、图像之间的关系以及推理跳数（表 S11和图 S16）。ASTERISM

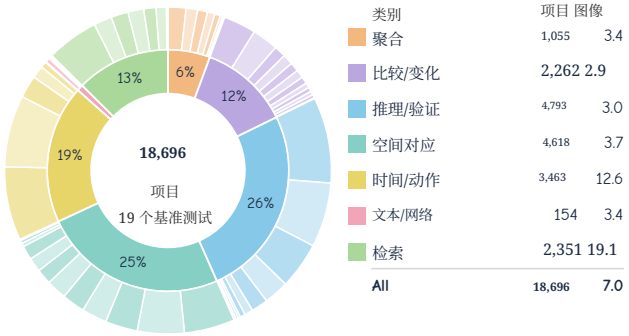


图 5. **ASTERISM 的组成**。 内环显示七类任务，外环显示其来源基准测试；两环均按项目数量排列；图像表示每项的平均图像数。

得分（表 2中的平均值，表 4中的条目）是七类任务的宏平均，因此较大的类别不会决定得分。表 S9 还给出了微平均值。

ASTERISM 使用一个不依赖大语言模型评审的基于规则的评分器为每个条目评分。对于 21 个模型的相同生成结果，使用八个基准测试的官方代码而非 ASTERISM 评分器进行评分，会使各基准测试的准确率最多相差 17.5 个百分点，并使 210 个模型对中的多达 25 对的排序发生逆转（补充材料E.2）。每个条目都带有从其提示中提取的答案格式，评分器会提取符合该格式的最后一个明确答案，并拒绝列出多个选项的答案。一名不知道评分器判定结果的标注员对 600 个输出中的 599 个与其判定一致（补充材料 E.1）。

3.3. ASTERISM-Lite

ASTERISM-Lite 从开源基础模型最难应对的 ASTERISM 条目中选取 400 项，并按照任务类别和图像数量设置固定配额；我们还会逐项人工核查图像及其答案（补充材料D）。58 个开源多模态大语言模型在 ASTERISM-Lite 上的排名与其在 ASTERISM 上的排名一致

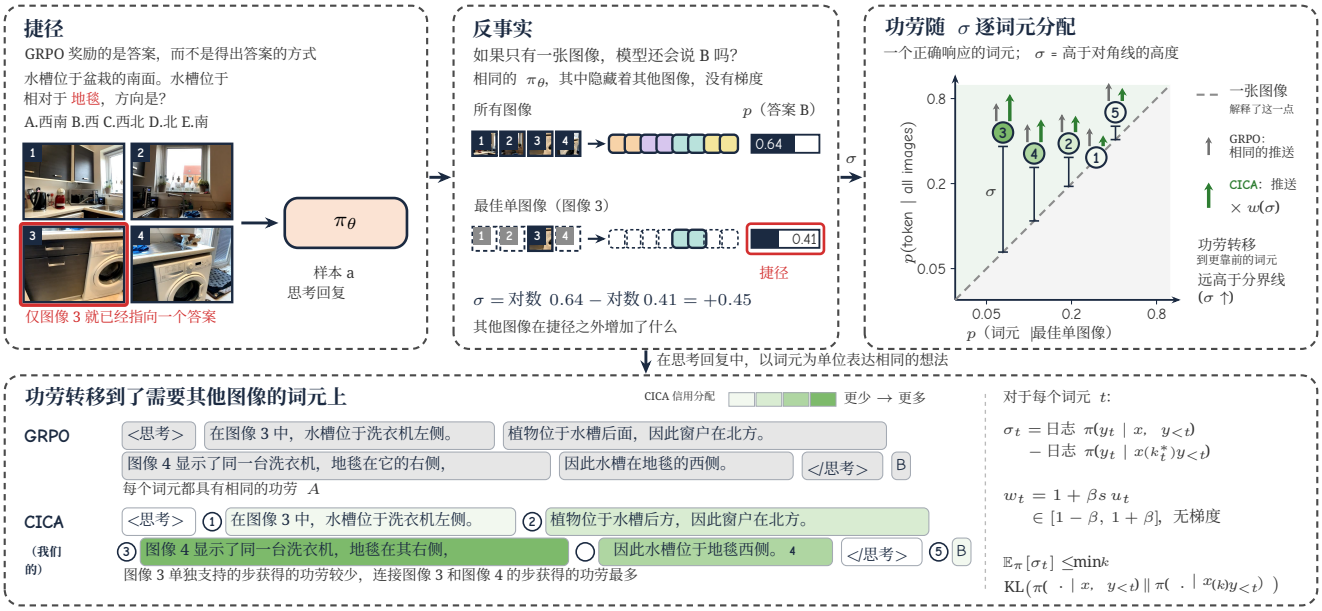


图 6。 **我们的跨图像信用分配（CICA）概览。**

斯皮尔曼相关系数为 0.93（表 S4）。我们还在 ASTERISM-Lite 上评测了专有多模态大语言模型（表 3）。

3.4. ASTERISM-Train

19个来源中的大多数仅发布评估条目，因此我们还收集了训练条目。ASTERISM-Train 包含 49,105 个具有可验证的闭式答案的多图像条目，这些条目来自 48 个来源，涵盖十个公共数据集的训练拆分、多图像指令集合，以及基于公开图像构建的合成任务（补充材料 G）。图像通过感知哈希和像素相关性与 ASTERISM 图像匹配的条目将被移除。对于采用非商业许可证的来源，我们发布条目标识符和构建脚本。

4. 跨图像信用分配提升需要多张图像的词元权重

RLVR 会奖励第 2节中的单图捷径，因为其奖励只检查最终答案，而 CICA 会将这一奖励的功劳转移给样本响应中无法由一张图像解释的词元。CICA 保留数据、可验证奖励和 RLVR目标，仅改变优势在词元之间的分配方式，使用策略在每次输入一张图像时自身的似然（图 6）。补充材料中的算法1 在补充材料 F中给出了完整的训练步。

GRPO [114] 为每个输入 x采样一组 G 个响应，为每个响应 y_i 计算可验证奖励 r_i，并在组内对奖励进行归一化，

$$A_i = \frac{r_i - \bar{r}}{\mathrm{std}(r)}. \tag{2}$$

每个词元 $y_{i,\,t}$ 都会在目标函数的策略项中获得这一优势。DAPO [157] 还会为每个响应保留一个优势，因此目标函数不会区分需要多张图像的词元与一张图像即可解释的词元。

4.1. 反事实跨图像对比

对于每个词元，CICA 衡量其他图像相较于最佳单图像额外提供了什么。设 $x^{(k)}$ 为输入 x，其中除 k 之外所有图像的视觉标记都已从上下文中移除。文本、图像分隔符和所有位置索引均保持与 x一致，因此模型看到的是相同的提示和图像槽位，只失去了其他图像的视觉证据。以这种方式移除其他图像，重现了第2.2 节中灰色遮罩所显示的差距（表S15）。对于样本响应 y_i中的每个词元，CICA 找出该词元在其下最有可能出现的单张图像，即最佳单图像$k^*_{i,\,t}$，并将该词元在所有图像下的对数似然与其在该图像下的对数似然进行对比，

$$k^*_{i,t} = \arg\max_k \log \pi_\theta(y_{i,t} \mid x^{(k)}, y_{i,<t}), \tag{3}$$

$$\sigma_{i,t} = \log \pi_\theta(y_{i,t} \mid x, y_{i,<t}) - \log \pi_\theta(y_{i,t} \mid x^{(k^*_{i,t})}, y_{i,<t}). \tag{4}$$

只有在最佳单图像也无法解释某个词元时，才将其视为需要多张图像。对 $y_{i,\,t}$ 在 π_θ 下，$\sigma_{i,\,t}$ 的期望值至多为 $\min_k \mathrm{KL}\big(\pi_\theta(\cdot \mid x,\ y_{i,\,<t}) \,\|\, \pi_\theta(\cdot \mid x^{(k)},\ y_{i,\,<t})\big)$（图 6）。

针对每个词元选择最佳单图像，可以严格检验这样一种响应：它在每一步都依据不同的图像作答。若以平均图像或随机选取的单张图像作为参照，那么从一张图像中读出的答案仍会得到较高的 σ，因为其他图像并不支持该答案（补充材料 F.5）。我们在两个基础模型的答案上测量 σ，并将每个答案的概率在各个选项上重新归一化。与最佳单图像相比，σ 对于

表 2. **ASTERISM 排行榜（代表性模型）**。各列依次为聚合（Agg）、比较/变化（Cmp）、推理/验证（Rsn）、空间对应（Spa）、时间/动作（Tmp）、文本/网络（Txt）和检索（Ret）。后训练模型（右侧）按其所基于的基础模型分组，各组顶部以灰色重复显示基础模型的得分；每组中，四舍五入前各列的最高值以粗体显示，第二高值加下划线。§专为多图像任务构建或微调。

基础模型

模型	规模	聚合	比较	推理	空间	时间	文本	检索	平均值
Gemma-4-12B [35]	120亿参数	53.6	42.6	46.5	43.3	36.1	75.3	49.2	49.5
GLM-4.1V-9B-Thinking [43]	9B	64.3	58.2	66.4	51.8	44.7	70.1	55.3	58.7
InternVL3-8B [170]	8B	49.5	43.5	39.9	37.7	42.8	44.8	38.3	42.4
LLaVA-Interleave-7B [67]§	7B	37.0	37.8	30.4	29.6	25.9	35.7	17.0	30.5
LLaVA-OneVision-2-80亿参数 [4]	8B	40.4	40.1	41.5	38.7	37.0	42.9	34.2	39.3
MiMo-VL-7B [147]	7B	65.0	63.9	67.3	52.5	40.0	72.7	54.5	59.4
Pixtral-12B [3]	120亿参数	42.3	33.6	32.7	35.1	23.3	43.5	23.1	33.4
Qwen2.5-VL-7B [7]	7B	45.8	38.3	40.0	42.0	36.1	40.3	42.2	40.7
Qwen2.5-VL-32B [7]	32B	54.8	58.8	52.1	48.2	37.7	53.9	45.9	50.2
Qwen3-VL-8B [104]	8B	50.9	46.8	44.8	47.0	40.0	54.5	47.7	47.4
Qwen3-VL-32B [104]	32B	62.1	56.8	66.0	50.8	46.0	64.3	54.1	57.2
Qwen3.5-9B [105]	9B	77.9	64.3	71.9	55.3	51.2	77.9	61.3	65.7
Qwen3.6-27B [106]	270亿参数	78.7	64.1	72.6	58.4	49.6	77.3	62.9	66.2

后训练模型

模型	规模	汇总	比较	推理	空间	时间	文本	检索	平均值
来自 Qwen2.5-VL-7B [7]		45.8	38.3	40.0	42.0	36.1	40.3	42.2	40.7
Video-R1-7B [29]	7B	40.7	33.0	35.8	37.2	34.8	40.9	39.5	37.4
Visual-Jigsaw-70亿参数 [143]§	7B	44.0	40.6	40.2	40.7	36.1	40.3	40.4	40.3
Vision-Zero-7B [131]	7B	44.7	37.1	40.1	42.2	36.2	40.3	41.9	40.4
ThinkLite-VL-7B [135]	7B	45.9	40.0	40.0	42.3	36.1	39.0	41.8	40.7
VLAA-Thinker-7B [14]	7B	44.5	39.0	40.3	42.7	36.6	40.3	41.9	40.8
R1-ShareVL-7B [153]	7B	47.5	42.6	42.3	42.9	36.8	39.6	41.4	41.9
PeRL-7B [165]§	7B	46.6	44.0	41.7	40.8	36.8	44.2	40.7	42.1
CICA-Qwen2.5-VL-7B（我们的）	7B	**52.1**							
来自 GLM-4.1V-9B-Thinking [43]		64.3	58.2	66.4	51.8	44.7	70.1	55.3	58.7
CICA-GLM-4.1V-9B-Thinking（我们的）	9B	**69.3**							
来自 Qwen3.5-9B [105]		77.9	64.3	71.9	55.3	51.2	77.9	61.3	65.7
CICA-Qwen3.5-9B（我们的模型）	9B	**80.0**							

表3. **ASTERISM- 轻量版上的专有和开源多模态大语言模型**。七类任务的平均值。

专有		开源，仅限轻量版	
模型	Avg	模型	Avg
Claude Opus 5.5 [6]	67.9	Qwen3.8-Flash-Next [107]	60.7
Gemini 3.8 Flash [36]	64.1	Kimi K3.1 Pro [93]	53.2
GPT-6-Astra [101]	62.8	GLM-5.3-Flash [161]	48.9
GPT-6.1-Sol [100]	61.6	Muse-Glimmer-30B [94]	44.8
Qwen3.8-Max [108]	60.9	MiMo-V2.6-Pro [148]	39.3
Muse Spark 1.3 [91]	59.6	Nemotron-3.5-Super-VL [96]	33.6

表4. **CICA 在 ASTERISM 上相比其他强化学习目标提升了各个骨干网络的表现**。

方法	Qwen2.5-VL-7B	GLM-4.1V-9B-Thinking	Qwen3.5-9B
Base	40.7	58.7	65.7
GRPO [114]	41.7	57.8	66.7
DAPO [157]	43.0	58.1	65.7
APPO [24]	43.1	58.7	65.9
VGPO [138]	42.7	58.7	67.1
CFPO [158]	43.6	59.0	64.0
CICA（本文方法）	**46.7**	**61.3**	**70.6**

对一张图像即可回答的条目，正确答案的值接近于零（两个模型的中位数均为 0.0004）；而对 ASTERISM 条目的正确答案，该值为正（中位数分别为 0.12 和 0.19）。纯文本参照对这些答案的排序恰好相反，而单图均值参照对它们的区分度较低（补充材料 F.5）。

4.2. 词元级优势重加权

CICA 将 σ 映射为每个词元优势上的有界权重。对于一组 $\mathcal{T}_i$ 活跃词元，它会对 σ进行中心化、压缩，再次中心化，

$$w_{i,t} = 1+\beta s_i u_{i,t},\ \ u_{i,t} = \tfrac{1}{2}\big(h_{i,t}-\bar{h}_i\big),\ \ h_{i,t} = \tanh\tfrac{\sigma_{i,t}-\bar{\sigma}_i}{\tau}, \tag{5}$$

其中 $\beta \in [0,1)$ 决定强度，$\beta = 0$ 对应 GRPO，而 τ 决定权重的尺度（表S8）。指示变量s_i 当响应包含符合其条目格式的答案（补充材料 F.3），且该答案正确时取 $A_i > 0$；答案错误时取 $A_i < 0$；否则取 $s_i = 0$。 $\mathcal{T}_i$ 包含 y_i 的活跃词元，以及与其共同中心化的任何响应中的活跃词元；这些词元包括除答案标记和用于界定响应的词元以外的所有词元，例如回合结束标记和思考标签（补充材料 F.4）。在底层可验证奖励强化学习目标的策略项中，CICA 将每个词元的优势乘以 $w_{i,\ t}$ 并将 $w_{i\text{设为 }t} = 1$ 在.....之外 $\mathcal{T}_i$（补充材料 F）。

¯平均值 $\bar{\sigma}_i$ 和 h_i 根据每个响应在更新中的占比分配权重，因此 w 除以 $\mathcal{T}_i$ 的更新加权平均值为 1，而 CICA 会在 $\mathcal{T}_i$ 的词元之间分配功劳（图 6）。

使用 s_i，需要多张图像的词元会在正确响应的奖励和错误响应的惩罚中占据更大份额。由于 $|u| < 1$ 和 $\beta < 1$，CICA 保留每个词元系数的符号，并将其权重限制在 $[1-\beta, 1+\beta]$。

5. 实验

5.1. ASTERISM 上的 58 个多模态大语言模型基准测试

我们使用相同的提示词评估 58 个开源多模态大语言模型，每个模型均采用其默认图像分辨率或官方多图像设置（补充材料 E.3）。表2 列出了具有代表性的模型。

对于以相同模式运行的模型（补充材料E.3），每个系列的准确率均随模型规模增大而提高；对于大多数参数量不低于 4B 的模型，时间/动作类条目的得分在七个类别中最低（表2 和S9）。在不同模型间，跨图像操作为合并图像的条目比操作为比较图像的条目更难（表S11）。

ASTERISM-Lite 上的专有多模态大语言模型。Claude Opus 5.5 得分最高；在开源多模态大语言模型中，只有 Qwen3.8-Flash-Next 达到专有模型的范围（表 3）。

5.2. CICA 提升每个骨干网络的表现

我们使用 Qwen2.5-VL-7B [7],GLM-4.1V-9B-Thinking [43]和 Qwen3.5-9B [105] 从各自的公开检查点出发，在 ASTERISM-Train 上训练一个训练周期（补充材料 G），并将训练后的模型称为 CICA-Qwen2.5-VL-7B、CICA-GLM-

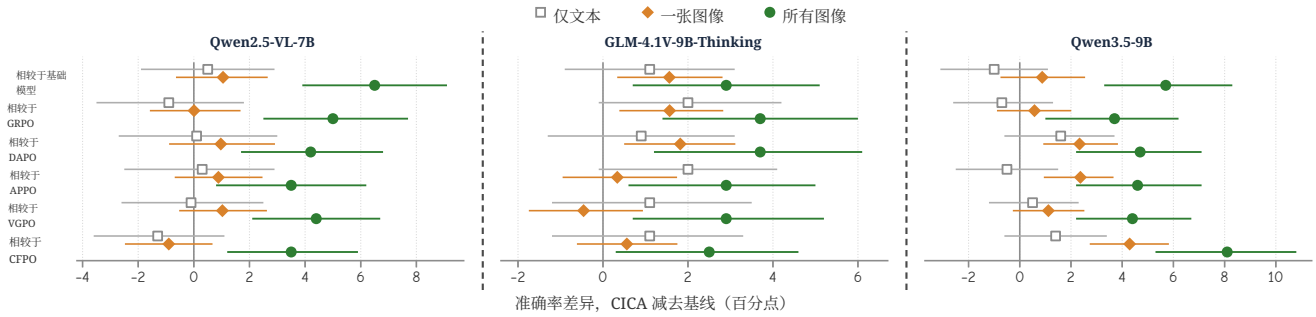


图 7. **所有图像均可见时，CICA 的提升最大。** 在第 2.2 节的灰色遮罩条件下，针对 1,000 个 ASTERISM 项目，CICA 的准确率减去各基线准确率所得的差值。 2.2，分别显示无图像、可见一张图像或可见所有图像的情况。柱状图显示 95% 置信区间，各面板的刻度不同。

4.1V-9B-Thinking 和 CICA-Qwen3.5-9B。在这些运行中，CICA 会重新加权 GRPO 的优势。表 4 中的五个强化学习基线，包括两个通用方法（GRPO、DAPO）和三个视觉专用方法（APPO、VGPO、CFPO），在数据、准确率奖励、提示词、组大小、批大小和学习率计划方面均与 CICA 保持一致，其余则遵循各自论文或官方代码（补充材料 G.3）。CICA 不增加参数、轨迹采样或反向传播。它仅针对每个响应中的每幅图像增加一次无梯度遍历；在 Qwen2.5-VL-7B 上，这约占一个训练步的 9%（补充材料 F.7）。

提升幅度超过强化学习基线。 CICA 使基础模型在 ASTERISM 上的得分提高了 2.6 至 6.0 分，并且在三种骨干模型上，其得分均超过所有强化学习基线，所有配对的 95% 置信区间均高于零（表 4；补充材料 H.3 和 H.7）。在每种骨干模型上，强化学习基线相较于基础模型的提升幅度都不到 CICA 的一半。在三种骨干模型上，CICA 在七类任务中的每一类得分都高于基础模型（表 2）。尽管 GLM-4.1V-9B-Thinking 和 Qwen3.5-9B 已经过强化学习后训练，CICA 仍进一步提升了二者的表现（表 4）。CICA-Qwen2.5-VL-7B 的得分高于所有其他基于 Qwen2.5-VL-7B 进行后训练的模型（表S9和S10）。

所有图像均可见时，收益最大。 在第2.2节的灰色遮罩条件下，CICA 与各基线之间的准确率差距在所有图像均可见时大于仅一张图像可见时（图7）。对于 Qwen2.5-VL-7B，当其他图像如 CICA 中一样被隐藏，或被替换为其他项目的图像时，相较于基础模型、GRPO 和 DAPO，结论同样成立（表 S17a 和S17b）。经过 CICA 训练后，各骨干网络的错误答案中重复单图像答案的比例，相较于经过 GRPO 训练后均有所降低，最多降低 7.7 个百分点（图S19）。在每个训练模型的正确响应中（图S14），与图像相关的步骤所携带的权重，高于描述单张图像的大多数步骤，如图6所示。

单图得分得以保留或提升。 尽管 CICA 仅使用多图像条目进行训练，但它在全部十个单图像基准测试中保留或提升了每个骨干网络的得分（表 S18）。它提升了每个骨干网络在幻觉基准 POPE [71]以及全部三个真实世界基准测试 [66, 145, 166]上的表现。

表 5。**CICA 的消融实验。**所有 – 一项是图7所示项目中，所有图像均可见时相较于 GRPO 的准确率增益减去仅一张图像可见时的准确率增益。

训练目标	ASTERISM	全部 – 一
GRPO（$\beta = 0$）	41.7	—
CICA，置换 σ	41.3	−1.0
CICA，针对错误回答进行 $w = 1$	44.1	+1.5
CICA，纯文本参照	44.4	+0.8
CICA，单图均值参照	45.6	+2.2
CICA（最佳单图像）	46.7	+5.0

5.3. CICA 在以 σ 对照最佳单图像进行测量时效果最佳

我们在 Qwen2.5-VL-7B 上对 CICA 进行消融实验，并使用相同的训练器、训练数据和奖励训练表 5 中的每一行（补充材料 G）。该表给出了 ASTERISM 得分以及全 – 一准确率：在所有图像可见时相对于 GRPO 的准确率增益，减去仅一张图像可见时的准确率增益。置换 σ变体会在形成权重之前，对 σ 在公式（4）中的数值进行打乱，并在 CICA 对其进行中心化的各个集合内分别执行。该变体在 ASTERISM 和全 – 一准确率上均低于 GRPO，因此 CICA 的增益来自 σ 与图像之间的联系。将 $w = 1$ 设置在错误回答上会使其继续留在 CICA 进行中心化的集合中 σ，并且仅对正确回答进行加权。该变体在 ASTERISM 和全 – 一准确率上均低于 CICA，因此错误回答上的权重对两者都有贡献。其得分仍高于表4中的所有强化学习基线；将 β降至 0.5 或将 τ加倍时，CICA 也是如此（表 S22）。纯文本参照保留了 CICA 约一半的 ASTERISM 得分增益，但几乎没有保留全 – 一准确率增益，因为它同样会将在一张图像中读出的答案的词元计入功劳。单图均值参照保留的全 – 一准确率增益不到一半，因此，以最佳单图像作为参照时，该增益最大（补充材料 F.5）。

6. 结论

许多多图像条目和可验证奖励强化学习都在奖励多模态大语言模型所采用的单图像捷径。在 19 个公开基准中，仅凭一张图像或问题本身就能解决五分之一以上的条目；即使一张图像不足以作答，多模态大语言模型的大多数

错误答案仍会重复单图像答案，远高于随机机会。ASTERISM 会在评估中移除那些会奖励这种捷径的条目，并保留 18.7k 个需要多张图像的条目。ASTERISM-Lite 选取 400 个难题条目，由我们逐一手动核验；ASTERISM-Train 则新增 49k 个带有可验证答案的训练条目。在训练中，CICA 将可验证奖励强化学习的功劳分配给需要多张图像的词元，其效果优于使用相同数据训练的强化学习基线。使用所有图像时，CICA 的提升最大，且不会降低单图得分。

参考文献

[1] Abdelrahman Abouelenin、Atabak Ashfaq、Adam Atkinson、Hany Hassan Awadalla、Nguyen Bach、Jianmin Bao、Alon Benhaim、Martin Cai、Vishrav Chaudhary、Congcong Chen等。Phi-4-Mini技术报告：通过LoRA混合实现紧凑而强大的多模态语言模型。CoRR，abs/2503.01743，2025年。3，S3[2] Aishwarya Agrawal、Dhruv Batra、Devi Parikh和Anirud-dha Kembhavi。不要只凭假设；观察并回答：克服视觉问答中的先验。在2018年IEEE计算机视觉与模式识别会议（CVPR 2018）上，盐湖城，犹他州，美国，2018年6月18日至22日，页码4971–4980。计算机视觉基金会/IEEE计算机学会，2018年。2，S1[3] Pravesh Agrawal、Szymon Antoniak、Emma Bou Hanna、Baptiste Bout、Devendra Singh Chaplot、Jessica Chud-novsky、Diogo Costa、Baudouin De Monicault、Saurabh Garg、Th´eophile Gervet、Soham Ghosh、Am´elie H´eliou、Paul Jacob、Albert Q. Jiang、Kartik Khandelwal、Timoth´ee Lacroix、Guillaume Lample、Diego de Las Casas、Thibaut Lavril、Teven Le Scao、Andy Lo、William Marshall、Louis Martin、Arthur Mensch、Pavankumar Muddireddy、Valera Nemychnikova、Marie Pellat、Patrick von Platen、Nikhil Raghuraman、Baptiste Rozi`ere、Alexandre Sablayrolles、Lu-cile Saulnier、Romain Sauvestre、Wendy Shang、Roman Soletskyi、Lawrence Stewart、Pierre Stock、Joachim Stud-nia、Sandeep Subramanian、Sagar Vaze、Thomas Wang和Sophia Yang。Pixtral 12B。CoRR，abs/2410.07073，2024年。7，S22，S28，S32[4] Xiang An、Yin Xie、Feilong Tang、Yunyao Yan、Huajie Tan、Didi Zhu、Changrui Chen、Xiuwei Zhao、Bin Qin、Kaicheng Yang、Yifei Shen、Yuanhan Zhang、Kaichen Zhang、Wenkang Zhang、Zheng Cheng、Nansen Zhang、Chunsheng Wu、Chun-jiang Ge、Zimin Ran、Dehua Song、Chunyuan Li、Shikun Feng、Ming Hu、Zhangquan Chen、Junbo Niu、Bo Li、Ziy-ong Feng、Ziwei Liu、Zongyuan Ge和Jiankang Deng。LLaVA-OneVision-2：迈向下一代感知智能。arXiv预印本，arXiv:2605.25979，2026年。7，S22，S28，S32[5] Anthropic。系统卡：Claude Fable 5与Claude Mythos 5。https://anthropic.com/claude-fable-5-mythos-5-system-card，2026年。系统卡，2026年6月。3，S3[6] Anthropic。系统卡：Claude Opus 5.5。https: //anthropic.com/claude-opus-5-5-system-卡，2026年。系统卡，2026年9月。7，S21，S22，S23[7] Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wen-bin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang、Humen Zhong、Yuanzhi Zhu、Ming-Hsuan 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。CoRR，abs/2502.13923，2025年。1，7，S22，S24，S25，S28，S29，S31，S32，S33，S34[8] Sule Bai、Mingxing Li、Yong Liu、Jing Tang、Haoji Zhang、Lei Sun、Xiangxiang Chu 和 Yansong Tang。UniVG-R1：由推理引导的通用视觉定位与强化学习。arXiv 预印本 arXiv:2505.14231，2025年。S3，S22，S28，S29，S32[9] Yoav Benjamini 和 Yosef Hochberg。控制假发现率：一种实用且强大的多重检验方法。《英国皇家统计学会杂志：B辑（方法论）》，57(1):289–300，1995年。S3[10] Shyamal Buch、Crist´obal Eyzaguirre、Adrien Gaidon、Jiajun Wu、Li Fei-Fei 和 Juan Carlos Niebles。重新审视视频-语言理解中的“视频”。载于 IEEE/CVF计算机视觉与模式识别会议，CVPR 2022，美国路易斯安那州新奥尔良，2022年6月18日至24日，页码 2907–2917。IEEE，2022年。S2[11] 字节跳动 Seed。Seed1.5-VL 技术报告。CoRR，abs/2505.07062，2025年。2，S2[12] Alex James Chan、Hao Sun、Samuel Holt 和 Mihaela van der Schaar。通过人类反馈进行强化学习，免费获得稠密奖励。载于第四十一届国际机器学习大会，ICML 2024，奥地利维也纳，2024年7月21日至27日，页码 6136–6154。PMLR / OpenReview.net，2024年。S2[13] Yingshan Chang、Guihong Cao、Mridu Narang、Jianfeng Gao、Hisami Suzuki 和 Yonatan Bisk。WebQA：多跳与多模态问答。载于 IEEE/CVF计算机视觉与模式识别会议，CVPR 2022，美国路易斯安那州新奥尔良，2022年6月18日至24日，页码 16474–16483。IEEE，2022年。S5，S25[14] Hardy Chen、Haoqin Tu、Fali Wang、Hui Liu、Xianfeng Tang、Xinya Du、Yuyin Zhou 和 Cihang Xie。监督微调还是强化学习？对训练 R1类推理大型视觉语言模型的初步研究。《机器学习研究汇刊》，2025年。7，S22，S28，S29[15] Jifan Chen 和 Greg Durrett。理解多跳推理的数据集设计选择。载于2019年计算语言学协会北美分会会议论文集（NAACL），页码 4026–4032，2019年。S1[16] Long Chen、Xin Yan、Jun Xiao、Hanwang Zhang、Shiliang Pu 和 Yueting Zhuang。合成反事实样本，实现稳健的视觉问答。载于2020年IEEE/CVF计算机视觉与模式识别会议，CVPR 2020，美国华盛顿州西雅图，2020年6月13日至19日，页码 10797–10806。计算机视觉基金会 / IEEE，2020年。S1，S3

[17] Lin Chen、Jinsong Li、Xiaoyi Dong、Pan Zhang、Yuhang Zang、Zehui Chen、Haodong Duan、Jiaqi Wang、乔宇、Dahua Lin 和 Feng Zhao。我们评估大型视觉语言模型的方向正确吗？载于《神经信息处理系统进展第37卷：2024年神经信息处理系统年会》，NeurIPS 2024，加拿大不列颠哥伦比亚省温哥华，2024年12月10日至15日，2024年。2，S1[18] Xi Chen、Mingkang Zhu、Shaoteng Liu、Xiaoyang Wu、Xiao-gang Xu、Yu Liu、Xiang Bai 和 Hengshuang Zhao。MiCo：用于强化视觉推理的多图像对比。载于《第38届神经信息处理系统进展：2025年神经信息处理系统年会》，NeurIPS 2025，美国加利福尼亚州圣迭戈，2025年12月2日至7日 5，2025，2025 30 2，3/ 墨西哥墨西哥城，11月至12月。S[19] Ziming Cheng、Binrui Xu、Lisheng Gong、Zuhe Song、Tian-shuo Zhou、Shiqi Zhong、Siyu Ren、Mingxiang Chen、Xi-angchao Meng、Yuxin Zhang、Yanlin Li、Lei Ren、Wei Chen、Zhiyuan Huang、Mingjie Zhan、Xiaojie Wang 和 Fangxi-ang Feng。使用多模态多图像推理基准评估多模态大语言模型。CoRR，abs/2506.04280，2025年。2，S1，S4，S5，S23，S34[20] Chanseul Cho、Noori Bae、Dongheon Lee 和 Youngmin Ro。GRIT：对富文本推理中视觉定位错误的词元进行梯度掩蔽。载于 IEEE/CVF冬季计算机视觉应用会议论文集（WACV），2027年。2，S2[21] Ganqu Cui、Lifan Yuan、Zefan Wang、Hanbin Wang、Yuchen Zhang、Jiacheng Chen、Wendi Li、Bingxiang He、Yuchen Fan、Tianyu Yu、Qixin Xu、Weize Chen、Jiarui Yuan、Huayu Chen、Kaiyan Zhang、Xingtai Lv、Shuo Wang、Yuan Yao、Xu Han、Hao Peng、Yu Cheng、Zhiyuan Liu、Maosong Sun、Bowen Zhou 和 Ning Ding。通过隐式奖励进行过程强化。《机器学习研究汇刊》，2026年。S2[22] Anurag Das、Adrian Bulat、Alberto Baldrati、Ioannis Maniadis Metaxas、Bernt Schiele、Georgios Tzimiropoulos 和 Brais Martinez。图像越多，问题越多？对视觉语言模型失效模式的对照分析。载于计算语言学协会论文集：ACL 2026，页码 7423–7442。计算语言学协会，2026年。2，S3[23] Matt Deitke、Christopher Clark、Sangho Lee、Rohun Tripathi、Yue Yang、Jae Sung Park、Mohammadreza Salehi、Niklas Muennighoff、Kyle Lo、Luca Soldaini 等。Molmo 和 PixMo：用于最先进大型视觉语言模型的开放权重与开放数据。载于 IEEE/CVF计算机视觉与模式识别会议，CVPR 2025，美国田纳西州纳什维尔，2025年6月11日至15日，页码 91–104。计算机视觉基金会 / IEEE，2025年。3，S3[24] Henghui Du、Chang Zhou、Xi Chen 和 Di Hu。APPO：面向视频推理的注意力引导感知策略优化。载于 IEEE/CVF计算机视觉与模式识别会议论文集（CVPR），2026年。2，7，S2，S27[25] Haodong Duan、Junming Yang、Yuxuan Qiao、Xinyu Fang、Lin Chen、Yuan Liu、Xiaoyi Dong、Yuhang Zang、Pan Zhang、Jiaqi Wang、Dahua Lin 和 Kai Chen。VLMEvalKit：一个用于评估大型多模态模型的开源工具包。在《第32届ACM国际多媒体会议》论文集（MM 2024），澳大利亚维多利亚州墨尔本，2024年10月28日至2024年11月1日，第11198–11201页。ACM，2024年。S21，S33[26] Bradley Efron。自助法：重新审视刀切法。《统计年鉴》，7(1):1–26，1979年。S21，S29[27] Kaixuan Fan、Kaituo Feng、Haoming Lyu、Dongzhan Zhou 和 Xiangyu Yue。SophiaVL-R1：通过思考奖励强化多模态大语言模型的推理。在《国际学习表征会议》，第9290–9310页，2026年。S22，S29[28] Yue Fan、Xuehai He、Diji Yang、Kaizhi Zheng、Ching-Chen Kuo、Yuting Zheng、Xinze Guan 和 Xin Wang。GRIT：教会多模态大语言模型借助图像进行思考。在《神经信息处理系统进展》，第116522–116543页。Curran Associates, Inc.，2025年。S2，S22，S29[29] Kaituo Feng、Kaixiong Gong、Bohao Li、Zonghao Guo、Yibing Wang、Tianshuo Peng、Junfei Wu、Xiaoying Zhang、Benyou Wang 和 Xiangyu Yue。Video-R1：强化多模态大语言模型中的视频推理。在《神经信息处理系统进展》，第99114–99137页。Curran Associates, Inc.，2025年。7，S22，S29[30] Jakob N. Foerster、Gregory Farquhar、Triantafyllos Afouras、Nantas Nardelli 和 Shimon Whiteson。反事实多智能体策略梯度。在《第三十二届人工智能 AAAI 会议》（AAAI-18）、《第三十届创新人工智能应用会议》（IAAI-18）和《第八届人工智能教育进展 AAAI 研讨会》（EAAI-18）论文集，新奥尔良，路易斯安那州，美国，2018年2月2日至7日，第2974–2982页。AAAI出版社，2018年。S2[31] Xingyu Fu、Yushi Hu、Bangzheng Li、Yu Feng、Haoyu Wang、Xudong Lin、Dan Roth、Noah A. Smith、Wei-Chiu Ma 和 Ranjay Krishna。BLINK：多模态大语言模型能够看见，却无法感知。在《计算机视觉—ECCV 2024—第十八届欧洲会议》论文集，意大利米兰，2024年9月29日至10月4日，第二十三部分，第148–166页。施普林格，2025年。2，S1，S2，S4，S34[32] Robert Geirhos、J¨orn-Henrik Jacobsen、Claudio Michaelis、Richard S. Zemel、Wieland Brendel、Matthias Bethge 和 Felix A. Wichmann。深度神经网络中的捷径学习。《自然·机器智能》，2(11):665–673，2020年。2[33] Aryo Pradipta Gema、Joshua Ong Jun Leang、Giwon Hong、Alessio Devoto、Alberto Carlo Maria Mancino、Rohit Saxena、Xuanli He、Yu Zhao、Xiaotang Du、Mohammad Reza Ghasemi Madani、Claire Barale、Robert McHardy、Joshua Harris、Jean Kaddour、Emile van Krieken 和 Pasquale Minervini。我们已经解决MMLU了吗？在《2025年美洲分会计算语言学协会会议：人类语言技术》论文集，NAACL 2025——第一卷：长篇论文，新墨西哥州阿尔伯克基，美国，2025年4月29日至5月4日，第5069–5096页。计算语言学协会，2025年。S1[34] Gemma 团队。Gemma 3技术报告。CoRR，abs/2503.19786，2025年。S22，S28，S32

[35] Gemma 团队。Gemma 4技术报告。arXiv预印本，arXiv:2607.02770，2026年。7，S22，S28，S32[36] Google DeepMind。Gemini 3.8 Flash模型卡片。https: / / deepmind . google / models / model - cards / gemini-3-8-flash/，2026年。模型卡片，2026年9月。7，S22，S23[37] Yash Goyal、Tejas Khot、Douglas Summers-Stay、Dhruv Batra 和 Devi Parikh。让VQA中的V发挥作用：提升图像理解在视觉问答中的作用。在《2017年IEEE计算机视觉与模式识别会议》，CVPR 2017，美国夏威夷州檀香山，2017年7月21日至26日，第6325–6334页。IEEE计算机学会，2017年。2，S1[38] Tianrui Guan、Fuxiao Liu、Xiyang Wu、Ruiqi Xian、Zongxia Li、Xiaoyu Liu、Xijun Wang、Lichang Chen、Furong Huang、Yaser Yacoob、Dinesh Manocha 和 Tianyi Zhou。HallusionBench：用于诊断大型视觉语言模型中交织的语言幻觉和视觉错觉的高级诊断套件。在IEEE/CVF计算机视觉与模式识别会议，CVPR 2024，美国华盛顿州西雅图，2024年6月16日至22日，第14375–14385页。IEEE，2024年。2，S33[39] Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song、Peiyi Wang、Qihao Zhu、Runxin Xu、Ruoyu Zhang、Shirong Ma、Xiao Bi 等。DeepSeek-R1通过强化学习激励大语言模型进行推理。《自然》，645 638，2025年8081 633 2，2 ( ): – 。S[40] Tianchen Guo、Chen Liu、Ling Chen 和 Xin Yu。SSMN-Bench：通过单视角充分性和多视角必要性诊断基于图像的跨视角人-物理解。在《欧洲计算机视觉会议》（ECCV），2026年。2，S2[41] Tianchen Guo、Chen Liu 和 Xin Yu。超越单视角充分性：面向跨视角人类理解的CVBench。在《IEEE/CVF计算机视觉与模式识别会议》（CVPR）论文集，第7154–7164页，2026年 2，2 – 。S[42] Yunzhuo Hao、Jiawei Gu、Huichen Will Wang、Linjie Li、Zhengyuan Yang、Lijuan Wang 和 Yu Cheng。多模态大语言模型能够进行多模态推理吗？EMMA：增强型多模态推理基准。在《第四十二届国际机器学习会议》，ICML 2025，加拿大不列颠哥伦比亚省温哥华，2025年7月13日至19日。PMLR / OpenReview.net，2025年。2，S1，S4，S23，S34[43] Wenyi Hong、Wenmeng Yu、Xiaotao Gu、Guo Wang、Guobing Gan、Haomiao Tang、Jiale Cheng、Ji Qi、Junhui Ji、Lihang Pan、Shuaiqi Duan、Weihan Wang、Yan Wang、Yean Cheng、Zehai He、Zhe Su、Zhen Yang、Ziyang Pan、Aohan Zeng、Baoxu Wang、Boyan Shi、Changyu Pang、Chenhui Zhang、Da Yin、Fan Yang、Guoqing Chen、Jiazheng Xu、Jiali Chen、Jing Chen、Jinhao Chen、Jinghao Lin、Jinjiang Wang、Junjie Chen、Leqi Lei、Letian Gong、Leyi Pan、Mingzhi Zhang、Qinkai Zheng、Sheng Yang、Shi Zhong、Shiyu Huang、Shuyuan Zhao、Siyan Xue、Shangqin Tu、Shengbiao Meng、Tianshu Zhang、Tianwei Luo、Tianxiang Hao、Wenkai Li、Wei Jia、Xin Lyu、Xuancheng Huang、Yanling Wang、Yadong Xue、Yanfeng Wang、Yifan An、Yifan Du、Yi ming Shi、Yiheng Huang、Yilin Niu、Yuan Wang、Yuanchang Yue、Yuchen Li、Yutao Zhang、Yuxuan Zhang、Zhanxiao Du、Zhenyu Hou、Zhao Xue、Zhengxiao Du、Zihan Wang、Peng Zhang、Debing Liu、Bin Xu、Juanzi Li、Minlie Huang、Yuxiao Dong 和 Jie Tang。GLM-4.1V-Thinking：通过可扩展的强化学习实现多样化的多模态推理。CoRR，abs/2507.01006，2025年。2，7，S2，S21，S22，S24，S25，S28，S29，S30，S31，S32，S33[44] Wenbo Hu、Jia-Chen Gu、Zi-Yi Dou、Mohsen Fayyaz、Pan Lu、Kai-Wei Chang 和 Nanyun Peng。MRAG-Bench：面向检索增强多模态模型的视觉中心评估。载于第十三届国际学习表征会议（ICLR 2025），新加坡，2025年4月24日至28日。OpenReview.net，2025年。2，S1，S4，S34[45] Siyuan Huang、Xiaoye Qu、Yafu Li、Yun Luo、Zefeng He、Daizong Liu 和 Yu Cheng。聚焦多模态强化学习中的词元感知。载于第十四届国际学习表征会议（ICLR 2026），巴西里约热内卢，2026年4月23日至27日。proceedings.iclr.cc，2026年。S2，S22，S27，S28[46] Wenxuan Huang、Bohan Jia、Shaosheng Cao、Zheyu Ye、Fei Zhao、Zhe Xu、Yao Hu 和 Shaohui Lin。Vision-R1：激励多模态大语言模型的推理能力。载于第十四届国际学习表征会议（ICLR 2026），巴西里约热内卢，2026年4月23日至27日。proceedings.iclr.cc，2026年。2，S2[47] Harsh Jhamtani 和 Taylor Berg-Kirkpatrick。学习描述相似图像对之间的差异。载于2018年自然语言处理实证方法会议论文集，比利时布鲁塞尔，2018年10月31日至11月4日，第4024–4034页。计算语言学协会，2018年。S4，S25[48] Mengzhao Jia、Wenhao Yu、Kaixin Ma、Tianqing Fang、Zhihan Zhang、Siru Ouyang、Hongming Zhang、Dong Yu 和 Meng Jiang。LEOPARD：面向文本丰富的多图像任务的视觉语言模型。《机器学习研究汇刊》，2025年。S25[49] Dongfu Jiang、Xuan He、Huaye Zeng、Cong Wei、Max Ku、Qian Liu 和 Wenhu Chen。Mantis：交错式多图像指令微调。《机器学习研究汇刊》，2024年。1，2，S1，S2，S3，S4，S22，S25，S28，S29，S32，S34[50] Yichen Jiang 和 Mohit Bansal。避免推理捷径：多跳问答的对抗评估、训练与模型开发。载于计算语言学协会第57届年会（ACL）论文集，第2726–2736页，2019年。S2[51] Justin Johnson、Bharath Hariharan、Laurens van der Maaten、Li Fei-Fei、C. Lawrence Zitnick 和 Ross B. Girshick。CLEVR：用于组合语言和基础视觉推理的诊断数据集。载于2017年IEEE计算机视觉与模式识别会议（CVPR 2017），美国夏威夷州檀香山，2017年7月21日至26日，第1988–1997页。IEEE计算机学会，2017年。S25[52] Mehran Kazemi、Nishanth Dikkala、Ankit Anand、Petar Devic、Ishita Dasgupta、Fangyu Liu、Bahare Fatemi、Pranjal Awasthi、Sreenivas Gollapudi、Dee Guo 和 Ahmed Qureshi。ReMI：用于多图像推理的数据集。载于神经信息处理系统进展第37卷：神经信息处理系统年度会议

2024年，NeurIPS 2024，加拿大不列颠哥伦比亚省温哥华，2024年12月10日至15日，2024年。2，S1，S4，S5，S21，S23，S34[53] Amirhossein Kazemnejad、Milad Aghajohari、Eva Portelance、Alessandro Sordoni、Siva Reddy、Aaron C. Courville 和 Nicolas Le Roux。VinePPO：改进大语言模型强化学习训练中的信用分配。载于第四十二届国际机器学习会议（ICML 2025），加拿大不列颠哥伦比亚省温哥华，2025年7月13日至19日。PMLR / OpenReview.net，2025年。S2[54] Aniruddha Kembhavi、Mike Salvato、Eric Kolve、Min Joon Seo、Hannaneh Hajishirzi 和 Ali Farhadi。一张图表胜过十几张图像。载于《计算机视觉——ECCV 2016——第十四届欧洲会议》，荷兰阿姆斯特丹，2016年10月11日至14日，论文集第四部分，第235–251页。施普林格，2016年。S33[55] Kimi团队。Kimi k1.5：利用大语言模型扩展强化学习。CoRR，abs/2501.12599，2025年。S2[56] Kimi团队。Kimi-VL技术报告。CoRR，abs/2504.07491，2025年。3，S2，S3[57] Neal Krawetz。差不多就是那样。https : / / www . hackerfactor . com / blog / index . php ? /archives / 529 - Kind - of - Like - That . html，2013年。The Hacker Factor Blog，2013年1月。S25[58] Woosuk Kwon、Zhuohan Li、Siyuan Zhuang、Ying Sheng、Lianmin Zheng、Cody Hao Yu、Joseph Gonzalez、Hao Zhang 和 Ion Stoica。使用PagedAttention实现大语言模型服务的高效内存管理。载于第二十九届操作系统原理研讨会论文集（SOSP 2023），德国科布伦茨，2023年10月23日至26日，第611–626页。ACM，2023年。S24[59]Nathan Lambert、Jacob Morrison、Valentina Pyatkin、Shengyi Huang、Hamish Ivison、Faeze Brahman、Lester James Validad Miranda、Alisa Liu、Nouha Dziri、Xinxi Lyu、Yuling Gu、Saumya Malik、Victoria Graf、Jena D. Hwang、Jiangjiang Yang、Ronan Le Bras、Oyvind Tafjord、Christopher Wilhelm、Luca Soldaini、Noah A. Smith、Yizhong Wang、Pradeep Dasigi 和 Hannaneh Hajishirzi。Tulu 3：推动开放式语言模型后训练的前沿。载于第二届语言建模会议，2025年。S2[60] Jordy Van Landeghem、Rafal Powalski、Rub`en Tito、Dawid Jurkiewicz、Matthew B. Blaschko、Lukasz Borchmann、Micka¨el Coustaty、Sien Moens、Michal Pietruszka、Bertrand Anckaert、Tomasz Stanislawek、Pawel J´oziak 和 Ernest Valveny。文档理解数据集与评估（DUDE）。载于IEEE/CVF国际计算机视觉会议（ICCV 2023），法国巴黎，2023年10月1日至6日，第19471–19483页。IEEE，2023年。2，S1，S3，S4，S21，S23，S34[61] Hugo Laurenc,on、L´eo Tronchon、Matthieu Cord 和 Victor Sanh。构建视觉语言模型时，哪些因素最重要？载于《神经信息处理系统进展》，第87874–87907页。Curran Associates, Inc.，2024年。S22，S28，S29[62] Jie Lei、Tamara L. Berg 和 Mohit Bansal。揭示视频与语言学习中的单帧偏差。载于计算语言学协会第61届年会论文集（第一卷：长篇论文），ACL *2023*年，多伦多，加拿大，*2023*年*7*月*9*日至*14*日，第487–507页。计算语言学协会，2023年。S2[63] Sicong Leng、Hang Zhang、Guanzheng Chen、Xin Li、Shijian Lu、Chunyan Miao 和 Lidong Bing。通过视觉对比解码缓解大型视觉语言模型中的物体幻觉。载于 IEEE/CVF 计算机视觉与模式识别会议（CVPR 2024），美国华盛顿州西雅图，2024年6月16日至22日，第13872–13882页。IEEE，2024年。2[64] Baiqi Li、Zhiqiu Lin、Wenxuan Peng、Jean de Dieu Nyandwi、Daniel Jiang、Zixian Ma、Simran Khanuja、Ranjay Krishna、Graham Neubig 和 Deva Ramanan。NaturalBench：在自然对抗样本上评估视觉语言模型。载于神经信息处理系统进展第37卷：2024年神经信息处理系统年度会议，NeurIPS 2024，加拿大不列颠哥伦比亚省温哥华，2024年12月10日至15日，2024年。S1[65] Bo Li、Yuanhan Zhang、Dong Guo、Renrui Zhang、Feng Li、Hao Zhang、Kaichen Zhang、Peiyuan Zhang、Yanwei Li、Ziwei Liu 和 Chunyuan Li。LLaVA-OneVision：轻松实现视觉任务迁移。《机器学习研究汇刊》，2025年。S21、S22、S28、S32[66] Chunyi Li、Jianbo Zhang、Zicheng Zhang、Haoning Wu、Yuan Tian、Wei Sun、Guo Lu、Xiongkuo Min、Xiaohong Liu、Weisi Lin、Xiao-Ping Zhang 和 Guangtao Zhai。R-Bench：你的大型多模态模型能否抵御真实世界的损坏？《IEEE精选信号处理主题期刊》，19(7)：1349–1361，2025年。8、S33[67] Feng Li、Renrui Zhang、Hao Zhang、Yuanhan Zhang、Bo Li、Wei Li、Zejun Ma 和 Chunyuan Li。LLaVA-Interleave：应对大型多模态模型中的多图像、视频和三维内容。载于第十三届国际学习表征会议，ICLR 2025，新加坡，2025年4月24日至28日。OpenReview.net，2025年。1、7、S3、S22、S25、S28、S32[68] Geng Li 和 Yuxin Peng。DiCoBench：通过差异性和共性视觉线索对多图像细粒度感知进行基准测试。载于计算机视觉——ECCV 2026——第19届欧洲会议，瑞典马尔默，2026年9月8日至12日，论文集第XVI部分，第169–184页。施普林格，2026年。2、S1、S4、S34[69] Hongxing Li、Dingming Li、Zixuan Wang、Yuchen Yan、Hang Wu、Wenqi Zhang、Yongliang Shen、Weiming Lu、Jun Xiao 和 Yueting Zhuang。SpatialLadder：面向视觉语言模型空间推理的渐进式训练。载于国际学习表征会议，第76566–76592页，2026年。S22、S29[70] Jiachun Li、Shaoping Huang、Zhuoran Jin、Chenlong Zhang、Pengfei Cao、Yubo Chen、Kang Liu 和 Jun Zhao。MMR-Life：拼接真实生活场景，实现多模态多图像推理。载于第十四届国际学习表征会议，ICLR 2026，巴西里约热内卢，2026年4月23日至27日。proceedings.iclr.cc，2026年。2、S1、S4、S5、S34[71] 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。载于2023年自然语言处理经验方法会议论文集

处理，EMNLP 2023，新加坡，2023年12月6日至10日，第292–305页。计算语言学协会，2023年。8、S33[72] You Li、Heyu Huang、Chi Chen、Kaiyu Huang、Chao Huang、Zonghao Guo、Zhiyuan Liu、Jinan Xu、Yuhua Li、Ruixuan Li 和 Maosong Sun。Migician：揭示多模态大语言模型中自由形式多图像视觉定位的奥秘。载于计算语言学协会论文集，ACL 2025，奥地利维也纳，2025年7月27日至8月1日，第9845–9867页。计算语言学协会，2025年。S3、S22、S28、S29、S32[73] Zongxia Li、Wenhao Yu、Chengsong Huang、Zhenwen Liang、Rui Liu、Fuxiao Liu、Jingxi Chen、Dian Yu、Jordan Boyd-Graber、Haitao Mi 和 Dong Yu。Vision-SR1：通过推理分解和多奖励策略优化打造自奖励视觉语言模型。载于国际学习表征会议，第81753–81772页，2026年。S2、S22、S29[74] Hunter Lightman、Vineet Kosaraju、Yuri Burda、Harrison Edwards、Bowen Baker、Teddy Lee、Jan Leike、John Schulman、Ilya Sutskever 和 Karl Cobbe。让我们逐步验证。载于第十二届国际学习表征会议，ICLR 2024，11，2024 7 2024 2维也纳，奥地利，五月 - 。OpenReview.net，。S[75] Chin-Yew Lin。ROUGE：用于摘要自动评估的软件包。载于《文本摘要的拓展》，第74–81页。计算语言学协会，2004年。S2[76] Fangyu Liu、Emanuele Bugliarello、Edoardo Maria Ponti、Siva Reddy、Nigel Collier 和 Desmond Elliott。跨语言和文化的视觉落地推理。载于2021年自然语言处理实证方法会议论文集，EMNLP 2021，虚拟活动/多米尼加共和国蓬塔卡纳，2021年11月7日至11日，第10467–10485页。计算语言学协会，2021年。2、S1、S3、S4、S34[77] Haowei Liu、Xi Zhang、Haiyang Xu、Yaya Shi、Chaoya Jiang、Ming Yan、Ji Zhang、Fei Huang、Chunfeng Yuan、Bing Li 和 Weiming Hu。MIBench：评估多张图像场景下的多模态大语言模型。载于2024年自然语言处理实证方法会议论文集，EMNLP 2024，美国佛罗里达州迈阿密，2024年11月12日至16日，第22417–22428页。计算语言学协会，2024年。S1[78] Xiangyan Liu、Jinjie Ni、Zijian Wu、Chao Du、Longxu Dou、Haonan Wang、Tianyu Pang 和 Michael Shieh。NoisyRollout：通过数据增强强化视觉推理。载于神经信息处理系统进展，第2923–2957页。Curran Associates, Inc.，2025年。S2、S22、S28、S29、S32[79] Yuliang Liu、Zhang Li、Mingxin Huang、Biao Yang、Wenwen Yu、Chunyuan Li、Xu-Cheng Yin、Cheng-Lin Liu、Lianwen Jin 和 Xiang Bai。OCRBench：大型多模态模型中光学字符识别的隐秘之谜。《中国科学：信息科学》，67(12)，2024年。S33[80] Yuan Liu、Haodong Duan、Yuanhan Zhang、Bo Li、Songyang Zhang、Wangbo Zhao、Yike Yuan、Jiaqi Wang、Conghui He、Ziwei Liu、Kai Chen 和 Dahua Lin。MMBench：你的多模态模型是全能选手吗？载于计算机视觉——ECCV 2024——第十八届欧洲会议，意大利米兰，2024年9月29日至10月4日，论文集，第六部分，第216–233页。施普林格，2025年。3, S3[81] Yuqi Liu、屈天元、钟志胜、彭博皓、刘曙、于蓓和贾佳亚。VisionReasoner：通过强化学习实现统一的推理集成视觉感知。载于国际学习表征会议，第94069–94086页，2026年。S22, S29[82] 刘宇宏、张北辰、Yuhang Zang、Yuhang Cao、Long Xing、Xiaoyi Dong、Haodong Duan、Dahua Lin和Jiaqi Wang。Spatial-SSRL：通过自监督强化学习增强空间理解。载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第9570–9581页，2026年。S22, S29[83] Ziyu Liu、孙泽毅、Yuhang Zang、Xiaoyi Dong、Yuhang Cao、Haodong Duan、Dahua Lin和Jiaqi Wang。Visual-RFT：视觉强化微调。载于IEEE/CVF国际计算机视觉会议，ICCV 2025，美国夏威夷州檀香山，2025年10月19日至25日，第2034–2044页。IEEE，2025年。2, S2[84] Ilya Loshchilov和Frank Hutter。SGDR：带热重启的随机梯度下降。载于第五届国际学习表征会议，ICLR 2017，法国土伦，2017年4月24日至26日，会议论文集。OpenReview.net，2017年。S24[85] Ilya Loshchilov和Frank Hutter。解耦权重衰减正则化。载于第七届国际学习表征会议，ICLR 2019，美国路易斯安那州新奥尔良，2019年5月6日至9日。OpenReview.net，2019年。S23[86] 马无非、Yu-Cheng Chou、刘启昊、王星睿、Celso de Melo、谢建文和Alan Yuille。Spatial-Reasoner：迈向明确且可泛化的三维空间推理。载于神经信息处理系统进展，第140751–140774页。Curran Associates, Inc.，2025年。S22, S29[87] 马宇博、Yuhang Zang、陈良宇、陈美琪、焦一竹、李新泽、卢新源、Ziyu Liu、马岩、Xiaoyi Dong、Pan Zhang、潘良明、姜育刚、Jiaqi Wang、曹一新和孙爱新。MMLONGBENCH-DOC：通过可视化对长上下文文档理解进行基准测试。载于神经信息处理系统进展第37卷：2024年神经信息处理系统年度会议，NeurIPS 2024，加拿大不列颠哥伦比亚省温哥华，2024年12月10日至15日，2024年。2, S1, S4, S21, S23, S34[88] 孟凡清、王瑾、李传昊、卢全峰、田昊、杨天硕、廖嘉琦、朱西洲、戴继峰、乔宇、罗平、张凯鹏和邵文琪。MMIU：用于评估大型视觉语言模型的多模态多图像理解。载于第十三届国际学习表征会议，ICLR 2025，新加坡，2025年4月24日至28日。OpenReview.net，2025年。S1, S2, S29, S34[89] 孟家豪、李祥泰、王昊辰、谭悦、张涛、孔令栋、童云海、王安然、滕志扬、王玉璟和王卓辰。Open-o3-Video：

基于证据的视频推理，具有明确的时空证据。载于第43届国际机器学习会议论文集，第87958–87979页。PMLR，2026年。S22, S29[90] Thomas Mesnard、Theophane Weber、Fabio Viola、Shantanu Thakoor、Alaa Saade、Anna Harutyunyan、Will Dabney、Thomas S. Stepleton、Nicolas Heess、Arthur Guez、Eric Moulines、Marcus Hutter、Lars Buesing和R´emi Munos。模型无关强化学习中的反事实信用分配。载于第38届国际机器学习会议论文集，ICML 2021，2021年7月18日至24日 *2021,* 7654 7664 2021 2 虚拟活动，第–[91] 1 3 页。PMLR，。S Meta。介绍 Muse Spark。。https://research.meta.ai/blog/introducing-muse-spark-1-3，2026年。发布博客，2026年9月。7, S22, S23[92] Sewon Min、Eric Wallace、Sameer Singh、Matt Gardner、Hannaneh Hajishirzi和Luke Zettlemoyer。组合式问题并不必然需要多跳推理。载于计算语言学协会第57届年会（ACL）论文集，第4249–4257页，2019年。S1[93] Moonshot AI。介绍 Kimi K3.1 Pro。https://www.kimi.ai/blog/kimi-k3-1-pro，2026年。技术博客，2026年10月。模型卡片：https://huggingface.co/moonshotai/Kimi-K3.1-Pro。7, S22[94] Muse Glimmer团队。Muse Glimmer 30B：端侧智能体模型。https://arxiv.org/abs/2610.24628，2026年。技术报告，arXiv:2610.24628，2026年10月。模型卡片：https://huggingface.co/meta-models/Muse-Glimmer-30B。7, S22[95] Curtis G. Northcutt、Anish Athalye和Jonas Mueller。测试集中的普遍标签错误会使机器学习基准测试失稳。载于神经信息处理系统数据集与基准测试方向论文集第1卷，NeurIPS 数据集与基准，2021年12月，虚拟活动，2021年。S1[96] NVIDIA。Nemotron 3.5 Super和Lightning：面向智能体人工智能的开放、高效混合专家混合Mamba-Transformer模型。https://arxiv.org/abs/2610.38095，2026年。技术报告，arXiv:2610.38095，2026年10月。模型卡片：https://huggingface.co/nvidia/NVIDIA-Nemotron-3.5-Super-VL-120B-A12B-BF16。7, S22[97] Masanari Oi、Koki Maeda、Ryuto Koike、Daisuke Oba、Nakamasa Inoue和Naoaki Okazaki。从对应关系到动作：多模态大语言模型中的类人多图像空间推理。载于第43届国际机器学习会议论文集，第94383–94399页。PMLR，2026年。S3, S22, S28, S29, S32[98] OpenAI。介绍 ChatGPT。https://openai.com/index/chatgpt/，2022年。S2[99] OpenAI。GPT-5.6 系统卡片。https://deploymentsafety.openai.com/gpt-5-6，2026年。系统卡片，2026年7月。3, S3[100] OpenAI。GPT-6 Astra 系统卡片附录：GPT-6.1 Sol。https://deploymentsafety.openai.com/gpt-6-1-sol，2026年。系统卡片附录，2026年9月。7, S22, S23[101] OpenAI。《GPT-6 Astra系统卡》。https : / / deploymentsafety . openai . com / gpt - 6 - astra，2026年。系统卡，2026年9月。7，S22，S23[102] Dong Huk Park、Trevor Darrell和Anna Rohrbach。《鲁棒变化图注》。载于2019年IEEE/CVF国际计算机视觉会议（ICCV 2019），韩国首尔，2019年10月27日至11月2日，第4623–4632页。IEEE，2019年。S21，S25[103] Viorica Patraucean、Lucas Smaira、Ankush Gupta、Adri`a Re- casens、Larisa Markeeva、Dylan Banarse、Skanda Koppula、Joseph Heyward、Mateusz Malinowski、Yi Yang、Carl Do- ersch、Tatiana Matejovicova、Yury Sulsky、Antoine Miech、Alexandre Fr´echette、Hanna Klimczak、Raphael Koster、Jun- lin Zhang、Stephanie Winkler、Yusuf Aytar、Simon Osindero、Dima Damen、Andrew Zisserman和Jo˜ao Carreira。《感知测试：多模态视频模型的诊断基准》。载于《神经信息处理系统进展》第36卷：神经信息处理系统年度会议2023，NeurIPS 2023，美国路易斯安那州新奥尔良，2023年12月10日至16日，2023年。S25[104] Qwen团队。《Qwen3-VL技术报告》。CoRR，abs/2511.21631，2025年。2，7，S2，S21，S22，S28，S32[105] Qwen团队。《Qwen3.5：迈向原生多模态智能体》。https://qwen.ai/blog?id=qwen3.5，2026年。发布博客，2026年2月。模型卡片：https : / / huggingface.co/Qwen/Qwen3.5-9B。7，S21，S22，S24，S25，S28，S29，S31，S32，S33，S34[106] Qwen团队。《Qwen3.6-27B：旗舰级编程能力，搭载270亿参数稠密模型》。https://qwen.ai/blog?id=qwen3. 6- 27b，2026年。发布博客，2026年4月。模型卡片：https://huggingface.co/Qwen/Qwen3.6-27B。7，S22，S28，S32[107] Qwen团队。《Qwen3.8-Next架构设计：评估、效率与训练稳定性》。https: //arxiv.org/abs/2608.30320，2026年。技术报告，arXiv:2608.30320，2026年8月。发布博客：https://qwen.ai/blog?id=qwen3.8- flash- next。模型卡片：https://huggingface.co/ Qwen/Qwen3.8-Flash-Next。7，S21，S22[108] Qwen团队。《Qwen3.8-Max：编程与协同工作的新标杆》。https://qwen.ai/blog?id = 3 8, qwen . . 2026 2026 7, 22, 23 发布博客，8月。S S[109] Arijit Ray、Jiafei Duan、Ellis L. Brown, II、Reuben Tan、Dina Bashkirova、Rose Hendrix、Kiana Ehsani、Aniruddha Kemb- havi、Bryan A. Plummer、Ranjay Krishna、Kuo-Hao Zeng和Kate Saenko。《SAT：面向多模态语言模型的动态空间能力训练》。载于第二届语言建模会议，2025年。S25[110] Gabriel Sarch、Snigdha Saha、Naitik Khandelwal、Ayush Jain、Michael Tarr、Aviral Kumar和Katerina Fragkiadaki。《面向视觉推理的基于接地的强化学习》。载于《神经信息处理系统进展》，第150977–151013页。Curran Associates, Inc.，2025年。S2[111] John Schulman。《近似KL散度》。http://joschu.net/blog/kl-approx.html，2020年。博客文章，2020年3月。S24[112] John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Rad-

ford和Oleg Klimov。《近端策略优化算法》。CoRR, abs/1707.06347，2017年。S2[113] Ziyao Shangguan、Chuhan Li、Yuxuan Ding、Yanan Zheng、Yilun Zhao、Tesca Fitzgerald和Arman Cohan。《TOMATO：评估多模态基础模型的视觉时序推理能力》。载于第十三届国际学习表征会议，ICLR 2025，新加坡，2025年4月24日至28日。OpenReview.net，2025年。S2[114] Zhihong Shao、Peiyi Wang、Qihao Zhu、Runxin Xu、Junxiao Song、Mingchuan Zhang、Y. K. Li、Y. Wu和Daya Guo。《DeepSeekMath：推动开放语言模型数学推理能力的极限》。CoRR，abs/2402.03300，2024年。2，6，7，S2，S22，S25，S27[115] Mohit Shridhar、Jesse Thomason、Daniel Gordon、Yonatan Bisk、Winson Han、Roozbeh Mottaghi、Luke Zettlemoyer和Dieter Fox。《ALFRED：解读日常任务接地指令的基准测试》。载于2020年IEEE/CVF计算机视觉与模式识别会议，CVPR 2020，美国华盛顿州西雅图，2020年6月13日至19日，第10737–10746页。计算机视觉基金会 / IEEE，2020年。S4[116] Dingjie Song、Shunian Chen、Guiming Hardy Chen、Fei Yu、Xiang Wan和Benyou Wang。《MileBench：长上下文中的多模态大语言模型基准测试》。载于首届语言建模会议，2024年。2，S1，S2，S4，S5，S21，S34[117] Alane Suhr、Stephanie Zhou、Ally Zhang、Iris Zhang、Huajun Bai和Yoav Artzi。《基于照片进行自然语言推理的语料库》。载于计算语言学协会第57届会议论文集，ACL 2019，意大利佛罗伦萨，2019年7月28日至8月2日，第一卷：长篇论文，第6418–6428页。计算语言学协会，2019年。S1，S25[118] Hai-Long Sun、Zhun Sun、Houwen Peng和Han-Jia Ye。《通过随行视觉条件输入缓解多模态长链式思维推理中的视觉遗忘》。载于计算语言学协会第63届年会论文集（第一卷：长篇论文），第5158–5171页。计算语言学协会，2025年。S22，S29[119] Hao Tan、Franck Dernoncourt、Zhe Lin、Trung Bui和Mohit Bansal。《通过语言表达视觉关系》。载于计算语言学协会第57届会议论文集，ACL 2019，意大利佛罗伦萨，2019年7月28日至8月2日，第一卷：长篇论文，第1873–1883页。计算语言学协会，2019年。S4[120] Huajie Tan、Yuheng Ji、Xiaoshuai Hao、Xiansheng Chen、Pengwei Wang、Zhongyuan Wang和Shanghang Zhang。《Reason-RFT：面向视觉语言模型视觉推理的强化微调》。载于《神经信息处理系统进展》，第5772–5822页。Curran Associates, Inc.，2025年。S2[121] Xinyu Tian、Shu Zou、Zhaoyuan Yang和Jing Zhang。《识别并缓解多图像视觉语言模型的位置偏差》。载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第10599–10609页，2025年。2[122] Peter Tong、Ellis Brown、Penghao Wu、Sanghyun Woo、Adithya Iyer、Sai Charitha Akula、Shusheng Yang、Ji han Yang、Manoj Middepogu、Ziteng Wang、Xichen Pan、Rob Fergus、Yann LeCun 和 Saining Xie。Cambrian-1：对多模态大语言模型进行完全开放的视觉中心探索。载于《神经信息处理系统进展第37卷：2024年神经信息处理系统进展年度会议》，NeurIPS 2024，加拿大不列颠哥伦比亚省温哥华，2024年12月10日至15日，2024年。2，S1[123] Sang T. Truong、Yuheng Tu、Michael Hardy、Anka Reuel-Lamparth、Zeyu Tang、Jirayu Burapacheep、Jonathan Perera、Chibuike Uwakwe、Ben Domingue、Nick Haber 和 Sanmi Koyejo。人工智能基准测试中的奇妙漏洞及其寻找方法。载于《第38届神经信息处理系统进展：2025年神经信息处理系统进展年度会议》，NeurIPS 2025，美国加利福尼亚州圣迭戈，2025年12月2日至7日/墨西哥墨西哥城，2025年11月30日至12月5日，2025年。S1[124] Joshua Vendrow、Edward Vendrow、Sara Beery 和 Aleksander Madry。大语言模型基准测试是否检验可靠性？CoRR, abs/2502.03461，2025年。S1[125] Fei Wang、Xingyu Fu、James Y. Huang、Zekun Li、Qin Liu、Xiaogeng Liu、Mingyu Derek Ma、Nan Xu、Wenxuan Zhou、Kai Zhang、Tianyi Lorena Yan、Wenjie Jacky Mo、Hsiang-Hui Liu、Pan Lu、Chunyuan Li、Chaowei Xiao、Kai-Wei Chang、Dan Roth、Sheng Zhang、Hoifung Poon 和 Muhao Chen。MuirBench：用于稳健多图像理解的综合基准测试。载于《第十三届国际学习表征会议》，ICLR 2025，新加坡，2025年4月24日至28日。OpenReview.net，2025年。2，S1，S2，S4，S5，S34[126] Haozhe Wang、Chao Qu、Zuming Huang、Wei Chu、Fangzhen Lin 和 Wenhu Chen。VL-Rethinker：通过强化学习激励视觉语言模型进行自我反思。载于《神经信息处理系统进展》，第30865–30891页。Curran Associates, Inc.，2025年。S2[127] Jiaqi Wang、Kevin Qinghong Lin、James Cheng 和 Mike Zheng Shou。思考还是不思考？通过强化学习实现视觉语言模型的选择性推理。载于《神经信息处理系统进展》，第111196–111234页。Curran Associates, Inc.，2025年。S22，S29[128] Peng Wang、Shuai Bai、Sinan Tan、Shijie Wang、Zhihao Fan、Jinze Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge、Yang Fan、Kai Dang、Mengfei Du、Xuancheng Ren、Rui Men、Dayiheng Liu、Chang Zhou、Jingren Zhou 和 Junyang Lin。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。CoRR，abs/2409.12191，2024年。S22，S28，S29，S32[129] Peijie Wang、Zhong-Zhi Li、Fei Yin、Dekang Ran 和 Cheng-Lin Liu。MV-MATH：在多视觉上下文中评估多模态数学推理。载于 IEEE/CVF 计算机视觉与模式识别会议，CVPR 2025，美国田纳西州纳什维尔，2025年6月11日至15日，第19541–19551页。计算机视觉基金会/IEEE，2025年。2，S1，S2，S4，S23，S34[130] Qi Wang、Yanrui Yu、Ye Yuan、Rui Mao 和 Tianfei Zhou。VideoRFT：通过强化微调激励多模态大语言模型的视频推理能力。载于《神经信息处理系统进展》，

第4350–4376页。Curran Associates, Inc.，2025年。S22，S29[131] Qinsi Wang、Bo Liu、Tianyi Zhou、Jing S hi、Yueqian Lin、Yiran Chen、Hai Li、Kun Wan 和 Wentian Zhao。Vision-Zero：通过多智能体自博弈实现可扩展的视觉语言模型自我进化。载于《国际学习表征会议》，第106716–106744页，2026年。7，S22，S29[132] Qineng Wang、Baiqiao Yin、Pingyue Zhang、Jianshu Zhang、Kangrui Wang、Zihan Wang、Jieyu Zhang、Keshigeyan Chandrasegaran、Han Liu、Ranjay Krishna、Saining Xie、Jiajun Wu、Li Fei-Fei 和 Manling Li。从有限视角进行空间心理建模。载于《国际学习表征会议》，第82559–82622页，2026年。S3，S22，S25，S27，S28，S29，S32[133] Shenzhi Wang、Le Yu、Chang Gao、Chujie Zheng、Shixuan Liu、Rui Lu、Kai Dang、Xiong-Hui Chen、Jianxin Yang、Zhenru Zhang、Yuqiong Liu、An Yang、Andrew Zhao、Yang Yue、Shiji Song、Bowen Yu、Gao Huang 和 J unyang Lin。超越80/20法则：高熵少数令牌推动大语言模型推理的有效强化学习。载于《第38届神经信息处理系统进展：2025年神经信息处理系统进展年度会议》，NeurIPS 2025，美国加利福尼亚州圣迭戈，2025年12月2日至7日 5，2025年，2025年 30 2/ 墨西哥墨西哥城，11月至12月。S[134] Wenbin Wang、Liang Ding、Minyan Zeng、Xiabin Zhou、Li S hen、Yong Luo、Wei Yu 和 Dacheng Tao。分而治之再合并：用于多模态大语言模型高分辨率图像感知的免训练框架。载于第三十九届 AAAI 人工智能会议、第三十七届人工智能创新应用会议、第十五届人工智能教育进展研讨会，AAAI 2025，美国宾夕法尼亚州费城，2025年2月25日至3月4日，第7907–7915页。AAAI出版社，2025年。S33[135] Xiyao Wang、Zhengyuan Yang、Chao Feng、Hongjin Lu、Linjie Li、Chung-Ching Lin、Kevin Lin、Furong Huang 和 Lijuan Wang。以更少实现最先进水平：MCTS 引导的样本选择，用于数据高效的视觉推理自我改进。载于《神经信息处理系统进展》，第118818–118850页。Curran Associates, Inc.，2025年。7，S22，S29[136] Ye Wang、Ziheng Wang、Boshen Xu、Yang Du、Kejun Lin、Zihan Xiao、Zihao Yue、Jianzhong Ju、Liang Zhang、Dingyi Yang、Xiangnan Fang、Zewen He、Zhenbo Luo、Wenxuan Wang、Junqi Lin、Jian Luan 和 Qin Jin。Time-R1：用于时间视频定位的后训练大型视觉语言模型。载于《神经信息处理系统进展》，第83330–83364页。Curran Associates, Inc.，2025年。S22，S29[137] Zhenhailong Wang、Xuehang Guo、Sofia Stoica、Haiyang Xu、Hongru Wang、Hyeonjeong Ha、Xiusi Chen、Yangyi Chen、Ming Yan、Fei Huang 和 Heng Ji。面向感知的多模态推理策略优化。载于《第十四届国际学习表征会议》，ICLR 2026，27，2026年 23 2026年 2，22，27，28 巴西里约热内卢，4月至。proceedings.iclr.cc，。S S S S[138] Zengbin Wang、Feng Xiong、Liang Lin、Xuecai Hu、Yong Wang、Yanlin Wang、Man Zhang 和 Xiangxiang Chu。面向多模态推理的视觉引导策略优化。载于第64届计算语言学协会年会论文集（第1卷：长篇论文），第6646–6664页。计算语言学协会，2026年。2、7、S2、S27[139] Edwin B. Wilson。概率推断、继承律与统计推断。《美国统计学会期刊》，22(158)：209–212，1927年。S5[140] Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chaumond、Clement Delangue、Anthony Moi、Pierric Cistac、Tim Rault、Rémi Louf、Morgan Funtowicz、Joe Davison、Sam Shleifer、Patrick von Platen、Clara Ma、Yacine Jernite、Julien Plu、Canwen Xu、Teven Le Scao、Sylvai n Gugger、Mariama Drame、Quentin Lhoest 和 Alexander M. Rush。Transformer：自然语言处理的先进技术。载于2020年自然语言处理经验方法会议论文集：系统演示，EMNLP 2020 - 演示，线上，2020年11月16日至20日，第38–45页。计算语言学协会，2020年。S24[141] Bo Wu、Shoubin Yu、Zhenfang Chen、Josh Tenenbaum 和 Chuang Gan。STAR：真实世界视频情境推理基准。载于神经信息处理系统数据集与基准专题论文集第1卷，NeurIPS 数据集与基准 2021，2021年12月，线上，2021年。S25[142] Penghao Wu 和 Saining Xie。V*：引导式视觉搜索是多模态大语言模型的核心机制。载于IEEE/CVF计算机视觉与模式识别会议，CVPR 2024，美国华盛顿州西雅图，2024年6月16日至22日，第13084–13094页。IEEE，2024年。S33[143] Penghao Wu、Yushan Zhang、Haiwen Diao、Bo Li、Lewei Lu 和 Ziwei Liu。视觉拼图后训练提升多模态大语言模型。载于第十四届国际学习表征会议，ICLR 2026，巴西里约热内卢，2026年4月23日至27日。proceedings.iclr.cc，2026年。2、7、S3、S22、S28、S29、S32[144] Zhiyu Wu、Xiaokang Chen、Zizheng Pan、Xingchao Liu、Wen Liu、Damai Dai、Huazuo Gao、Yiyang Ma、Chengyue Wu、Bingxuan Wang、Zhenda Xie、Yu Wu、Kai Hu、Jiawei Wang、Yaofeng Sun、Yukun Li、Yishi Piao、Kang Guan、Aixin Liu、Xin Xie、Yuxiang You、Kai Dong、Xingkai Yu、Haowei Zhang、Liang Zhao、Yisong Wang 和 Chong Ruan。DeepSeek-VL2：用于高级多模态理解的混合专家视觉语言模型。CoRR，abs/2412.10302，2024年。3、S3[145] xAI。Grok-1.5 视觉预览版。https://x.ai/news/ grok-1.5v，2024年。介绍 Real-WorldQA 基准的发布博客，2024年4月 9，2026年 8，33 。访问日期：11月 。S[146] Tong Xiao、Xin Xu、Zhenya Huang、Hongyu Gao、Quan Liu、Qi Liu 和 Enhong Chen。Perception-R1：通过视觉感知奖励提升多模态大语言模型的多模态推理能力。载于国际学习表征会议，第26868–26898页，2026年。S2、S22、S29[147] 小米 LLM-Core 团队。MiMo-VL 技术报告。arXiv预印本 arXiv:2506.03569，2025年。2、7、S2、S21、S22、S28、S32

[148] 小米 LLM-Core 团队。MiMo-V2.6：通过扩展强化学习实现自我改进。https: //arxiv.org/abs/2610.11959，2026年。技术报告，arXiv:2610.11959，2026年10月。模型卡片：https://huggingface.co/XiaomiMiMo/MiMo- V2.6-Pro-RL。7、S22[149] Long Xing、Xiaoyi Dong、Yuhang Zang、Yuhang Cao、Jianze Liang、Qidong Huang、Jiaqi Wang、Feng Wu 和 Dahua Lin。CapRL：通过强化学习激发密集图像描述能力。载于国际学习表征会议，第13066–13093页，2026年。S22、S29[150] Zihui (Sherry) Xue、Romy Luo 和 Kristen Grauman。大型多模态模型中的时间之箭感知。载于神经信息处理系统进展，第90925–90955页。Curran Associates, Inc.，2025年。S22、S29[151] Sihan Yang、Runsen Xu、Yiman Xie、Sizhe Yang、Mo Li、Jingli Lin、Chenming Zhu、Xiaochen Chen、Haodong Duan、Xiangyu Yue、Dahua Lin、Tai Wang 和 Jiangmiao Pang。MMSI-Bench：多图像空间智能基准。载于第十四届国际学习表征会议，ICLR 2026，巴西里约热内卢，2026年4月23日至27日。proceedings.iclr.cc，2026年。2、S1、S2、S4、S5、S34[152] Huanjin Yao、Jiaxing Huang、Wenhao Wu、Jingyi Zhang、Yibo Wang、Shunyu Liu、Yingjie Wang、YuXin Song、Haocheng Feng、Li Shen 和 Dacheng Tao。Mulberry：通过集体蒙特卡洛树搜索，赋予多模态大语言模型类 o1 推理与反思能力。载于神经信息处理系统进展，第29918–29952页。Curran Associates, Inc.，2025年。S22、S29[153] Huanjin Yao、Qixiang Yin、Jingyi Zhang、Min Yang、Yibo Wang、Wenhao Wu、Fei Su、Li Shen、Minghui Qiu、Dacheng Tao 和 Jiaxing Huang。R1-ShareVL：通过 Share-GRPO 激励多模态大型语言模型的推理能力。载于神经信息处理系统进展，第80424–80451页。Curran Associates, Inc.，2025年。7、S22、S28、S29、S32[154] Jiabo Ye、Haiyang Xu、Haowei Liu、Anwen Hu、Ming Yan、Qi Qian、Ji Zhang、Fei Huang 和 Jingren Zhou。mPLUG-Owl3：面向多模态大语言模型的长图像序列理解。载于国际学习表征会议，第98891–98913页，2025年。S22、S28、S32[155] Xiaojun Ye、Guanbao Liang、Chun Wang、Liangcheng Li、Pengfei Ke、Rui Wang、Bingxin Jia、Gang Huang、Qiao Sun 和 Sheng Zhou。M4Bench：面向多模态大语言模型的多领域、多粒度、多图像理解基准。载于第三十四届国际人工智能联合会议论文集，IJCAI 2025，2025年22日，16 6848 6856 2025年 2，1，4，34 加拿大蒙特利尔，8月 - 第 – 页。ijcai.org，。S S S[156] Kexin Yi、Chuang Gan、Yunzhu Li、Pushmeet Kohli、Jiajun Wu、Antonio Torralba 和 Joshua B. Tenenbaum。CLEVRER：用于视频表征与推理的碰撞事件。载于第八届国际学习表征会议，ICLR 2020，埃塞俄比亚亚的斯亚贝巴，2020年4月26日至30日。OpenReview.net，2020年。S25

[157] Qiying Yu、郑章、Ruofei Zhu、Yufeng Yuan、Xiaochen Zuo、Yu Yue、Weinan Dai、Tiantian Fan、Gaohong Liu、Juncai Liu、Lingjun Liu、Xin Liu、Haibin Lin、Zhiqi Lin、Bole Ma、Guangming Sheng、Yuxuan Tong、Chi Zhang、Mofan Zhang、Ru Zhang、Wang Zhang、Hang Zhu、Jinhua Zhu、Jiaze Chen、Jiangjie Chen、Chengyi Wang、Hongli Yu、Yuxuan Song、Xiangpeng Wei、Hao Zhou、Jingjing Liu、Wei- Ying Ma、Ya-Qin Zhang、Lin Yan、Yonghui Wu 和 Mingx- uan Wang。《DAPO：大规模开源大语言模型强化学习系统》。载于《神经信息处理系统进展》第38卷：2025年神经信息处理年会，NeurIPS 2025，美国加利福尼亚州圣迭戈，2025年12月2日至7日 / 墨西哥墨西哥城，2025年11月30日至12月5日，2025年。2、6、7、S2、S27[158] Zhangyuan Yu、Wanran Sun、Guangjing Yang、Xiaohu Wu 和 Qicheng Lao。《CFPO：用于多模态推理的反事实策略优化》。载于第43届国际机器学习会议论文集，第151070–151086页。PMLR，2026年。2、7、S2、S27[159] Yifu Yuan、Haiqin Cui、Yaoting Huang、Yibin Chen、Fei Ni、Zibin Dong、Pengyi Li、Yan Zheng、Hongyao Tang 和 Jianye Hao。《Embodied-R1：面向通用机器人操作的强化具身推理》。载于《国际学习表征会议》，第8984–9009页，2026年。S22、S29[160] Xiang Yue、Tianyu Zheng、Yuansheng Ni、Yubo Wang、Kai Zhang、Shengbang Tong、Yuxuan Sun、Botao Yu、Ge Zhang、Huan Sun、Yu Su、Wenhu Chen 和 Graham Neubig。《MMMU-Pro：更稳健的多学科多模态理解基准测试》。载于《第63届计算语言学协会年会论文集》（第1卷：长篇论文），ACL 2025，奥地利维也纳，2025年7月27日至8月1日，第15134–15186页。计算语言学协会，2025年。S1[161] Z.ai。《GLM-5.3-Flash》。https://huggingface.co/zai-org/GLM-5.3-Flash，2026年。模型卡片，2026年8月。7、S22[162] Jianshu Zhang、Dongyu Yao、Renjie Pi、Paul Pu Liang 和 Yi R. Fung。《VLM2-Bench：深入考察视觉语言模型在多大程度上能隐式关联明确匹配的视觉线索》。载于《第63届计算语言学协会年会论文集》（第1卷：长篇论文），ACL 2025，奥地利维也纳，2025年7月27日至8月1日，第7510–7545页。计算语言学协会，2025年。2、S1、S4、S23、S34[163] Kai Zhang、Lingbo Mo、Wenhu Chen、Huan Sun 和 Yu Su。《MagicBrush：用于指令引导图像编辑的人工标注数据集》。载于《神经信息处理系统进展》第36卷：2023年神经信息处理年会，NeurIPS 2023，美国路易斯安那州新奥尔良，2023年12月10日至16日，2023年。S25[164] Mingtian Zhang。《MMLongBench-Doc-V2：MMLongBench-Doc的修正标注、语义感知修订版》。CoRR，abs/2608.03397，2026年。S1[165] Yizhen Zhang、Yang Ding、Shuoshuo Zhang、Xinchen Zhang、Haoling Li、Zhong-Zhi Li、Peijie Wang、Jie Wu、Lei Ji、Yeyun Gong、Yelong Shen 和 Yujiu Yang。《PeRL：

用于交错式视觉-语言推理的置换增强强化学习。载于《神经信息处理系统进展》第38卷：2025年神经信息处理年会，NeurIPS 2025，美国加利福尼亚州圣迭戈，2025年12月2日至7日 / 墨西哥墨西哥城，2025年11月30日至12月5日，2025年。7、S22、S28、S29[166] Yifan Zhang、Huanyu Zhang、Haochen Tian、Chaoyou Fu、Shuangqing Zhang、Junfei Wu、Feng Li、Kun Wang、Qing- song Wen、张章、王亮和金榕。《MME-RealWorld：你的多模态大语言模型能否应对对人类而言难度极高的高分辨率真实世界场景？》。载于《第十三届国际学习表征会议》，ICLR 2025，新加坡，2025年4月24日至28日。OpenReview.net，2025年。8、S33[167] Zicheng Zhang、Haoning Wu、Erli Zhang、Guangtao Zhai 和 Weisi Lin。《Q-Bench⁺：面向多模态基础模型的低层视觉基准测试，从单张图像到图像对》。IEEE模式分析与机器智能汇刊，46(12)：10404–10418，2024年。2、S1、S4、S5、S34[168] Bingchen Zhao、宗永硕、张乐天和 Timothy M. Hospedales。《视觉与语言模型多图像理解基准测试：感知、知识、推理与多跳推理》。CoRR，abs/2406.12742，2024年。2、S1、S2、S4、S23、S34[169] 郑立豪、陈佳伟、沈新天、马昊和韦涛。《MIRG-RL：通过强化学习进行多图像推理与定位》。载于 ICASSP 2026——2026年 IEEE声学、语音与信号处理国际会议，第12292–12296页。IEEE，2026年。2、S3[170] 朱金国、王伟云、陈哲、刘朝阳、Shenglong Ye、顾立新、田昊、段宇辰、苏伟杰、邵杰、高章伟、崔尔飞、王学辉、曹越、刘扬州、魏星光、张宏杰、王浩敏、徐伟业、李昊、王家豪、邓念辰、李松泽、何一南、蒋坦、罗佳鹏、王毅、Conghui He、石博天、张星成、邵文琪、何军军、熊英彤、屈文文、孙鹏、Penglong Jiao、吕涵、吴立军、张凯鹏、邓辉鹏、葛家业、Kai Chen、王利民、窦敏、Lewei Lu、朱西洲、卢彤、Dahua Lin、乔宇、戴继峰和王文海。《InternVL3：探索开源多模态模型的高级训练与测试时方案》。CoRR，abs/2504.10479，2025年。7、S21、S22、S28、S32

ASTERISM：多模态大语言模型真的会查看所有图像吗 ？

补充材料

目录

A. 相关工作 S1
A.1. 多图像基准 .
S1A.2. 基准审计和捷径 .S1A.3. 面向大语言模型和多模态大语言模型的可验证奖励强化学习 . S2A.4. 令牌级信用分配 .S2 A.5. 面向视觉证据的强化学习 .S2A.6. 多图像输入训练 .S3

B. ASTERISM 的构建 S3
B.1. 审计协议 .
S3B.2. 按来源计数条目 .S3B.3. 人工验证 .S5B.4. 任务类别 .S5

C. 示例条目 S6
C.1. 删除和更正的条目 .
S6C.2. 改写条目 .S6C.3. 按任务类别划分的项目 .S13

D. ASTERISM-Lite S21
D.1. 选择 .
S21D.2. 验证 .S21D.3. 排名一致性和配对区间 . . . . . S21 D.4. 专有多模态大语言模型 .S21

E. 评分与评估 S21
E.1. 基于规则的评分器 .
S21E.2. 官方评分代码 .S21 E.3. 评估设置 .S21

F. CICA 的完整规范 S21
F.1. CICA 目标 .S22 F.2. CICA 的单图像输入 .S23 F.3. CICA 中的答案词元 .S23F.4. 活跃令牌与 CICA 权重 .S24 F.5. CICA 对比项的选择 .S24F.6. 真实响应上的 CICA 权重 . S25 F.7. CICA 成本 .S25

G. 训练细节 S25
G.1. 骨干模型 .
S25G.2.训练数据 .S25 G.3.强化学习基线 .S27G.4.消融实验 .S27

H. 补充结果 S27
H.1. 评估多模态大语言模型 .
S27H.2. 单图恢复与错误答案 . . . . . S27 H.3. 使用 CICA 训练 .S29 H.4. 单图基准测试 .S29 H.5.MMIU . S29H.6. 训练后的捷径 .S30 H.7. 增益的可靠性 .S30

I. 局限性与未来工作 S30

J. 数据卡 S31
J.1. 许可证与发布 .
S31J.2. 组成与预期用途 .S32

A. 相关工作

A.1. 多图像基准

多图评测最初用于验证关于两张照片的陈述，例如 NLVR2 及其多语言、多文化版本 MaRVL [76, 117], , 如今已发展为测试多种跨图像技能的基准套件。Mantis-Eval、MuirBench、MMIU、MIBench、MIRB 和 M4Bench 涵盖图像间的比较、排序、对应、计数和多跳推理 [49, 77, 88, 125, 155, 168], 。MileBench 测试较长的多图像上下文 [116], , ReMI、MMRB 和 MMR-Life 则针对多张图像上的推理 [19, 52, 70]。其他基准侧重于某一类证据，例如 MMSI-Bench 中不同视图之间的空间关系 [151], 、MV-MATH 中对多个图形进行数学运算 [129]、VLM2-Bench 中视觉线索的匹配 [162], 、DiCoBench 中差异与共性 [68] ， 以及 Q-Bench 中低层次的视觉比较2 [167]。视觉感知、多模态推理、检索增强式问答和多页文档基准中也包含多图像条目 [31, 42, 44, 60, 87]。ASTERISM 从其中 19 个基准中收集多图像条目，并在有官方说明的情况下保留这些说明。一种审计协议会针对每个条目分别仅使用文本和单张图像进行测试；ASTERISM 保留那些需要多张图像的条目，并使用同一个基于规则的评分器为所有条目评分（表 S1）。

A.2. 基准审计与捷径

对测试集的审计发现，标签错误会改变模型排名 [95], ；对语言和文档基准测试的重新标注则发现了错误标签，以及含糊不清或无法作答的问题 [33, 124, 164]。模型响应的统计数据可以标记此类条目，以供专家审查 [123]。在视觉问答中，仅凭文本就能得出答案的问题，促使研究者构建了将相似图像与不同答案配对，或改变训练集与测试集之间答案先验的基准测试 [2, 37], , 并使用反事实样本进行训练 [16]。许多 MLLM 基准中的条目也可以在不看图像的情况下作答 [17, 122],；而 MMStar、MMMU-Pro 和 NaturalBench 则会选择、筛选或配对条目，使作答必须依赖图像 [17, 64, 160]。在多跳阅读理解中，一次阅读一个段落或一个句子的模型，能够回答许多本应需要阅读两个段落才能作答的问题 [15, 92],；而提供

表 S1. **ASTERISM 逐条审计公开基准测试**。条目遵循各基准论文中的说明（n/r，即未报告），而表 S2 统计了我们从各来源公开发布内容中纳入样本池的条目数量。

基准测试	项目来源	项目核查	评分
BLINK [31]	3,807 条计算机视觉数据 + new	作者弃用含糊不清 ones	规则 + GPT-3.5 [98]
MuirBench [125]	2,600 7 个数据集 + new	4 位专家，每个条目	规则
MMIU [88]	11,698 个现有数据集	转换后的数据已检查	选项匹配
MileBench [116]	6,440 个，21 个数据集	10% 样本；所有新集合	准确率，ROUGE-L [75]
MIRB [168]	925 个新 + 3 数据集	图像筛选器，1 task	准确率
Mantis-Eval [49]	217 个新	标注员-精选	n/r
MMSI-Bench [151]	1,000 新版	3 位审阅者，每个条目	精确；大语言模型回退机制
MV-MATH [129]	2,009 个新增（K-12教育）	≥2标注员 大语言模型评审	
ASTERISM	18,696 19 基准测试	2 个多模态大语言模型，每个条目，人工裁决；4 位审阅者：重新标注，改写；1,000 按两个	规则，没有 LLM

同一捷径也能得出第二个答案，这会降低强大模型的准确率 [50]。视频基准测试中也存在类似现象：单帧图像，或顺序错误的帧，往往就足以作答 [10, 62, 113]。CVBench 和 SSMNBench 构建了针对人类场景的多摄像头基准测试，其中的条目需要结合多个视角作答，并分析了依赖单一视角的多模态大语言模型所犯的错误 [40, 41]。ASTERISM 将这些测试应用于 19 个公开基准测试中的多图像条目。它会对从中抽取的每个条目进行检查，判断仅凭文本或单张图像是否就能作答，同时进行标签、含糊性、重复项和评分检查（补充材料B.1）。CICA 在训练期间将相同的单图比较应用于响应中的词元。

A.3. 面向大语言模型和多模态大语言模型的可验证奖励强化学习

可验证奖励强化学习用检查最终答案的规则（例如精确匹配或测试用例）替代学习得到的奖励模型。T¨ulu 3 命名了这一方法，并将其应用于数学和指令遵循 [59], 而 DeepSeekMath 引入了 GRPO，该方法对同一提示的响应组内奖励进行归一化，且不需要价值模型 [114]。DeepSeek-R1 和 Kimi k1.5 将带有答案检查奖励的强化学习扩展到数学和编程中的长思维链推理 [39, 55]。对于多模态大语言模型，可验证奖励强化学习提升了视觉感知 [83]和多模态推理 [46, 120, 126], , 近期的多模态大语言模型在后训练中加入了强化学习阶段，并经常结合来自偏好模型的奖励 [11, 43, 56, 104, 147]。奖励会检查最终答案，而 GRPO 和 DAPO [157] 会为响应中的每个词元赋予相同的优势，因此更新不会区分从一张图像中读取出的答案与结合多张图像得出的答案。CICA 保留这一奖励，并改变其在词元之间的分配方式（第4节）。

A.4. 令牌级信用分配

每个响应只有一个奖励时，功劳分配决定每个词元获得多少奖励。近端策略优化学习一个价值函数，为每个词元赋予各自的优势 [112], ; 而 VinePPO 则用从中间步骤重新开始的采样轨迹所得的蒙特卡洛估计取代价值函数 [53]。过程奖励模型通过人工步骤标签 [74] 或仅通过结果标签 [21], 进行训练，用于为中间步骤评分。在基于人类反馈的强化学习中，奖励模型的注意力权重可以将序列奖励分摊到各个词元上 [12], ; 而在可验证奖励强化学习中，将更新限制在响应中的高熵词元上，其效果不亚于甚至优于更新所有词元 [133]。反事实功劳分配通过将某个动作与替代方案进行比较，衡量该动作的贡献，例如在多智能体评论器中对某个智能体的动作进行边缘化 [30]，或根据关于轨迹的事后信息对基线进行条件化 [90]。CICA 从输入图像中构造反事实。它保持采样得到的响应不变，改变策略所看到的图像，并比较每个词元在所有图像条件下的似然与在最佳单张图像条件下的似然，因此无需价值模型、奖励模型或额外的轨迹采样。

A.5. 面向视觉证据的强化学习

若干方法改变多模态大语言模型的可验证奖励强化学习更新，使其依赖于视觉输入。VGPO 通过相似度定位视觉线索，并将优势重新加权，使其偏向视觉激活较高的词元，以及偏向累积更多视觉激活的采样轨迹 [138], ; APPO 则为在一组响应中关注同一关键视频帧的词元提供词元级奖励 [24]。另一种方法会在文本丰富的推理中屏蔽视觉定位错误的词元的梯度 [20]。VPPO 根据每个响应的词元相对于随机遮挡图像所测得的平均图像依赖度，对该响应的优势进行缩放，并且只更新依赖度最高的词元 [45]。CFPO 通过正则化策略，使其不同于在关键视觉线索被抑制的反事实状态下所作出的预测 [158], ; PAPO 增加一项损失，使策略针对原始图像的输出偏离其针对遮挡图像的输出 [137], ; NoisyRollout 则混合来自干净图像和失真图像的采样轨迹 [78]。其他方法训练模型将其推理步骤锚定到图像坐标或边界框 [28, 110], , 或者直接奖励感知：通过将响应中的视觉内容与参考标注进行比较来判断 [146] , 或检查自生成的视觉描述是否足以回答问题 [73]。反事实视觉输入也出现在强化学习之外，例如用于视觉问答训练的样本中，后者

遮蔽关键物体 [16]。CICA 在多图像输入的图像上定义其反事实状态。它利用这一反事实状态重新加权 GRPO 的优势，无需额外的损失或奖励。

A.6. 多图像输入的训练

交错式指令数据（如 Mantis-Instruct 和 M4-Instruct）教会多模态大语言模型接收多张图像 [49, 67],Migician 使用大型指令集训练跨图像视觉定位 [72], 而 Das 等人 [22]将单图像标注组合成多图像训练示例，并对多图像输入应用注意力掩码。强化学习方法加入带有可验证答案的多图像任务，例如区分同一图像的两个增强视图与一张相似图像 [18], 在图像、视频和三维数据上完成拼图 [143], 跨图像定位物体 [8, 169] 以及根据多个视图回答空间问题 [97, 132]。CICA 不添加任务。它使用带有各自可验证答案的多图像条目进行训练，只改变响应的奖励如何分配给其中的词元。

B. ASTERISM 的构建

B.1. 审计协议

判定结果。 每位审计员从以下八种判定结果中选择一种；若同时适用多种，则以列表顺序最靠前者为准。标签错误 指根据图像和问题，真实标注有误，且审计员能够给出正确答案。无法回答 指无法根据输入得出任何答案。评分错误 指标签正确，但官方评分器拒绝了符合所要求格式的正确答案。含糊不清 指有两个或更多答案均有合理依据。重复项 指该条目重复了另一条目的图像和问题。可由文本回答 指审计员仅凭文本就能有把握地确定答案，且该答案与真实标注一致。单图像足以回答 指仅凭一张图像就能回答问题，且审计员指出了该图像。有效 指以上情况均不适用。

流程。 两个多模态大语言模型 Claude Fable 5 [5] 和 GPT- 5.6-Sol [99], 独立标注每个经审计的条目。它们对每个条目作答两次：第一次仅依据问题和选项，第二次则结合带编号的图像；并为每个条目写下一项判定、两次答案、任何足以作答的图像、置信度评级和简短理由。两个多模态大语言模型在 92.2% 的情况下，对保留还是移除某个条目的意见一致。若判定不同，人工裁定者会对照图像阅读两份判定、答案和理由，并作出最终决定。升级处理的判定（指出标签或条目存在缺陷的判定，以及置信度较低但认为有效的判定）也会交由人工裁定者处理。每项标签修正都会接受人工审核（补充材料B.3）。

单图筛选器。 单图筛选器的两个阶段都使用一个由来自四个开源多模态大语言模型模型家族的 Phi-4-多模态指令 [1], DeepSeek-VL2-Small [144], Kimi-VL-A3B-Instruct [56] 和 Molmo- 70亿参数-O [23]组成的委员会。第一阶段在审核前对源条目运行，并在人工审核前对每个改写运行；第二阶段在标签、格式和选项顺序最终确定后运行一次，涵盖所有项目，包括改写。该委员会与我们训练的骨干模型不共享任何模型家族，也不与排行榜共享任何模型，因此没有任何参与评估的模型参与筛选。对于一个项目的每个图像，每位成员仅根据该图像作答，并按照选项的循环轮换顺序各进行一次展示， C如 CircularEval 中所示 [80]。在每次展示中，成员会收到该项目的评估提示，其中仅包含该图像，并以贪心方式进行解码；随后，补充材料E.1中的基于规则的评分器会检查其答案。对于选项涉及其他选项的项目，例如“以上都不是”，保留其原始顺序，并向每位成员展示一次。如果对于某个图像，每位成员在每次展示中都回答正确，并且委员会正确作答次数相对于随机机会率 $1/C$ 的单侧二项检验达到显著，则移除该项目。该检验会针对项目的图像进行邦弗朗尼校正，并在该阶段的所有项目上以 5% 的假发现率进行 Benjamini–Hochberg 控制 [9]。对于没有选项的项目，随机机会率为零，因此四位成员都必须根据同一图像正确作答。

B.2. 每个来源的条目数

表S2列出了全部 32,688 个源条目。每个来源均纳入其公开评估数据中带标签的条目，但 BLINK 纳入其验证集中的多图像条目，EMMA 纳入其多图像条目，MV-MATH 纳入其多项选择题，VLM2-Bench 纳入其图像条目，MileBench 则不纳入其 GPR1200 任务；此外，MaRVL [76] 和 DUDE [60]各纳入随机抽取的 1,000 个条目；MaRVL 的样本按语言和标签分层抽取，DUDE 的样本则按答案类型从其验证集中分层抽取。该集合还排除了 573 个图像出现在公开训练数据中的条目（MMRB 的 249 个、MuirBench 的 239 个、MIRB 的 34 个、MMR-Life 的 27 个、Q-Bench2 的 16 个、MRAG-Bench 的 4 个，以及 VLM2-Bench 和 MileBench 各 2 个）、359 个与其他来源条目共用图像文件的 MMRB 条目、9 个因我们所持副本缺少文档而无法纳入的 MMLongBench-Doc 条目、2 个因发布内容中缺少图像而无法纳入的条目（MMRB 和 MMR-Life 各 1 个），以及 4 行标识符与其他行重复的 VLM2-Bench 数据。DUDE 的单页条目随其样本一并纳入，并根据一条规则予以移除。

单图筛选器的第一阶段在审计前移除 545 个项目。在经过审计的 32,143 个项目中，首次审核轮次的判定结果移除 8,857 个项目（有一张图像即可 6,434 个、文本即可 1,007 个、含糊不清或无法回答 1,233 个、重复项 170 个、由裁决员移除 13 个），规则移除 1,333 个项目（表的图注S2 列出了这些规则）。首次审核轮次纠正了 250 个项目的标签，并为 350 个存在评分错误的项目提供了一个能够接受

表 S2. **每个源条目的去向**。对于每个源基准，表中各列分别统计：条目总数（项目）、审核前被单图筛选器移除的条目（筛选器）、已审核条目（已审核）、保留在除检索类别外六个任务类别中的条目（主要）或检索类别中的条目（检索）、根据首轮审核判定移除的条目（移除：判定）或根据规则移除的条目（移除：规则）、未能通过转换、转换审核或首轮筛选阶段的自由形式条目，以及副本、模糊题项和被其改写条目取代的源条目（移除：自由形式）、第二轮审核移除的条目（移除：第二轮）、因一张图像即可作答而在单图筛选器第二阶段移除的条目（单图），以及源条目衍生、通过人工审核和最终步骤的改写条目（R）。每行均满足：项目 = 主要 + 检索 + 移除：判定 + 移除：规则 + 移除：自由形式 + 移除：第二轮 + 1-图像 + 筛选器。R 条目是新增条目，不计入此总和。主要类别包含 MileBench 中转换后的 83 个自由形式条目。规则会移除 DUDE 中的单页条目、其他带有自由形式答案的条目、控制任务，以及一个标签未通过审核的子任务。最后三列给出了该行条目中仅凭一张图像或问题本身即可得出答案的百分比；这些比例分别根据两轮审核的判定结果（审核）、加入首轮单图筛选器移除的条目后（+筛选器），以及进一步加入单图条目后（+1-图像）计算。检索条目和多页文档条目（其任务是找到包含答案的图像或页面），以及其他来源的条目副本，均不计入统计。短横线表示零。

												一张图像或文本（%）		
来源	项目	筛选器	已审计	Main	保留项	删除项	判定	删除项	规则	删除项	自由	删除项	第二个	单图像	R 审计	+筛选器	+1-图像
MileBench [116]	5,838	2	5,836	1,508	318	2,522	260	519	382	327	218	34.3	34.3	37.0			
MMRB [19]	4,141	112	4,029	1,069	191	1,373	68	249	955	124	–	21.6	23.4	26.2			
VLM2-Bench [162]	2,854	37	2,817	1,728	–	521	–	29	381	158	82	29.2	30.5	36.0			
MMR-Life [70]	2,618	81	2,537	2,093	–	394	1	30	16	3	158	9.0	12.1	12.2			
ReMI [52]	2,600	–	2,600	1,956	–	213	–	259	153	19	273	13.3	13.3	14.0			
MuirBench [125]	2,361	121	2,240	979	226	710	227	22	15	61	–	22.4	26.3	27.8			
MRAG-Bench [44]	1,349	83	1,266	–	291	947	4	–	5	19	–	0.2	0.2	0.2			
M4Bench [155]	1,314	–	1,314	865	–	150	230	3	4	62	–	11.0	11.0	15.7			
MV-MATH [129]	1,109	–	1,109	857	–	206	1	39	5	1	36	17.3	17.3	17.4			
BLINK [31]	1,108	–	1,108	879	–	62	–	141	1	25	145	5.4	5.4	7.7			
MMLongBench-Doc [87]	1,082	–	1,082	–	963	91	1	24	–	3	–	0.8	0.8	0.8			
DUDE [60]	1,000	–	1,000	–	334	93	529	35	–	9	–	3.3	3.3	3.3			
MaRVL [76]	1,000	–	1,000	316	–	522	–	120	–	42	147	51.6	51.6	55.8			
MMSI-基准 [151]	1,000	–	1,000	928	–	19	1	37	12	3	23	2.1	2.1	2.4			
Q-Bench2 [167]	984	64	920	391	–	405	2	33	1	88	102	38.6	45.1	54.1			
MIRB [168]	935	45	890	432	4	345	2	50	34	23	177	33.6	38.4	40.9			
DiCoBench [68]	765	–	765	547	–	119	–	46	32	21	22	6.4	6.4	9.2			
EMMA [42]	413	–	413	309	–	96	4	4	–	–	–	22.8	22.8	22.8			
Mantis-Eval [49]	217	–	217	82	24	69	3	9	3	27	23	31.8	31.8	44.2			
总计	32,688	545	32,143	14,939	2,351	8,857	1,333	1,649	1,999	1,015	1,406	20.6	21.8	24.2			

所请求形式的答案格式，该格式在项目中设定，或由评分器的规则设定。另外，1,092 个项目包含自由形式答案、复制其他来源的保留项目，或含糊不清。

Spot-the-Diff 的自由形式条目 [47], IEdit [119]和 ALFRED [115]任务（属于 MileBench） [116]在图像支持一个可验证答案时，会被转换为多项选择题。一个多模态大语言模型会为每个此类条目编写四个选项：一个正确选项，说明参考答案中经图像验证的一项变更、编辑或下一步，以及三个被图像证明为错误的干扰项。另一个模型家族的多模态大语言模型会根据补充材料B.1中的判定结果，对每个转换后的条目进行审计；只有通过此次独立审核的转换条目才会进入 ASTERISM。在 1,092 个条目中，404 个为自由形式条目，其中 91 个通过了转换、独立审核和单图筛选器的第一阶段；其中 83 个还通过了后续检查并进入 ASTERISM。未通过前三步的 313 个条目，以及 688 个模糊题项和副本，共计 1,001 个条目，均已移除。

Claude Fable 5 会改写项目，其中 1,546 条改写通过了 GPT-5.6-Sol 检查和单图筛选器的第一阶段。经过人工审核和自动检查后，我们移除了 1,546 条改写中未通过任一项检查的 95 条（见第 3.1 节）。

第二轮定向审核将补充材料B.1 中的判定结果和流程应用于人工审核员标记的各类条目。该轮移除 2,008 个条目（1,999 个源条目和 9 个改写），其中包括 805 个复制了另一来源保留条目的 MMRB 条目，并修正 ASTERISM 保留的 86 个条目的标签，以及其中 321 个源条目的问题或选项。单图筛选器的第二阶段在标签、格式和选项顺序最终确定后运行，移除可由一张图像回答的 1,038 个条目（1,015 个源条目，即表S2中的 1-img，以及 23 个改写）。最后，为避免源条目与其改写同时进入，我们移除改写进入 ASTERISM 的 648 个源条目（表S2中“移除自由形式”部分），并删除源条目属于 ASTERISM-Lite 的 13 个改写。随后，ASTERISM 在七类任务中保留 18,696 个条目，其中包括 1,406 个改写和检索类别的 2,351 个条目。在所有步骤中，单图筛选器的两个阶段共移除 1,560 个源条目，两轮审核移除 10,856 个，规则移除 1,333 个，处理自由形式条目、副本和模糊题项移除 1,001 个，以改写替换源条目移除 648 个，因此 32,688 个源条目中有 17,290 个进入 ASTERISM。在之后的

表 S3。**任务类别**。每个类别涵盖的操作、一个来自 ASTERISM 的示例问题（来源、图像数量 K）以及 ASTERISM 条目的数量，包括每个类别的改写条目。

类别	操作	示例	项目
聚合	对图像中的值进行计数或合并	“图像中总共有多少个瓶子？”（MuirBench [125], $K=2$）	1,055
比较/变化	判断图像中的某项属性，或找出对齐视图之间的差异	“与第一张图像相比，第二张图像的清晰度如何？”（Q-Bench2 [167], $K=2$）	2,262
推理/验证	整合图形、图表或示意图中的数量，或验证关于一组内容的陈述	“以下图像展示了两个图……这两个图是否同构？”（ReMI [52], $K=2$）	4,793
空间对应关联三维中的相机视图，或匹配图像的	图像中的物体、区域或风格	“这些图像由车载摄像机连续拍摄。在车辆正朝哪个方向行驶？”（MMSI-Bench [151], $K=2$）	4,618
时间/动作	为帧或步排序，追踪状态随时间的变化，选择下一个动作	“请按事件的正确排序排列这些图像。”（MMR-Life [70], $K=4$）	3,463
文本/网络	结合文本和图像给出答案，以图像作为证据	“哈默史密斯的 Black Lion 和 Hop Poles 是否都有黑色背景，位于底层上方的主招牌上？”（MileBench [116] WebQA [13], $K=2$）	154
检索	从多张图像中找到一张图像，或阅读一份多页文档	“请在候选图像中找出与此最相关的图像描述。”（MMRB [19], $K=4$）	2,351

移除步骤之后，ASTERISM 中有 222 个源条目的标签需要修正；总计 569 个条目（包括转换而来的 83 个自由形式条目）在标签、答案格式、问题或选项方面存在超出选项顺序的差异（表中的修正项 S21）。

第 2.1 节和图 2 中的比例是在这 32,688 项上计算得出的，并统计了两轮审核和两个筛选阶段的判定结果。检索项、多页文档项、仅由一张图像或问题即可作答的项，以及复制其他来源项目的项，均不计入，即使这些项目已被移除。表 S2 的最后三列给出了每个来源中由一张图像或问题确定答案的项目比例。

我们平衡正确选项的位置。在每个选项计数中，每个字母作为答案出现的频率大致相同；为此，我们在 834 个项目中将正确选项与另一个选项进行了交换。选项指向其他选项、答案类似于“以上都不是”，或选项只是字母的项目，保留其排序。

B.3. 人工验证

共有四位审阅者参与。一位审阅者会检查每个修正后的标签和每个改写条目；对于包含 1,000 个条目的协议样本，每个条目则由两位审阅者独立检查。审阅者在网页界面中工作，界面会显示图像、问题和选项。审阅者首先在看不到标签的情况下回答条目，随后查看标签并记录判定，还可以选择添加备注。对于改写条目，审阅者在查看标签前有 94% 的正确回答率，并且认为其中 98% 需要多张图像；他们标记的条目会从 ASTERISM 中移除（补充材料 B.2）。在包含 1,000 个条目的样本中，审阅者的回答与审计判定有 82.6% 一致（Wilson [139] 95% 区间 [80.9, 84.2]），彼此之间在 82.3% 的条目上意见一致。

B.4. 任务类别

每个条目都属于七类任务之一，类别名称依据回答所需的跨图像操作而定（表S3）。GPT-5.6-Sol 为 19 个来源的条目标注了操作类型，并通过固定的对应表将操作类型映射到各类别。五条结构规则用于确定对应表划分的类型（搜索一张图像的图库、找到指针后再读取、跨三帧或更多帧进行跟踪、提示中包含表格，以及科学和编程问题）。在某些来源子集中，类别则由适用于该子集的规则、其来源类别或问题模式决定；例如，MileBench 的 WebQA 源条目属于文本/网络，除非第二轮审核更正其类别。改写条目保留其改写所依据条目的对应表类别。

任务是从多张图像中找出一张图像或阅读多页文档的条目属于检索类别。上述规则，而非审计判定，决定哪些条目属于检索项。检索条目要求模型从其他图像或页面中找出包含答案的图像或页面，因此第 2.1 节的统计不会将此类条目计作由一张图像或问题作答，即使审计判定将其移除也是如此（图 2中以斜线阴影标出）。因此，检索测试的是从图像中寻找证据，而其他六个类别测试的是跨图像整合信息。图S1中的 MIRB 条目询问两张地标照片的拍摄地点，按照规则，这不算检索。任一照片都能回答该问题，因此第2.1 节的占比统计将其计入。相比之下，图 S13中的 MMLongBench-Doc 条目询问一份文档中的页面，按照规则，这算作检索。除 83 个转换后的自由形式条目外，每个条目都采用相同的标注方式，记录跨图像操作、图像之间的关系、操作次数与推理跳数，以及时间尺度。表 S11 和图 S16使用了这些标签。我们随 ASTERISM 一并发布这些逐条标签。

C. 示例条目

C.1. 删除和更正的条目

本节的图展示了由一张图像作答的条目（图 S1）、由文本作答的条目（图 S2）、含糊不清或无法回答的条目（图 S3）、保留但修正了标签的条目（图 S4），以及保留但修正了答案格式的条目（图 S5）。图中展示了各类条目及其图像、问题、选项（如有）和源基准的标签。图 S1 至 S5 中的条目均来自不同的来源基准或子集。

C.2. 改写条目

改写保留源条目的图像，并替换问题和选项，使每个选项都结合来自不同图像的证据（第 3.1 节）。图 S6 展示了改写及其源条目。每个选项背后图像证据的斜体说明由我们添加，不属于这些条目。

MIRB（观光）· 2 张图像





以下观光景点在哪里？标签 **新加坡**

Qwen3.5-9B
仅文本 A ×仅图像1 新加坡 ✓仅图像2 新加坡 ✓
所有图像 新加坡 ✓

MileBench（多模态问答）· 2 张图像

1

	Home Team	Score	Away Team	Date
0	Albion Rovers	5–2	East Stirlingshire	12 August 1961
1	Forfar Athletic	1–3	Queen's Park	12 August 1961
2	Morton	2–1	Forfar Athletic	16 August 1961
3	Queen's Park	1–3	Albion Rovers	16 August 1961
4	East Stirlingshire	4–1	Morton	19 August 1961
5	Forfar Athletic	3–5	Albion Rovers	19 August 1961
6	East Stirlingshire	1–0	Forfar Athletic	26 August 1961
7	Morton	1–3	Queen's Park	26 August 1961
8	Albion Rovers	3–1	Morton	2 September 1961
9	Queen's Park	7–2	East Stirlingshire	2 September 1961

2



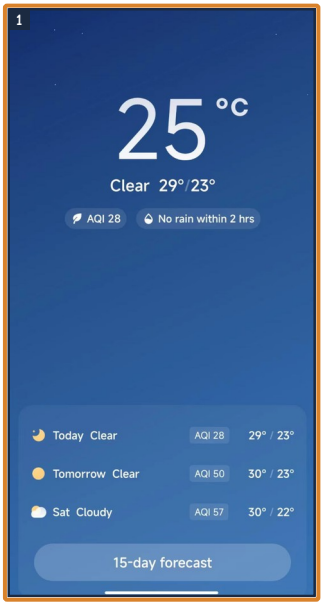
全局表格： [illegible]下文： 问[illegible]：在 1961-62 年苏格兰联赛杯中，哪支主队的队徽是黑白色的?
A. 希伯尼安足球俱乐部 B. 东斯特林郡足球俱乐部 C. 圣约翰斯通足球俱乐部 D. 中洛锡安之心足球俱乐部

标签 **B**

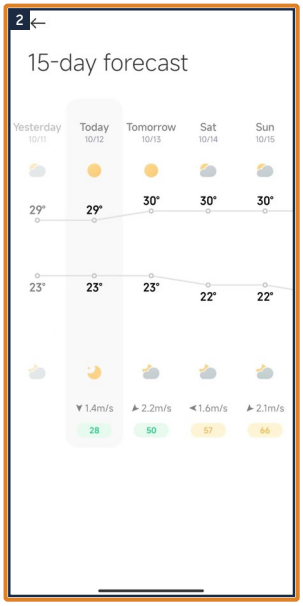
Qwen3.5-9B仅文本
D ×仅图像 1 B ✓
仅图像 2 B ✓所有图像 B ✓

MMRB（视觉定位）· 2 张图像

1



2



今天的天气信息如下： 以下是最新的[illegible]天气预报： 最近适合进行户外运[illegible]吗? (A) 以上选项均不符合 (B) 不确定 (C) 适合 (D) 不适合，因为空气质量较差

标签 **C**

Qwen3.5-9B
仅文本 A ×仅图像 1 C ✓仅图像2 C ✓
所有图像 C ✓

图 S1。**因一张图像即可回答而移除的项目**。每个面板展示一个源条目及其图像、问题（如有）选项和标签。橙色边框标记出每张单独即可回答该条目的图像，符号 ✓ 和 × 表示 Qwen3.5-9B 在仅文本、保留一张图像以及保留所有图像的情况下，在第 2.2节的灰色遮罩下回答正确还是错误。提示词中的答案格式说明和通用前言未在此处或图 S2 至 S6中显示。MIRB 条目的任一地标——图像 1 中的滨海湾金沙或图像 2 中的鱼尾狮——都能给出新加坡；仅使用文本时，Qwen3.5-9B 对这一开放式问题回答了字母 A。在 MileBench MultiModalQA 条目中，图像 2 里的队徽为黑白色，并标注为东斯特灵郡足球俱乐部；而在选项中的各俱乐部里，只有东斯特灵郡足球俱乐部作为主队出现在图像 1 的表格中。MMRB 条目的图像 1 显示天气晴朗，空气质量指数为 28；图像 2 显示晴天或局部多云天气的预报，因此任一屏幕都给出“适宜”。

MMRB（网页问答）· 2 个图像



图像图注 #1：瞭望塔，布拉尼城堡，宅邸与花园，布拉尼（ 506704）（28363314801） 图像图注 #2：厄克特城堡塔楼宅邸 问题：克特城堡塔楼宅邸和布拉尼城堡的瞭望塔主要由哪种材料制成？（A）克特城堡塔楼宅邸和布拉尼城堡的瞭望塔主要由石材制成。（C）厄克特城堡塔楼宅邸和布拉尼城堡的瞭望塔主要由木材制成。（D）厄克特城堡塔楼

标签

BQwen3.5-9B仅

文本 B ✓

MileBench（WebQA）· 2 个图像



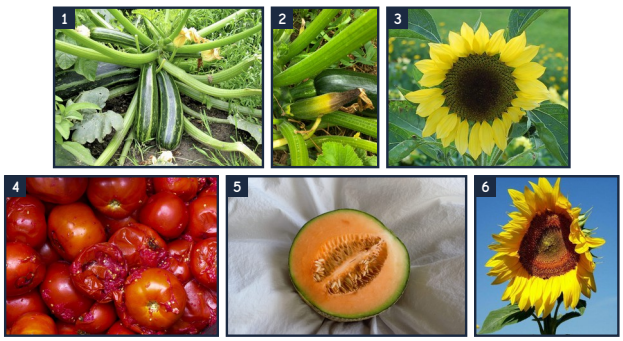
图像图注 #1：吉卜赛蘑菇 Kitsemampel 图像图注 #2：干枯纤孔正在渗出液体问题：；干枯纤孔菌还是吉卜赛蘑菇？A. 吉卜赛蘑菇看起来渗出得更多。B. 干枯纤孔菌看起来渗出得更多。C. 两者看起来都没有渗出。D. 两者看起来渗出的量相同。

标签

BQwen3.5-9B仅

文本 B ✓

MRAG-Bench · 6 个图像



在列出的特征中，哪一项不太可能出现在这种水果氧化后？A：其表皮形成黏滑质地。B：其表面被白色霉菌覆盖。C：其果肉出现褐色变色。D：其果肉变成绿色偏蓝的颜色。

标签

DQwen3.5-9B仅

文本 D ✓

MuirBench · 4 个图像



以下哪张图像是由这段 Python 代码生成的：a = rint("b is greater than a")?
A. [图像 1]B. [图像 2]C. [图像 3]
D. 所提供的选项均不符合 E. [图像 4]标签 DQwen3.5-9B仅文本

图 S2. **因文本已给出答案而移除的项目**。每个面板都展示一个源条目及其图像、问题、选项和标签，并高亮文本中泄露答案的部分以及正确选项；同时展示 Qwen3.5-9B 将所有图像替换为灰色后的回答（仅文本）。在 MMRB 项目中，两座城堡塔楼的名称在没有图像的情况下就指向了石材。在 MileBench 项目中，图像 2 的图注称干枯纤孔菌正在渗液。在 MRAG-Bench 项目中，第一张图像是查询，其余图像是检索示例；三个选项描述了常见的腐败情况，因此剩下选项 D。在 MuirBench 项目中，由于 200 不大于 333，代码不会打印任何内容，因此即使没有图像，选项 D 也很可能是正确答案。

MRAG-Bench · 6 张图像



你能告诉我这是什么花吗?
A：刺苞菜蓟
B：朝鲜蓟
C：菜蓟属
D：洋蓟

标签	D
同样合理	B

Artichoke 和 *globe artichoke* 指的是同一种植物，而 *Cynara* 是它所属的属。

MuirBench · 5 个图像

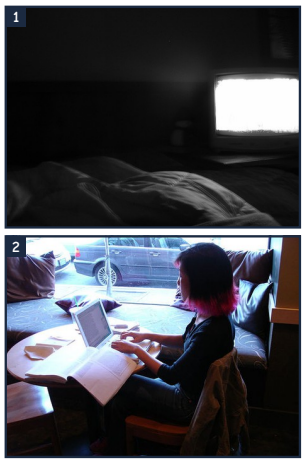


[图像 1] 以下哪张图像与其处于同一场景
给定的图像中是否包含灯这一物体?
A. [图像 2]
B. 所提供的选项中没有符合项
C. [图像 3]
D. [图像 4]
E. [图像 5]

标签	A
也说得通	C, D

图像 *2*、*3* 和 *4* 都显示了一间有灯的客厅。

Q-Bench2 · 2 张图像

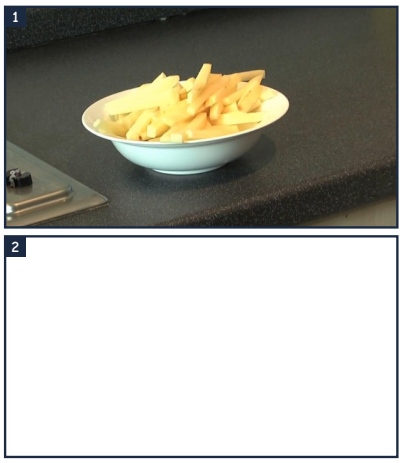


这两张图像是否都过度曝光?
A. 是
B. 否

标签	A
同样合理	B

图像 *1* 除了过曝的电视屏幕外都很暗，因此只有其高光过度曝光。

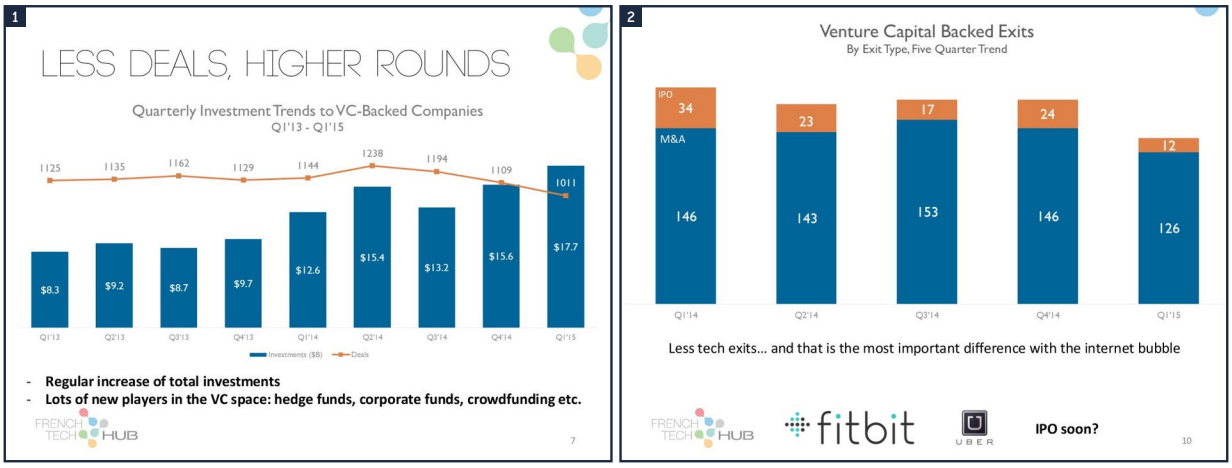
M4Bench（状态比较）· 2 张图像



第一张图像和第二张图像中的土豆是否处于相同状态? (A) 否 (B) 是标签B缺失 图像 2 的内容图像 *2* 是空白的，因此看不到第二张图像中的土豆

图 S3。 **含糊不清或无法回答的项目。** 审计会移除这样的源条目：对于这些条目，有两个或更多答案都说得通，或者无法从输入中得出答案（补充材料 B.1）。每个面板都展示一个源条目及其图像、问题、选项和标签、其他说得通的答案或缺失的信息，以及图像证据。在 MRAG-Bench 条目中，第一张图像是查询，其他图像是检索示例；artichoke 和 globe artichoke 指的是同一种植物。在 MuirBench 条目中，图像 2、3 和 4 都展示了一个带灯的客厅，与图像 1 中的房间相似。在 Q-Bench2 条目中，图像 1 除了过曝的电视屏幕外一片昏暗，因此按照对问题的一种理解，它算作过度曝光；按照另一种理解，则不算。M4Bench 条目中的图像 2 是空白的，因此无法比较其中本应展示的土豆。

MileBench（SlideVQA）· 2 张图像



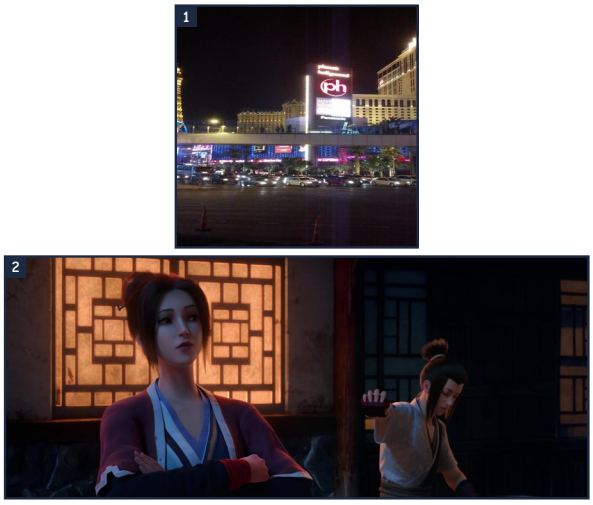
问题：关于风险投资支持公司的季度投资，交易数量较少的是 2013 年第二季度，还是首次公开募股型风险投资支持退出数量为 24 的那个季度？A. 2013 年第二季度

B. 首次公开募股型风险投资支持退出数量为 24 的那个季度

原始标签 A ✗
修正后的标签 B ✓

图像 *2* 中的 *24* 笔 *IPO* 退出发生在 *2014* 年第四季度，该季度在图像 *1* 中有 *1,109* 笔交易，少于 *2013* 年第二季度的 *1,135* 笔。

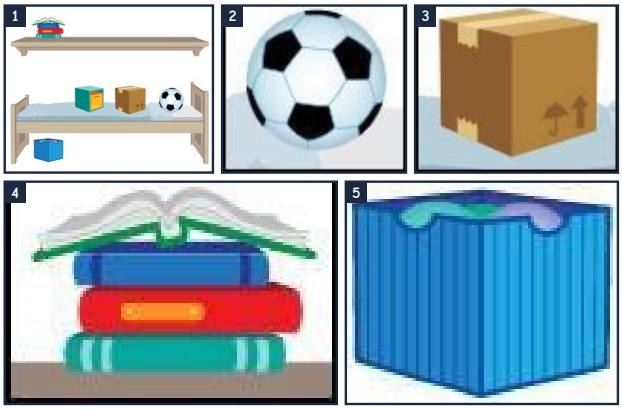
Q-Bench2 · 2 张图像



第一张图像比第二张图像更真实吗？
A. 是
B. 否
原始标签 B（否）✗
修正后的标签 A（是）✓

图像 *1* 是一张城市街道的夜间照片，图像 *2* 是一帧静止画面来自一段三维动画。

MuirBench · 5 张图像



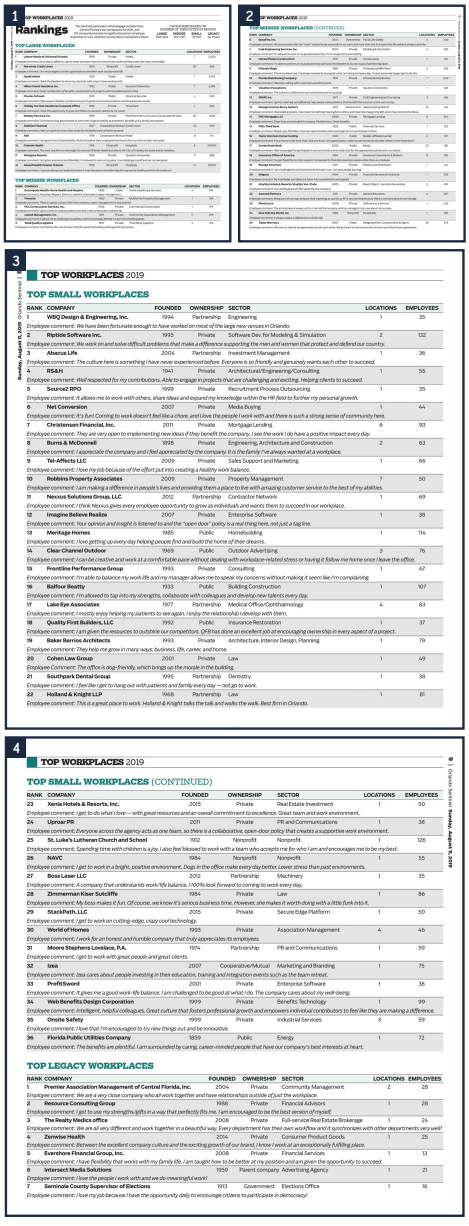
[图像 1] 哪个物体在球体形状的物体旁边？
A. [图像 2]
B. [图像 3]
C. [图像 4]
D. [图像 5]
E. 未提供任何选项

原始标签 A ✗
修正后的标签 B ✓

源标签 *Image 2* 指的是球本身，而纸板箱则是 *Image 3* 中的物体，即 *Image 1* 中球旁边的物体。

图 S4。**保留并采用修正后标签的项目**。每个面板展示一个 ASTERISM 项目及其图像、问题和选项、源基准的标签（×）、修正后的标签（✓）以及支持修正的图像证据。在 MileBench 项目中，源标签为 2013 年第二季度，但图像 2 中的 24 起 IPO 退出属于 2014 年第四季度，而图像 1 中该季度有 1,109 起交易，少于 2013 年第二季度的 1,135 起。在 Q-Bench2 项目中，图像 1 是一张夜间照片，而图像 2 是一帧 3D 动画，因此第一张图像更加逼真。在 MuirBench 项目中，源标签指的是球本身，而图像 1 中球旁边的物体就是图像 3 中的纸板箱。一名人工审核员接受了每项修正（补充材料 B.3）。

DUDE · 4 张图像



问题：COLE ENGINEERING SERVICES INC 成立于哪一年？

标签	2004-01-01
Qwen3.5-9B	2004
官方评分器	拒绝 ✗
ASTERISM 评分器 接受	✓

标签是一个完整日期，*ASTERISM* 也接受问题所要求的年份问题所询问的内容。

MIRB（arXiv）· 4 张图像



第 13 页上引文 [14] 的论文标题是什么？

标签	别走捷径：基于集成的规避已知数据集偏差的方法
Qwen3.5-9B	别走捷径：基于集成的规避已知数据集偏差的方法。
官方评分器	拒绝 ✗
ASTERISM 评分器 接受	✓

两种评分器都会去掉末尾的句点。随后，官方评分器会拒绝直引号，而 *ASTERISM* 会将其规范化。

图 S5。 **采用修正后的答案格式的条目。** 每个条目的标签都正确，但其来源所用的官方评分器会拒绝以问题要求的形式给出的正确答案；而 ASTERISM 会采用能够接受这种形式的答案格式为该条目评分，该格式可以在条目中指定，也可以通过评分器的规则设定（补充材料 B.1 和 B.2）。每个面板展示一个 ASTERISM 条目及其图像、问题和标签、Qwen3.5-9B 的答案，以及官方评分器和 ASTERISM 评分器是否接受该答案。DUDE 的问题询问成立年份，标签则是完整日期，而 ASTERISM 也接受年份。MIRB 条目的答案给出了所引标题，其中使用了直引号，而标签中使用的是弯引号；官方 MIRB 评分器采用的精确匹配会拒绝该答案，ASTERISM 则会将其规范化。

MileBench（WebQA）· 2 张图像





源条目

[redacted] 图像图注 #1：Lyric Theater，布莱克斯堡 Lyric 电影院，市中心，布莱克斯堡，弗吉尼亚州，美国 图像图注 [redacted]：Lyric Theater，韦克罗斯历史街区，佐治亚州，美国 Lyric Theater，韦克罗斯历史街区，佐治亚州，美国问题：布莱克斯堡 Lyric Theater 和佐治亚州 Lyric Theater 标牌上“lyric”一词的颜色是否不同？

A. 颜色可能不同，但也可能相似。B. 我不确定颜色是否不同。C. 不，颜色没有不同。D. 是的，颜色不同。

标签 C

被标记为含糊不清，因为一个标牌上的字母是黄色的，而另一个上的字母是奶油色的，并且选项 *A* 和 *C* 都表示颜色相同。

重写· 比较/变化

比较两个“LYRIC”剧院标牌。哪个描述正确地将每个标牌面板的背景颜色与其建筑物的立面相匹配？A. 浅色石材立面上的紫色面板；红砖立面上的深蓝色面板（图像 *1* 中为石材上的紫色，图像 *2* 中为砖上的蓝色）B. 红砖立面上的紫色面板；浅色石材立面上的深蓝色面板（两个立面互换）

C. 两个面板均为紫色，一个位于石材上，一个位于砖上 *(*图像 *2* 面板为深蓝色*)*D. 两个面板均为深蓝色，一个位于石材上，一个位于砖上 *(*图像 *1* 面板为紫色*)*

标签 A

仅图像 1，筛选器 B ×仅图像 2，筛选器 B ×所有图像，Qwen3.5-9B A ✓

MIRB（观光）· 2 张图像





源条目

以下观光景点在哪里？

标签 **巴黎**

已被单图筛选器移除，因为仅图像 *1* 中的埃菲尔铁塔或图像 *2* 中的凯旋门就足以显示巴黎。

重写 · 聚合

以下哪项在两张图像中都可见？A. 两个人举起双臂比出心形（仅图像 *1*）B. 树木 （图像 *1* 中塔的下方，图像 *2* 中沿着大道）C. 带有雕刻浮雕板的石拱门 （仅图像 *2*）D. 棕褐色（单色）色调 （仅图像 *2*）

标签 B

仅图像 1，筛选器 A ×仅图像 2，筛选器 C ×所有图像，Qwen3.5-9B B ✓

Mantis-Eval · 2 张图像





源条目

要将第一张图像变成第二张图像，需要进行什么操作？A. A：翻转方形白色碗 B. C：移除圆形白色碗 C. A：将粉色碗放在白色碗上面 D. D：以上都不是

标签 C

已移除，因为选项 *B* 和 *C* 以另一个字母开头，容易让人选错字母。

重写 · 时间/动作

图像 1 和图像 2 之间，恰好有一道菜肴改变了位置。是哪一道菜肴？它是如何移动的？A. 粉色碗：从桌上堆叠物旁移入白色圆碗（图像 *1* 中在起始位置，图像 *2* 中在结束位置）B. 白色圆碗：从方盘上移除（图像 *2* 中仍在盘子上）C. 粉色碗：从白色碗内移到桌面上 （两个状态正好相反）D. 白色方盘：从粉色托盘下方移到托盘上方（在两张图像中都位于托盘上）

标签 A

仅图像 1，筛选器 C ×仅图像 2，筛选器 C ×所有图像，Qwen3.5-9B A ✓

图 S6。**改写条目与其源条目并列展示**。每个面板都展示一个源条目、该条目被标记或移除的原因，以及其在 ASTERISM 中的改写和改写所属的任务类别。标记为筛选器的行给出了筛选委员会仅依据该图像得出的最常见答案。选项后的斜体注释说明了支持各选项的图像证据。MileBench 源条目含糊不清，因为图像 1 中的 LYRIC 是黄色的，而图像 2 中是奶油色的，并且选项 A 和 C 都表示颜色相近。改写要求将每个标志面板与其立面相匹配：图像 1 中为浅色石材上的紫色，图像 2 中为红砖上的深蓝色。仅依据 MIRB 源条目，图像 1 中的埃菲尔铁塔和图像 2 中的凯旋门分别可以得出答案。在改写的四个选项中，只有树木同时出现在两幅图像中；爱心手势仅出现在图像 1 中，雕刻拱门和棕褐色调仅出现在图像 2 中。Mantis-Eval 源条目的选项 B 和 C 都以另一个选项的字母开头，而改写询问哪道菜移动了以及移动方向，因此需要图像 1 中的起始状态和图像 2 中的结束状态。

C.3. 每个任务类别的条目

对于七类任务（补充材料 B.4），图 S7至S13 展示了该类别的 ASTERISM 条目。我们仅选取审计判定为有效、且其标签和文本在审计后均未发生变化的条目。

VLM2-Bench · 2 张图像



输入图像中有多少种不同的玩具?

标签 2

MuirBench · 2 张图像



有多少台售卖食品或饮料的自动售货机?
A. 八台
B. 四
C. 十
D. 所提供的选项均不正确
E. 七

标签 E

MuirBench · 2 张图像



这两张图像中有多少只鸟?
A. 以上选项均不正确
B. 四只
C. 两只
D. 三只
E. 五只

标签 B

VLM2-Bench · 3 张图像



输入图像中有多少个不同的杯子?

标签 2

图 S7. **聚合条目**。 该类别的四个 ASTERISM 条目，以及每个条目的来源、图像数量、问题、选项（如有）和标签。第一个 VLM2-Bench 条目中，图像 1 和图像 2 展示了两个不同的擎天柱模型。第一个 MuirBench 项目中，图像 1 展示了五台自动售货机，图像 2 展示了两台。第二个 MuirBench 项目中，每张图像都展示了两只绯红金刚鹦鹉，合计四只。第二个 VLM2-Bench 条目中，图像 2 和图像 3 展示了同一个漩涡图案纸杯，而图像 1 展示了一个棕色纸杯，因此这些图像中共有两种不同的杯子。此处以及图 S8 至 S13 中均未展示提示词的回答格式说明和通用前言。S8 至 S13。

MMRB（时尚问答）· 4 张图像



这四张图片中的衣服颜色相同吗？
（A）是（B）否标签 B

Q-Bench2 · 2 张图像



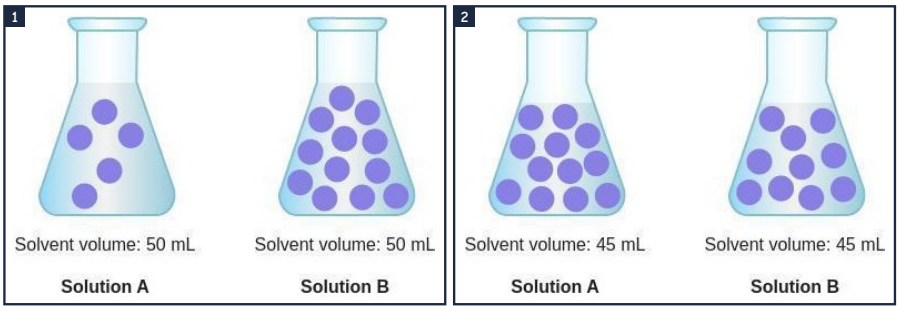
与第一张图像相比，
第二张图像的色彩鲜艳度
如何？
A. 更生动
B. 不那么生动
C. 差不多
标签 B

Mantis-Eval · 2 张图像



你想要哪种促销活动
考虑折扣后该去哪里？
A. 第一
B. 第二
C. 无法确定
标签 A

MuirBench · 2 张图像



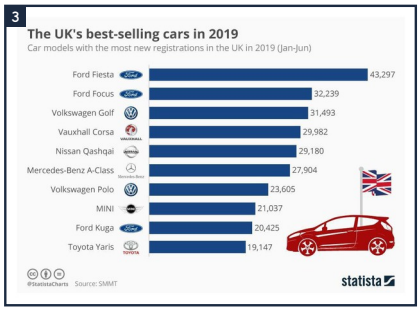
在哪张图像中，方案 A 的
紫色颗粒浓度更高
？
A. 以上选项均不正确
B. [图像 1]
C. [图像 2]
标签 C

图 S8。**比较/变化条目**。该类别的四个 ASTERISM 条目，如图 S7 所示。在 MMRB 条目中，图像 1 至图像 3 展示的是黑色服装，而图像 4 展示的是深红色牛仔裤，因此这些衣服并非同一种颜色。在 Q-Bench2 条目中，图像 1 中红黄相间的球衣比图像 2 中的棕色猫更加鲜艳。在 Mantis-Eval 条目中，第一项促销最高可打五折，第二项最高可打七折，且均可额外享受七五折优惠。在 MuirBench 条目中，与图像 1 相比，图像 2 中的溶液 A 含有更多颗粒，且溶剂更少。

MuirBench · 3 张图像







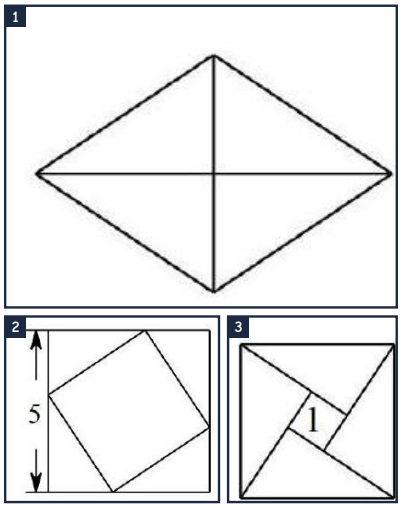
这些是汽车的图片。
图像中的车型?
A. 以上选项均不正确
B. 48941
C. 75536
D. 53276

标签 B

2这是英国畅销汽车的排名图表。

请问这些汽车的总销量是多少

MV-MATH · 3 张图像



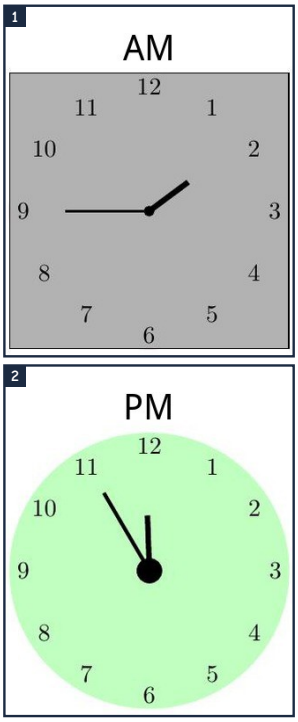
将图1中的菱形分成四份
全等直角三角形，沿着它的
对角线切开，然后将这四个
直角三角形拼成正方形，如图所示
见图2和图3。图1中的
菱形的面积为（）

1 图1 图2 图3

A. 12
B. 14
C. 16
D. 18

标签 A

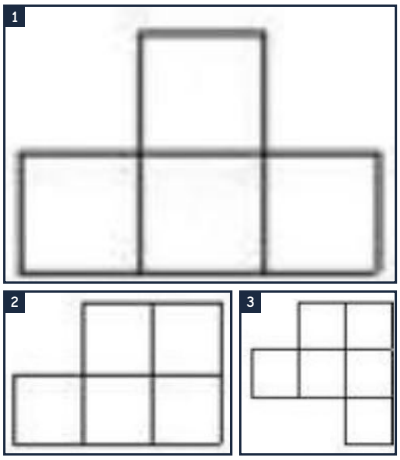
ReMI（时钟）· 2 张图像



图像中的两个时钟显示的是同一天里的两个
不同时间。求这两个时间之间的绝对差，
以分钟为单位
即两个时钟所显示的时间之差。

标签 1330.0

MV-MATH · 3 张图像



如图所示，这是
一个由若干相同的小立方体组成的三维结构的
三个视图。
形成这个结构用了多少个小立方体
三维结构？（）

1 左视图 视图 俯视图

A. 6
B. 7
C. 9
D. 8

标签 D

图 S9。 **推理/验证条目**。 该类别的四个 ASTERISM 条目，如图 S7 所示。在 MuirBench 项目中，图像 1 展示了一辆梅赛德斯-奔驰 A 级，图像 2 展示了一辆 MINI，而图像 3 的图表列出的它们的销量分别为 27,904 和 21,037，相加为 48,941。在第一个 MV-MATH 项目中，图像 2 中边长为 5 的正方形给出了每个三角形两条直角边之和，图像 3 中的单位正方形给出了它们的差，因此两条直角边分别为 3 和 2，菱形的面积为 12。在 ReMI 项目中，时钟显示凌晨 1:45 和晚上 11:55。在第二个 MV-MATH 项目中，三张图像中的左视图、正视图和俯视图共同确定了立方体的数量。

BLINK（多视图推理）· 2 张图像



这些图像是视频中的帧。该视频拍摄的是一个静态场景。摄像机正绕着物体顺时针（左）或逆时针（右）移动。第一张图像来自视频开头，第二张图像来自视频结尾。拍摄视频时，摄像机是向左还是向右移动？
A. 左 B. 右标签 A

MMSI-Bench · 2 张图像



当我看到图1中的场景就在我面前时，楼梯间区域位于我的
什么方位？

位置？
A. 我身后
B. 左前方
C. 右侧
D. 左后方

标签 C

VLM2-Bench · 4 张图像



哪些图像展示了与输入图像中相同的一双鞋？
A) 3 和 4；
B) 1、2 和 3；
C) 1、3 和 4；
D) 1 和 4

标签 D

图 S10. **空间对应项目**。 该类别的三个 ASTERISM 条目，如图所示 S7。在 BLINK 条目中，摄像机的方向仅体现在两个帧之间的视图变化上。在 MMSI-Bench 条目中，图像 1 右侧带圆点靠垫的皮革沙发再次出现在图像 2 中；沙发旁敞开的门显示出带楼梯的门厅，因此楼梯间位于第一个视图的右侧。在 VLM2-Bench 条目中，只有图像 1 和 4 显示了带斑点的白色鞋底的针织鞋。

MuirBench · 3 张图像



婴儿试图穿上衣服。根据给定的上下文，图像的正确排序是什么？ A. 第三张图像、第一张图像、第二张图像 B. 以上选项均不正确 C. 第二张图像、第一张图像、第三张图像 D. 第二张图像、第三张图像、第一张图像

标签 A

MMR-Life · 6 张图像



问题：我想做蒜香面包，请生成正确的图像排序，以组成完整的食谱
A. 6-4-3-5-2-1
B. 6-4-2-5-3-1
C. 6-3-2-5-4-1
D. 6-3-2-4-5-1
E. 6-4-3-2-5-1

标签 E

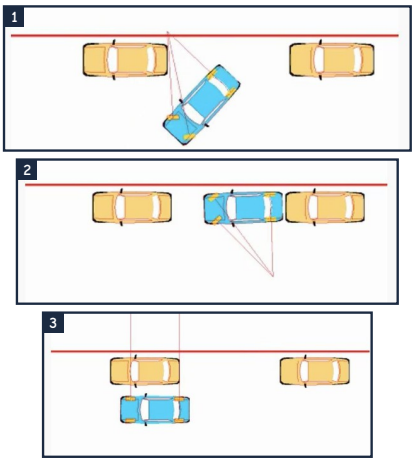
MMR-Life · 6 张图像



问题：我想做芦苇披萨，请生成正确的排序这些图像，以组成一份完整的食谱
A. 6-2-4-5-3-1
B. 6-4-1-5-3-2
C. 2-4-1-5-3-6
D. 6-2-4-1-5-3
E. 4-2-3-5-1-6

标签 D

MuirBench · 3 张图像



蓝色汽车试图驶离停车位。根据给定的上下文，正确的图像排序是什么？
A. 第二张图像、第一张图像、第三张图像
B. 第三张图像、第一张图像、第二张图像
C. 以上选项均不正确
D. 第三张图像、第二张图像、第一张图像

标签 A

图 S11。**时间/动作项目**。该类别的四个 ASTERISM 条目，如图所示。S7。在第一个 MuirBench 项目中，外套在图像 3 中被举过头部，在图像 1 中穿了一半，在图像 2 中则已穿好。在第一个 MMR-Life 项目中，图像 6 中的食材最先出现，接着是图像 4 中的裸切片、图像 3 中的黄油、图像 2 中的大蒜、图像 5 中的香草，以及图像 1 中的烤面包片。在第二个 MMR-Life 项目中，披萨制作过程从图像 6 中的食材开始，经过图像 2、4 和 1 中的面粉、揉面和压制饼皮，最后到图像 5 中的酱汁和图像 3 中的配料。在第二个 MuirBench 项目中，蓝色汽车停在图像 2 中，在图像 1 中驶出车位，并在图像 3 中离开了车位。

MileBench (WebQA) · 2 张图像




图像图注 #1：英国伦敦斯皮塔菲尔德基督堂管风琴 - Diliff 英格兰伦敦斯皮塔菲尔德基督堂的管风琴。

图像图注 #2：韦尔什普尔基督堂。1815年由 Gray 制作的管风琴及陶土拱门 1815年由 Gray 制作的管风琴及陶土拱门。该管风琴从韦尔什普尔教区教堂迁移而来。问题：韦尔什普尔基督堂和英国伦敦的斯皮塔菲尔德基督堂的管风琴音管是否都镀金？ A. 只有英国伦敦斯皮塔菲尔德基督堂管风琴的音管镀金。B. 不，韦尔什普尔基督堂和英国伦敦斯皮塔菲尔德基督堂的管风琴音管都没有镀金。C. 是的，韦尔什普尔基督堂和英国伦敦斯皮塔菲尔德基督堂的管风琴音管都镀金。D. 不，只有韦尔什普尔基督堂管风琴的音管镀金。

标签 C

MileBench（WebQA）· 2 个图像



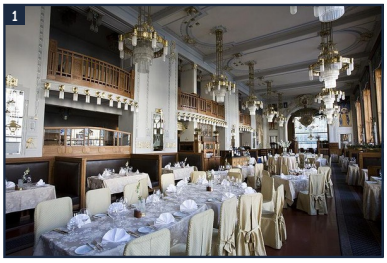

图像图注 #1：阿拉巴马州休斯敦县多森的多森市政中心

图像图注 #2：维斯塔市政中心 维斯塔市政中心入口，从市政中心大道方向观看。问题：多森市政中心还是维斯塔市政中心前方的旗帜更多？ A. 两座建筑前方的旗帜数量相同。

B. 维斯塔市政中心前方的旗帜比多森市政中心更多。
C. 多森市政中心前方的旗帜比维斯塔市政中心更多。
D. 两座建筑前方都没有旗帜。

标签 B

MileBench（WebQA）· 2 张图像



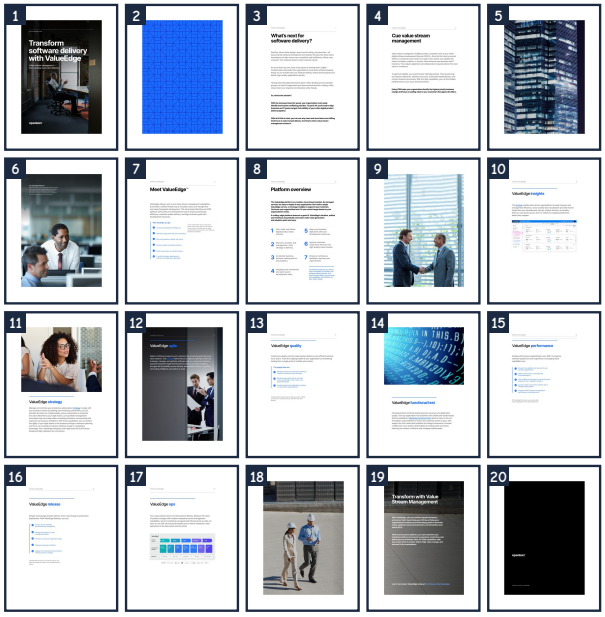

图像图注 #1：市民会馆餐厅，布拉格 - 8377 市民会馆餐厅，布拉格

图像图注 #2：OldTown White Coffee 门店 Taman Permata 的 OldTown White Coffee 门店，吉隆坡 问题：市民会馆餐厅（布拉格）和 Taman 的 OldTown White Coffee 门店的桌子上都有桌布吗？ A. 没有，只有市民会馆餐厅的桌子上有桌布。B. 没有，市民会馆餐厅和 OldTown White Coffee 门店的桌子上都没有桌布。C. 只有 Taman 的 OldTown White Coffee 门店有桌布。D. 有，两家餐馆的桌子上都有桌布。

标签 A

图 S12. **文本/网页项目**。 该类别的三个 ASTERISM 条目，如图 S7 所示。第一个 MileBench 条目中，图注写出了两座器官的名称，只有图像显示它们都有镀金管道：图像 1 中的是伦敦的器官，图像 2 中的是威尔士浦的器官。第二个 MileBench 条目中，图注写出了两座市民中心的名称，只有图像显示图像 1 中有一面旗帜，图像 2 中有三面旗帜。第三个 MileBench 条目中，图注写出了两家餐馆的名称，只有图像显示图像 1 中有桌布，而图像 2 中是裸桌。

MMLongBench-Doc · 20 张图像



问题：包含截图的页面标题是什么？标签 ValueEdgeinsights

MuirBench · 3 张图像



下面哪张图片更符合以下描述：一间带有灰色台面的白色厨房。

A. 所提供的选项中没有符合项
B. [图像 1]
C. [图像 2]
D. [图像 3]

标签 D

MMRB（文本到图像检索）· 4 张图像



请从候选图像中找出与此描述最相关的一张。

描述：这只鸟的身体是白色的，臀部是灰色的，且有一只黄色的 喙。
这只鸟是白色的，鸟喙很大，翅膀是黑色的。
这只鸟是白色的，翅膀是黑色的，且有一只橙色的长喙。
这只鸟的头部是白色的，喙又长又弯，且有一只白色腹部和黑色翅膀。
一种中等大小的鸟，整体为白色，带有一个非常大的钩状喙
一种体型非常大的鸟，身体大部分为白色，并有一个大鸟喙。
这只鸟的翅膀是黑色的，鸟喙是黄色的
这只鸟有黑色翅膀、白色胸部、白色颈部，以及白色头部，以及黄色的鸟喙。
这只鸟的身体上有白色羽毛，翅膀上有黑色翅膀羽毛，它有一只粗壮的橙色喙。
这只鸟的胸部颜色雪白，喙又长又弯曲

A：第一张图像
B：第二张图像
C：图像 3
D：图像 4

标签 B

图 S13。 **检索项**。 该类别的三个 ASTERISM 条目，如图 S7所示。在 MMLongBench-Doc 项目中，20 个页面中只有图像 10 显示了产品截图，其标题为 ValueEdge insights。在 MuirBench 项目中，三个厨房全部为白色，只有图像 3 中的厨房有灰色台面。在 MMRB 项目中，图像 2 显示了四只鸟中唯一一只具有白色身体、深色翅膀和长钩状喙的鸟。

D. ASTERISM-Lite

表 S4 给出了表中多模态大语言模型的准确率。3 按任务类别划分的结果，以及表S5 列出了 ASTERISM-Lite 的构成。

D.1. 选择

ASTERISM-Lite 在固定配额下选取 ASTERISM 中最难的条目。我们通过在 ASTERISM 上进行评估的开源基础模型在某条目上的平均准确率来衡量该条目的难度，也就是说，这些模型不是从其他受评估模型中经过后训练得到的模型。条目按难度从高到低纳入，难度相同的条目则按照固定的随机顺序打破平局。后训练多模态大语言模型（包括本文训练的模型）以及专有多模态大语言模型均不参与选择。每个任务类别包含 61 个条目，但 ASTERISM 中规模最小的文本/网络类别除外，该类别包含 34 个条目。在每个类别中，条目按照五个图像数量区间（2、3–4、5–8、9–16 以及 17 张或更多图像；表S5）的目标占比进行填充，仅当符合条件的条目过少时才放宽这些占比。每个源基准在某一类别中提供的条目数不超过 19 个，在文本/网络类别中不超过 11 个。

D.2. 验证

我们针对图像和标签，手动验证 ASTERISM-Lite 的全部 400 个项目，检查标签是否成立、是否存在无法排除的第二个答案、是否有单张图像即可回答该项目，以及是否有图像存在缺陷。对于验证失败的项目，我们按照相同配额，用其类别中的下一个项目替换，并以相同方式验证替换项目。

D.3. 排名一致性与配对区间

在 ASTERISM-Lite 上，对两个基准测试均进行了评估的 58 个开源多模态大语言模型（包括未参与选择的后训练模型）的排名与其在 ASTERISM 上的排名一致；它们在两个基准测试上的准确率之间的斯皮尔曼等级相关系数为 0.93（表 S4、S9和S10）。在针对 400 个项目进行的配对比较中（按类别分层的项目自助法 [26]，10,000 次重采样），Claude Opus5。5 [6] 领先于 Qwen3.8-Flash-Next， [107]，后者是最强的开源多模态大语言模型，领先幅度为 7.2 个百分点（95% 区间 $[2.0, 12.4]$）。

D.4. 专有多模态大语言模型

这六个专有多模态大语言模型通过应用程序编程接口接收开源实验使用的提示词和图像文件，并采用表 S6 中的标识符和设置；补充材料 E.1 中的基于规则的评分器会从返回文本中提取每个答案。它们均使用各自应用程序编程接口的高推理力度设置。

E. 评分与评估

E.1. 基于规则的评分器

每个项目都带有一种根据其提示生成的答案格式。该格式可以是选项字母（附选项文本）、是/否或真/假、符合问题允许的单位的整数或数字，或带有可接受别名列表的短文本；规则则根据来源任务、问题和选项生成该格式。除非任务的官方规则允许一定容差，否则数字必须与其标签完全匹配；例如，ReMI [52]对数字（1%）、几何（3%）和时钟读数（十分钟）允许容差，MMLongBench-Doc [87] 对小数答案允许1%的容差。列表答案只有在按任意顺序准确列出其标签所包含的项目时才计分。评分器会提取最后一个明确答案（例如“答案：”行、框选答案，或提示要求给出裸答案时的裸答案），拒绝列出多个选项的含糊答案，并且在提取过程中绝不查阅真实标注。它不使用大语言模型评审。在266万条合成对抗性输出（正确、错误、空白、含糊及顺序打乱的答案）上，它没有出现提取错误。在GLM-4.1V-9B-Thinking [43],的600条输出中，该模型会在每个答案前进行长篇思考；一名不知道评分器判定结果的标注员有599条与评分器判定一致。

E.2. 官方评分代码

对于 21 个模型的同一批生成结果，与补充材料 E.1 中的基于规则的评分器相比，官方代码会使各基准测试的准确率朝两个方向变化，并颠倒同一基准测试中模型对的排序（表 S7）。字符串相似度会对 DUDE 中的自由形式答案进行评分 [60]，也会对 [102] MileBench [116], 子集 CLEVR-Change 中的答案进行评分，它会比较答案措辞与参考答案，因此，格式不同但正确的答案可能比一个差之毫厘的答案得分更低。在多项选择基准测试上，官方代码和基于规则的评分器提取部分答案的方式也有所不同。

E.3. 评估设置

所有模型使用相同的提示词。每条提示词都遵循其来源的官方指令（如有），并要求按照条目的格式作答（补充材料 E.1）。在 ASTERISM 上评估的所有开源模型，每次回答的输出词元预算至少为 8,192。Qwen3.5-9B [105]及其训练出的模型均在启用思考的情况下进行评估，并采用模型卡片中的思考模式采样设置（温度 0.6、top-p 0.95、Top-k 20）。它们在第 2.2 节和图7的灰色遮罩测试中使用相同的解码方式。在 ASTERISM 上评估的所有其他开源模型，包括 Qwen2.5-VL-7B、GLM-4.1V-9B-Thinking 及其训练出的模型，均采用贪婪解码。GLM-4.1V-9B-Thinking [43], Qwen3-VL-4B-Thinking [104] 和 MiMo-VL-7B [147] 在每次回答前都会进行思考。InternVL3 [170]模型采用 VLMEvalKit 的多图像平铺规则 [25],LLaVA-OneVision [65] 采用其官方多图像设置，其他模型则使用各自的默认图像处理方式。

F. CICA 的完整规范

本节介绍算法1的规则，表S8 列出了相关设置。

表 S4. **ASTERISM-Lite 排行榜。** 顶部区块列出了表 3 中多模态大语言模型在各任务类别上的准确率、宏平均值（平均值）及宏平均值的 95% 区间，并列出混合专家模型的总参数量和激活参数量。底部区块列出了表 S9 和 S10 中的 58 个开源多模态大语言模型以及使用 CICA 训练的三个模型在 ASTERISM-Lite 上的平均准确率（平均值），并按准确率排序。各列含义同表 2，顶部区块每列的最佳值以粗体显示。Qwen3.5-9B 和 CICA-Qwen3.5-9B 在启用思考功能的情况下作答（补充材料 E.3）。

模型	Size	Agg	Cmp	Rsn	Spa	Tmp	Txt	Ret	Avg	95% 区间
专有										
Claude Opus 5.5 [6]	–	80.3	77.0	75.4	31.1	**44.3**	**100.0**	67.2	**67.9**	[64.2, 71.9]
Gemini 3.8 Flash [36]	–	**82.0**	80.3	67.2	21.3	36.1	91.2	70.5	64.1	[60.1, 68.1]
GPT-6-Astra [101]	–	75.4	83.6	**78.7**	31.1	39.3	85.3	45.9	62.8	[58.6, 67.3]
GPT-6.1-Sol [100]	–	73.8	**85.2**	77.0	26.2	37.7	85.3	45.9	61.6	[57.4, 65.8]
Qwen3.8-Max [108]	–	73.8	77.0	72.1	32.8	31.1	85.3	54.1	60.9	[56.6, 65.2]
Muse Spark 1.3 [91]	–	73.8	75.4	52.5	23.0	26.2	91.2	**75.4**	59.6	[55.4, 63.8]
开源，仅在 ASTERISM-Lite 上进行评测										
Qwen3.8-Flash-Next [107]	125B-A6B	68.9	72.1	63.9	**36.1**	31.1	94.1	59.0	60.7	[56.1, 65.2]
Kimi K3.1 Pro [93]	2.8T-A104B	65.6	70.5	54.1	19.7	27.9	85.3	49.2	53.2	[48.9, 57.8]
GLM-5.3-Flash [161]	320B-A18B	54.1	70.5	31.1	23.0	42.6	73.5	47.5	48.9	[44.0, 53.6]
Muse-Glimmer-30B [94]	30B	29.5	62.3	18.0	24.6	27.9	94.1	57.4	44.8	[40.7, 49.0]
MiMo-V2.6-Pro [148]	1T-A42B	39.3	54.1	26.2	13.1	21.3	85.3	36.1	39.3	[35.0, 43.6]
Nemotron-3.5-Super-VL [96]	120B-A12B	29.5	34.4	21.3	11.5	11.5	94.1	32.8	33.6	[29.7, 37.3]

模型	Avg
CICA-Qwen3.5-9B（我们的方法）	47.0
Qwen3.5-9B [105]	43.3
CICA-GLM-4.1V-9B-Thinking（我们的）	40.2
Qwen3.6-27B [106]	40.0
GLM-4.1V-9B-Thinking [43]	37.7
MiMo-VL-7B [147]	36.3
Qwen3-VL-4B-Thinking [104]	35.9
Qwen3-VL-32B [104]	35.8
Qwen3-VL-30B-A3B [104]	32.0
CICA-Qwen2.5-VL-7B（我们的）	29.9
Gemma-4-12B [35]	29.6
Qwen2.5-VL-32B [7]	28.3
Time-R1-7B [136]	27.9
PAPO-D-7B [137]	27.7
Migician-7B [72]	27.5
NoisyRollout-7B [78]	27.3
Qwen2.5-VL-7B [7]	27.2
VisionReasoner-7B [81]	27.2
PeRL-7B [165]	27.0
Qwen3-VL-8B [104]	26.9
Vision-Zero-7B [131]	26.3
Perception-R1-7B [146]	26.2

模型	Avg
ThinkLite-VL-7B [135]	26.2
SophiaVL-R1-7B [27]	26.1
VLAA-Thinker-7B [14]	25.7
Qwen3-VL-4B [104]	25.5
PAPO-G-H-7B [137]	25.3
VPPO-7B [45]	25.3
InternVL3-8B [170]	24.7
VideoRFT-7B [130]	24.6
Spatial-SSRL-7B [82]	24.4
R1-ShareVL-7B [153]	24.2
Video-R1-7B [29]	24.0
Visual-Jigsaw-7B [143]	23.6
ArrowRL-7B [150]	23.4
Gemma-3-12B [34]	22.9
Vision-SR1-7B [73]	22.7
TON-7B-Math [127]	22.6
Qwen3-VL-2B [104]	21.0
Open-o3-Video-7B [89]	20.1
LLaVA-OneVision-2-8B [4]	19.5
Qwen2-VL-7B [128]	18.7
UniVG-R1-70亿参数 [8]	18.1
SpatialReasoner-7B [86]	17.9

模型	Avg
Pixtral-12B [3]	17.5
Mulberry-7B [152]	16.6
TVC-7B [118]	16.4
InternVL3-2B [170]	15.1
mPLUG-Owl3-7B [154]	14.7
Gemma-3-4B [34]	14.2
Embodied-R1-3B [159]	13.3
SpatialLadder-3B [69]	12.7
LLaVA-OneVision-7B [65]	12.5
LLaVA-Interleave-7B [67]	12.3
Idefics2-8B [61]	11.5
CapRL-3B [149]	11.2
GRIT-3B [28]	10.5
Qwen2.5-VL-3B [7]	10.5
HATCH-3B [97]	10.1
Mantis-80亿参数-Idefics2 [49]	10.0
MindCube-3B [132]	9.6

表 S5. **ASTERISM-Lite 的组成。** 各任务类别及每项图像数，以及每个类别的来源基准数量。

		每项图像数					
类别	项目	2	3–4	5–8	9–16	17+	来源
聚合	61	28	21	11	1	0	8
比较/变化	61	43	11	4	3	0	10
推理/验证	61	39	11	10	0	1	9
空间对应	61	29	15	14	3	0	8
时间/动作	61	1	10	33	7	10	4
文本/网络	34	23	3	5	3	0	6
检索	61	8	11	19	4	19	5
总计	400	171	82	96	21	30	19

F.1. CICA 目标

每一步中，每个采样得到的响应 i 相对于其所在组的优势为 $A_i = (r_i - \bar{r})/(\mathrm{std}(r) + \epsilon)$，其中 $\epsilon = 10^{-6}$，且以 G 个响应的总体标准差为准（式（2）省略了 ϵ）。奖励 $r_i \in [0, 1]$ 是提取答案的准确率；当 $r_i = 1$ 时，响应计为正确。设 N 为响应数量，L_i 为 y_i 的长度（包括回合结束标记），并令 $\ell_{i,t} = \log \pi_\theta(y_{i,t} \mid x_i, y_{i,<t})$。每个骨干网络在每一步基于带有 CICA 权重的 GRPO [114]目标进行一次更新，

$$\mathcal{L} = -\frac{1}{N}\sum_i \frac{A_i}{L_i}\sum_t \eta_{i,t} w_{i,t} \ell_{i,t} + \frac{\beta_{\mathrm{KL}}}{T}\sum_i \sum_t \left(e^{\delta_{i,t}} - \delta_{i,t} - 1\right), \tag{S1}$$

表 S6. **ASTERISM-Lite 上的专有多模态大语言模型。** 表3中专有多模态大语言模型的应用程序编程接口标识符和接口。每个请求均携带开源运行所用的提示和图像文件，最多允许生成 65,536 个输出词元，且不设置温度，因此每个应用程序编程接口均按其默认设置进行采样。

模型	应用程序编程接口标识符	接口
Claude Opus 5.5 [6]	claude-opus-5-5	Anthropic API
Gemini 3.8 Flash [36]	google/gemini-3.8-flash	OpenRouter（Google AI Studio）
GPT-6-Astra [101]	gpt-6-astra	OpenAI API
GPT-6.1-Sol [100]	gpt-6.1-sol	OpenAI API
Qwen3.8-Max [108]	qwen/qwen3.8-max-0902	OpenRouter（阿里巴巴）
Muse Spark 1.3 [91]	meta/muse-spark-1.3	OpenRouter（Meta）

表 S7。**官方评分代码会使各基准测试的准确率向上或向下偏移。**在同一批生成结果上，官方评分代码与 ASTERISM 评分器的准确率之差。评估对象为 21 个模型在各基准测试上的表现；每个基准测试均使用固定的 n 个项目样本，且 21 个模型使用相同样本。表中给出了平均值、模型间的范围，以及排序发生逆转的模型对（共 210 对，记为 Rev.）。† 表示其 95% 区间不包含零的平均值。

基准测试	主要区别	Mean	最小/最大 Rev.	
EMMA [42] $n = 63$	信件来自推理文本	$+3.3^{\dagger}$	$-1.6/+17.5$	17
MMLongBench-Doc [87] $n = 198$	模糊匹配通过次数错误列表	$+2.7^{\dagger}$	$-0.5/+10.1$	5
MV-MATH [129] $n = 177$	信件来自推理文本	$+1.5^{\dagger}$	$0.0/+5.1$	5
DUDE [60] $n = 73$	差一点答对的情况通过，重新格式化答案失败	$+0.3$	$-5.5/+8.2$	5
ReMI [52] $n = 473$	拒绝替代方案黄金、通行证对冲	-0.2	$-3.6/+9.3$	25
MIRB [168] $n = 107$	拒绝数字词语释义是/否	$-2.3^{\dagger}$	$-10.3/0.0$	8
MMRB [19] $n = 469$	匹配为空图像选项文本	$-2.7^{\dagger}$	$-4.9/-1.3$	3
VLM2-Bench [162] $n = 464$	拒绝 A)，数字词语	$-3.4^{\dagger}$	$-8.6/0.0$	20

其中，$\delta_{i,\,t}$ 是参考模型 π_{ref} 相对于 π_θ的对数比，并裁剪至 $[-5, 5]$，$\beta_{\text{KL}} = 0.02$；T 是回合结束标记之前的响应词元数量。对 t 的求和，在第一项中遍历 y_i 的所有 L_i 词元，在第二项中则遍历回合结束标记之前的词元；$w_{i,\,t}$ 不携带梯度。这里，$\eta_{i,\,t} = \min(\pi_\theta/\mu,\ 2)$是每个词元经过截断的重要性权重，不携带梯度；其中 μ 是展开采样器生成该词元的概率，因此，对于与策略匹配的采样器，$\eta_{i,\,t} = 1$。CICA只改变策略项。KL项使用对相同响应进行的参考前向传播，且不按 w加权。

准确率奖励相同的组会被丢弃，并抽取新条目；保留组数量过少的步会被跳过，梯度会裁剪至范数 1（限制见表 S8）。每一步都会从打乱后的池中不重复地抽取条目，跳过提示本身超过 32,768 个词元的条目；如果提示与其任一响应的长度之和会超过该上限，则丢弃该组。

Alg算法 1 一次训练　　g step 使用 CICA。

要求：策略 π_θ，条目 $\mathcal{D}$，组大小 G，每步条目数 B，尝试次数上限 M，最小值 $B_{\min}$（表 S8），β，　　τ

1 $\mathcal{S} \leftarrow \emptyset$　　▷：保留组 2：当 $|\mathcal{S}| < B$ 且尝试次数少于 M 时：3：从 $\mathcal{D}$ 中抽取一个条目 x 4：从采样轨迹展开采样器 $\mu(\cdot \mid x)$ 中采样 $y_{1..G}$，并获得奖励 $r_{1..G}$ 5：如果所有 r_i 都相同，则继续 6：将组 $\{(x, y_i, A_i)\}_{i=1}^{G}$ 添加到 $\mathcal{S}$，其值为 $A_i = (r_i - \bar{r}) / (\text{标准差}(r) + \epsilon)$ 7：结束循环 8：如果 $|\mathcal{S}| < B_{\min}$，则跳过该步 9：对于输入 x_i 中包含 K_i 个图像的每个响应 y_i（属于 $\mathcal{S}$）：10：$\ell_{i,t} \leftarrow$ 记录 $\pi_\theta(y_{i,t} \mid x_i, y_{i,<t})$，并计算梯度 ▷ 11：$\ell_i^{(k)}{}_{,t} \leftarrow$ 记录 $\pi_\theta(y_{i,t} \mid x_i^{(k)}, y_{i,<t})$，$k$=1..$K_i$　　▷不计算梯度。12：$k^*_{i,t}$ 以及$\sigma_{i,t}$，根据公式（3）和（4）计算 13：结束循环 14：公式（5），在每组 $\mathcal{T}_i$ 活跃词元上进行中心化 15：使用固定的 w 对公式（S1）执行 AdamW 更新步 [85]

F.2. CICA 的单图像输入

输入 $x^{(k)}$ 保留图像 k 的视觉标记，并从上下文中移除其他图像的视觉标记。文本、每幅图像的起始和结束词元，以及所有保留词元的位置索引，均与 x相同，因此只有其他图像被移除。这样，视觉编码器只会对图像 k进行编码。在精确计算下，删除其他图像的视觉标记与在注意力层中屏蔽它们会得到相同的对数概率，同时也能避免它们进入线性注意力层的循环状态。所有响应和图像的序列会批量处理，长度最多为 32,768 个词元（包括填充）；所有单图处理轮次均使用正在训练的参数，并且不计算梯度。

F.3. CICA 中的答案词元CICA

对于每个响应，ans$_i$ 是奖励所评分的答案。如果响应中有一行“Answer: ...”（也包括“Final answer:”，可带项目符号或粗体），ans$_i$ 涵盖从最后一行此类内容中提取的答案；提取时去掉任何 **、圆括号、引号、花括号或 \boxed{，且不保留句末句号或方括号。如果没有此类行，ans$_i$ 就是最后一个\boxed{}中的内容；或者，当整个响应只有一行且不超过 40 个字符时，就是其中的裸答案。如果无法定位答案（例如，响应中的思考始终没有结束），

表 S8。**CICA 训练步的设置**。CICA 在每个骨干网络上均基于 GRPO 构建，跨越多个列的数值适用于每个骨干网络。

	Qwen2.5-VL-7B [7]	Qwen3.5-9B [105]	GLM-4.1V-9B-Thinking [43]
项目 B，组大小 G	8, 8		32，8
展开采样器	Transformer [140]		vLLM [58]
温度；top-p		1；	
Top-k；重复惩罚	50；1.05	0.95	20；无
展开预算	最多 3,072 个词元		8,192 个词元
学习率（余弦调度 [84]）	5×10^{-7}		3×10^{-7}
CICA β; τ	0.95，在前 32 步内逐渐增加；0.5		
单图像输入	删除其他图像，批处理（≤32，768 个词元）		
每步更新次数	1		
策略代理目标	无比率；采样器权重 $\min(\pi_\theta/\mu, 2)$		
归一化	每个响应，然后 N		
与参考模型的 KL散度	k3 [111], $\beta_{\mathrm{KL}} = 0.02$		
过长惩罚	none		
准确率相同的组	已丢弃，抽取下一条目		
尝试次数上限 M 每步	256		128
最少保留组数 $B_{\min}$	4		12
最大长度	32,768		
梯度裁剪；优化器	范数 1；AdamW，不使用权重衰减		
冻结	视觉编码器和投影器		

或者其裸答案不符合该条目的规范格式，则会获得 $s_i = 0$。对于先思考再作答的响应，ans_i 取自思考结束后的文本；如果聊天模板标记了答案框，则取自该答案框。答案标记从最后一行答案的开头延伸至 ans_i。每个词元都使用各自的最佳单张图像（公式（3））， 而ans_i 用于定位答案及其标记。

F.4. 活跃词元与 CICA 权重

在答案前进行推理的响应中，活跃词元是除答案标记、答案标签、回合结束标记和停止令牌、开启或关闭思考的词元，以及思考结束后紧接的词元之外的所有词元，而 $\mathcal{T}_i$ 保留这些词元。仅给出答案的响应也会将其答案 ans_i 作为活跃词元，但数量太少，无法进行比较。因此，集合$\mathcal{T}_i$ 会将这些词元与同组中其他以相同方式作答的仅含答案的响应中的词元合并，这些作答方式包括答案行（包括 \方框标注{}）或裸答案。同一条目的响应会一起处理，只有一个响应的词元池则会被赋予 $u = 0$。

指示变量 s_i 在响应正确（准确率为 1）时取 1， $A_i > 0$ 在响应错误时取 1。其他情况下取 0，例如上文提到的部分正确答案，其得分高于组均值，或响应没有可用答案。对于响应$^-$，若 $s_i = 0$，则保留 $w = 1$，且不加入任何池。在 $\mathcal{T}_i$ 中， $\bar{\sigma}_i$ 和 h_i 是平均值，其中响应 j 的每个词元都按$a_j = |A_j|/(NL_j)$ 计入，即其在更新中的份额，因此在单个响应内，它们就是普通平均值。由于 $\sum a_j u_{j,t} = 0$ 相对于$\mathcal{T}_i$，权重不会改变更新的总规模 $\sum a_j w_{j,t}$，即重要性权重 η应用前相对于 $\mathcal{T}_i$ 的更新规模；并且，对于进行推理的响应，权重在其活跃词元上的平均值为 1。因此，池化后的答案会相对于其他答案重新加权，这些答案对应其各自的条目；当完整输入相较于这些答案中的最佳单图像更支持某个答案时，该答案在该条目更新中所占的份额就更大。在第 n 步，权重采用 $\beta\min(1, (n-1)/32)$，并在第 33 步达到 0.95。

F.5. CICA 对比项的选择

CICA 将完整输入与其中的最佳单图像进行对比，而不是与平均图像、随机单张图像或减去一张图像后的完整输入进行对比。因为 $k^*_{i,t}$ 会最大化每个词元的似然（式（3）），$\sigma_{i,t}$ 是任意一张图像在该词元处所能产生的最小对比项；只有当完整输入使该词元比任何单张图像都更可能时，这个对比项才为正。跨多张图像进行推理的响应可以在每一步利用不同的图像。对于单令牌答案，最能解释该答案的图像就是该词元对应的最佳单图像。平均图像或随机单张图像（如单独的文本）会让从一张图像中直接读出的答案获得较高的σ，因为其他图像和单独的文本都无法支持该答案。留一对比项会隐藏一张图像；对于任何依赖该图像的词元，包括仅从该图像中直接读出的答案，它的值都很高。这些对比项都会把功劳归于单图捷径。与最佳单图像相比，只有当其他图像为信息量最大的图像提供补充时，σ 才会很高。隐藏其他图像会保留文本、图像分隔符以及所有位置，因此，简化后的输入在每个活跃词元上造成的对数概率变化相同，这种变化会在 $\sigma - \bar{\sigma}$中抵消。

基础模型给出的答案印证了这一点。对于表 S15中的条目，我们按照公式（σ）计算4中定义的），针对 Qwen2.5-VL-7B 或 Qwen3.5-9B 在输入所有图像时给出的答案计算，并将每个答案的概率在该条目的

有效答案范围内重新归一化。对于正确答案，在审计因一张图像即可作答而移除的条目中，两种模型的中位数σ 均为 0.0004，而在 ASTERISM 条目中分别为 0.12 和 0.19。在正确回答的“已移除条目”与正确回答的 ASTERISM 条目组成的配对中，有 72% 和 73% 的情况下，ASTERISM 答案的 σ更大。以单张图像对数似然的平均值为参照时，这一比例分别为 62% 和 66%；仅以文本为参照时，则为 39% 和 44%，因为这两种参照对从单张图像中读出的答案都赋予较高的 σ。表 5 中的两行分别用纯文本和单张图像对数似然的平均值替代最佳单图像。

F.6. 真实响应上的 CICA 权重

图 S14 展示了 CICA-Qwen2.5-VL-7B、CICA-Qwen3.5-9B 和 CICA-GLM-4.1V-9B-Thinking 在不同 ASTERISM 条目上的完整正确响应，并像图 6那样将其切分为若干步，每一步均根据其词元的平均权重进行着色w （公式（5））。该权重由训练代码使用训练后的模型计算。涉及图像间关联的步骤，例如比较三角形的边长和角度、对比不同照片中杯子上的文字，或追踪不同帧中的立方体，颜色最深；而大多数描述单张图像的步骤颜色较浅。

F.7. CICA 的成本

CICA 不增加参数，也不需要反向传播。对于每条响应，它会对 K 单图像输入进行计算，且不计算梯度。这些输入的视觉标记总数与 x相同，而文本会重复 K 次。GRPO 的一个训练步本来就会对响应进行采样，并对每条响应执行一次参考前向传播、一次前向传播和一次反向传播。单图处理轮次会复用这些采样得到的响应，因此训练步无需额外采样，也无需额外反向传播；只有这些前向传播会增加耗时。在 Qwen2.5-VL-7B 上，每个提示作为一个批次运行时，这些轮次耗时约占一个训练步的 9%。图像数量增加时，CICA 的单步计算量增长速度与 GRPO 的单步相近，而非达到 K 倍（图S15）。对于每个提示中的所有采样响应，视觉编码器只运行一次。每张图像只会进入 K 个单图像输入之一，因此单图处理轮次会对每条响应中的每张图像各处理一次；而 GRPO 的一个训练步本来就会对每张图像执行参考前向传播，以及每条采样响应各自的前向传播和反向传播。在一个训练步中，每增加一张图像，GRPO 训练步的耗时会增加其中位数的 1.1%；单图处理轮次则增加 0.14%，约为其八分之一。

单图处理轮次不保留任何激活值，并以有界批次运行；在 Qwen3.5-9B 上，CICA 的峰值已分配内存不会超过 GRPO。

G. 训练细节

所有运行均从公开检查点开始，并保持视觉编码器和投影器冻结；同一骨干网络上的运行共享奖励和提示词。CICA 在每个骨干网络上训练一个训练周期，遍历 ASTERISM-Train 的全部 49,105 个项目，设置见表 S8，强化学习基线见补充材料 G.3 。

G.1. 骨干网络

我们使用 CICA 训练 Qwen2.5-VL-7B [7], Qwen3.5-9B [105]和 GLM- 4.1V-9B-Thinking [43]。对于每个骨干模型，CICA 都会重新加权 GRPO 的优势 [114]，并对基础模型施加 0.02 的 KL 惩罚（公式（S1））。这三个骨干模型在语言模型、视觉编码器、位置编码和预训练方面各不相同。与 Qwen2.5-VL-7B 不同，Qwen3.5-9B 在每四个语言模型层中有三个使用线性注意力，并采用不同的视觉编码器和位置编码，以及早期融合多模态预训练。GLM-4.1V-9B-Thinking 属于另一个模型家族，并且像 Qwen2.5-VL-7B 一样，在每个语言模型层中都使用全注意力。对于每个骨干模型，CICA 都以相同的方式构造单图像输入，即删除其他图像的视觉标记（补充材料F）。

G.2. 训练数据

ASTERISM-Train（第 3.4 节）包含 49,105 个来自 48 个来源且答案可验证的多图像条目，并已移除与 ASTERISM 图像近重复的内容。其中 26,443 个条目来自十个公共数据集的训练拆分（MindCube [132],PerceptionTest [103], MagicBrush [163], NLVR2 [117],WebQA [13], SAT [109], CLEVRER [156], STAR [141],CLEVR-Change [102] 和 Spot-the-Diff [47]）。另有 9,570 个条目来自多图像指令集合（M4-Instruct [67] 6,467, Mantis-Instruct [49] 1,049, LEOP-ARD [48] 2,039, 以及其他 15 个来源），还有 13,092 个条目是基于公开图像构建的合成条目（拼图、取证、艺术风格与相似性匹配、视觉和语义对应、计数、大海捞针、找异类、大小比较和帧排序任务，以及针对常见物体图像的问题模板）。

当训练图像与一张 ASTERISM 图像的 64 位差异哈希 [57] 相差不超过 6 位时，二者构成一个候选对；将这两张图像转换为灰度并调整大小至 32×32 像素后，在减去平均值的情况下，若相关系数达到 0.99 或更高，则该候选对属于近重复项。近重复项包括完全相同的帧，以及同一场景的其他帧，例如 Spot-the-Diff 中由同一固定摄像机拍摄的帧，或同一 CLEVR [51] 场景的渲染图。此检查会从训练池中移除 1,756 个项目。

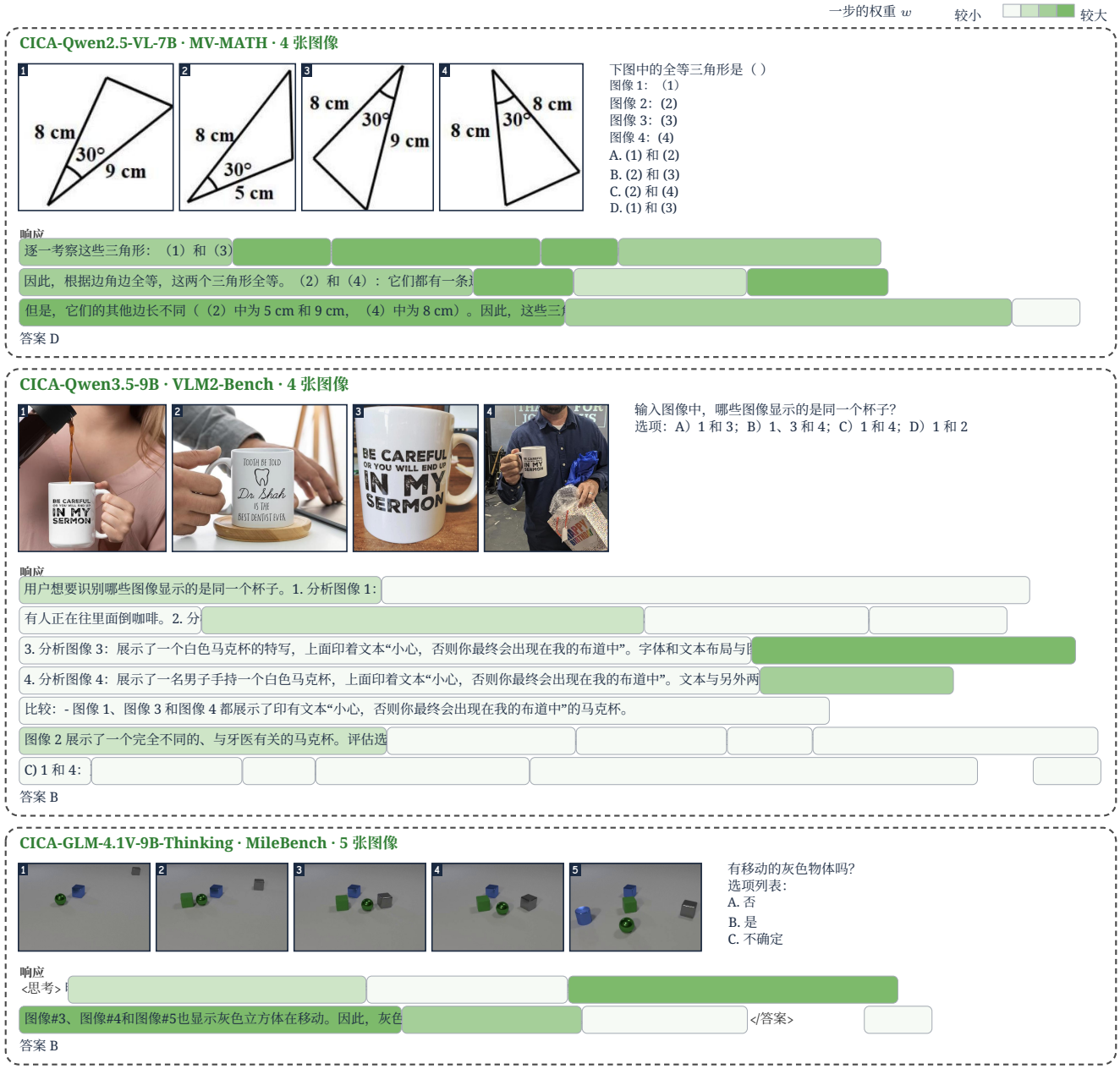


图 S14。**CICA 对完整响应的权重**。CICA-Qwen2.5-VL-7B（上）、CICA-Qwen3.5-9B（中）和 CICA-GLM-4.1V-9B-Thinking（下）在 ASTERISM 条目上的正确响应，按图中所示切分为若干步。图 6 中每一步都根据其词元的平均权重进行着色 w（公式（5））， 权重越大，颜色越深。灰色文本以及开启最后一个框的答案标记不计权重。图中未显示提示词中的答案格式说明。

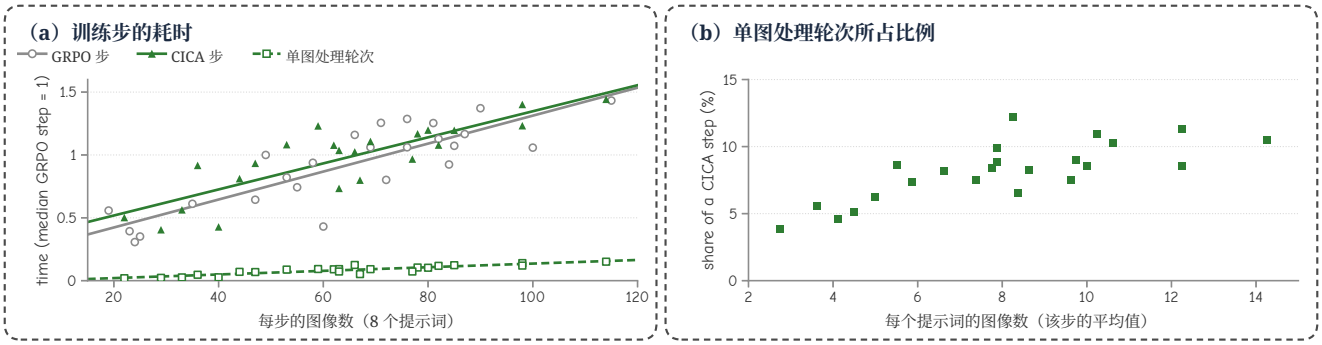


图 S15. **训练时间与图像数量**。在 Qwen2.5-VL-7B 上并行运行于同一节点的 GRPO 和 CICA 步数，与每步八个提示词所对应的图像数量之间的关系。(a) GRPO 中位步时长为单位的 CICA 每步时长和单图处理轮次时长，并绘出最小二乘拟合线。(b) CICA 每步中单图处理轮次所占比例，与该步每个提示词的平均图像数量之间的关系。

G.3. 强化学习基线

在每种骨干网络上，这五种强化学习基线在数据、准确率奖励、提示词、组大小、批量大小和学习率计划方面均与 CICA 保持一致，其余设置则遵循各自论文或官方代码。每次运行均报告最后一个检查点的结果。对于多图条目，VGPO [138],APPO [24] 和 CFPO [158] 会基于其中所有图像的视觉标记进行处理。它们的官方实现能够处理每个输入中的多张图像或视频帧。VPPO [45] 和 PAPO [137] 使用其发布的检查点进入表S9 。GRPO [114]和 DAPO [157] 会为响应中的每个词元赋予相同的优势；而 VGPO、APPO 和 CFPO 的视觉项则会偏向那些利用图像、关键帧或关键视觉线索的词元，但没有任何一项会特别区分直接从一张图像中读取答案的情况。

G.4. 消融实验

表 5 按照表 4中的设置，从 Qwen2.5-VL-7B 的公开检查点出发，在 ASTERISM-Train 的全部 49,105 个条目上进行训练。它们在此骨干网络上采用 CICA 的奖励、提示词、组大小、批次大小和学习率计划；除 GRPO 行外，还采用其 β 和 τ（表S8）。GRPO（$\beta = 0$）和 CICA 行是表 4中的 70亿参数模型。所有行均使用相同的运行器和硬件进行评估，评估数据包括 ASTERISM 全部内容，以及图 7 中在灰色遮罩条件下的 1,000 个条目。表 S22 改变 β 和 τ。

H. 补充结果

H.1. 评估多模态大语言模型

表 S9 列出了表 2 中大多数模型以及更多基础模型和后训练模型在各任务类别上的准确率。在 58 个开源多模态大语言模型中，按宏平均值计算的排序与按微平均值计算的排序之间的肯德尔相关系数为 0.95（平均值†），与不含文本/网络类别的宏平均值排序之间的肯德尔相关系数为 0.96（斯皮尔曼相关系数分别为 0.994 和 0.996）。在这三种平均值下，Qwen3.6-27B、Qwen3.5- 9B 和 MiMo-VL-7B 均是这 58 个模型中排名前三的模型，而 CICA-Qwen2.5-VL-7B 的排名高于其他所有模型基于 Qwen2.5-VL-7B 后训练。表S10 比较了基于公开基座模型后训练的开源多模态大语言模型与这些基座模型，两者均在相同硬件上使用相同运行器进行评测。表 S11 和图S16 按标签类别细分了 58 个开源多模态大语言模型的准确率。

H.2. 单图恢复与错误答案

表 S12 显示，在每个图像数量区间内，单张图像在已移除条目上恢复的全图增益比例都高于在 ASTERISM 条目上的比例；这一结论是在对图像数量分布进行标准化后得出的，且不包含答案为图像索引的条目。图 S17展示了 Qwen3.5-9B 和 Qwen2.5-VL-7B 在图 3所列条目上的纯文本、单图和全图准确率。在 15 个模型中，这些模型在两组条目上的全图准确率都至少比纯文本准确率高 10 个百分点，因此公式（1）的分母较大；其中位数 ρ 在已移除条目上为 0.77，在 ASTERISM 条目上为 0.18；而 ρ 在 ASTERISM 条目上保持在 0.09 至 0.31 之间。第 2.2 节中两个中位数之差的 95% 区间，是通过先对条目、再对模型进行重采样，并重复 1,000 次得出的。

单图移除条目。审计中认定可仅凭一张图像作答而移除的条目，其在 ρ 第2.2节的23个多模态大语言模型上的中位数为0.80，而 ASTERISM 条目的中位数为0.18；每个模型在已移除条目上的恢复表现都更好。

使用其他条目的图像时，这一差距依然存在。表S13列出了ρ 在23个多模态大语言模型上的结果：用其他条目的图像替换灰色图像，并与灰色遮罩下的结果并列。对于这两种设置，每个模型在已移除条目上的恢复表现都更好。

错误答案也与单张图像有关。表 S14 给出了第 2.2 节对31个开放源代码多模态大语言模型的匹配比较，即图S18中的23个模型（不包括 MindCube-3B） [132], 以及另外九个模型。对于表1中的 ASTERISM 条目，在双选项条目上，匹配与随机机会一致；而在包含三个或更多选项的条目上，Qwen3.5-9B、Qwen2.5-VL-7B 和 GLM-4.1V 的匹配率分别高出随机机会30.0、41.0和30.0个百分点——

表 S9. **ASTERISM 按任务类别划分的排行榜。** 表 2 中的大多数模型，以及其他基础模型和后训练模型，在 ASTERISM-Lite（轻量版）上的平均表现。列、分组和 § 标记与表 2相同；每组中每列四舍五入前的最佳值以粗体显示，次佳值加下划线。 †微平均值，即 ASTERISM 所有项目的准确率。Qwen3.5-9B 和 CICA-Qwen3.5-9B 在启用思考的情况下作答。

		ASTERISM									Lite
模型	Size	Agg	Cmp	Rsn	Spa	Tmp	Txt	Ret	Avg	Avg†	Avg
基础模型											
Gemma-3-4B [34]	4B	33.1	25.9	38.8	31.1	25.7	41.6	30.8	32.4	31.6	14.2
Gemma-3-12B [34]	12B	45.4	34.9	44.2	37.7	28.6	46.8	38.5	39.4	38.0	22.9
Gemma-4-12B [35]	12B	53.6	42.6	46.5	43.3	36.1	75.3	49.2	49.5	44.3	29.6
GLM-4.1V-9B-Thinking [43]	9B	64.3	58.2	66.4	51.8	44.7	70.1	55.3	58.7	56.3	37.7
Idefics2-8B [61]	8B	25.1	25.1	31.6	27.3	25.0	33.1	18.1	26.5	26.5	11.5
InternVL3-2B [170]	2B	34.2	30.1	30.2	27.9	34.7	41.6	27.6	32.3	30.4	15.1
InternVL3-8B [170]	8B	49.5	43.5	39.9	37.7	42.8	44.8	38.3	42.4	40.7	24.7
LLaVA-Interleave-7B [67]§	7B	37.0	37.8	30.4	29.6	25.9	35.7	17.0	30.5	29.0	12.3
LLaVA-OneVision-7B [65]	7B	36.4	30.8	33.0	31.1	22.1	35.7	18.9	29.7	28.7	12.5
LLaVA-OneVision-2-80亿参数 [4]	8B	40.4	40.1	41.5	38.7	37.0	42.9	34.2	39.3	38.8	19.5
MiMo-VL-7B [147]	7B	65.0	63.9	67.3	52.5	40.0	72.7	54.5	59.4	56.5	36.3
mPLUG-Owl3-7B [154]§	7B	29.1	26.5	30.4	27.8	30.6	33.1	20.0	28.2	28.0	14.7
Pixtral-12B [3]	12B	42.3	33.6	32.7	35.1	23.3	43.5	23.1	33.4	31.1	17.5
Qwen2-VL-7B [128]	7B	40.6	33.6	36.9	36.9	34.2	40.3	38.9	37.3	36.5	18.7
Qwen2.5-VL-3B [7]	3B	32.8	24.2	33.2	29.7	34.6	44.2	33.0	33.1	31.5	10.5
Qwen2.5-VL-7B [7]	7B	45.8	38.3	40.0	42.0	36.1	40.3	42.2	40.7	40.2	27.2
Qwen2.5-VL-32B [7]	32B	54.8	58.8	52.1	48.2	37.7	53.9	45.9	50.2	48.7	28.3
Qwen3-VL-2B [104]	2B	38.0	30.8	34.7	35.2	32.6	46.8	37.4	36.5	34.6	21.0
Qwen3-VL-4B [104]	4B	44.3	45.4	42.5	43.6	37.0	47.4	44.7	43.6	42.5	25.5
Qwen3-VL-4B-思考 [104]	4B	62.1	56.1	56.6	47.1	35.2	61.7	48.2	52.4	49.5	35.9
Qwen3-VL-80亿参数 [104]	8B	50.9	46.8	44.8	47.0	40.0	54.5	47.7	47.4	45.5	26.9
Qwen3-VL-30B-A3B [104]	30B	55.8	54.4	52.4	44.8	41.0	62.3	51.0	51.7	48.8	32.0
Qwen3-VL-32B [104]	32B	62.1	56.8	66.0	50.8	46.0	64.3	54.1	57.2	55.7	35.8
Qwen3.5-9B [105]	9B	77.9	64.3	71.9	55.3	51.2	77.9	61.3	65.7	62.1	43.3
Qwen3.6-27B [106]	27B	78.7	64.1	72.6	58.4	49.6	77.3	62.9	66.2	63.0	40.0
后训练模型											
来自Idefics2-8B [61]		25.1	25.1	**31.6**	27.3	25.0	33.1	18.1	26.5	**26.5**	**11.5**
Mantis-80亿参数-Idefics2 [49]§	8B	**35.0**	**31.1**	13.6	**30.6**	**26.4**	**35.1**	**19.0**	**27.3**	24.4	10.0
来自Qwen2-VL-7B [128]		40.6	33.6	**36.9**	36.9	34.2	40.3	38.9	37.3	36.5	18.7
Migician-7B [72]§	7B	**45.2**	**39.1**	35.7	**38.2**	**36.2**	**47.4**	**40.6**	**40.3**	**38.1**	**27.5**
UniVG-R1-70亿参数 [8]§	7B	38.5	32.2	35.5	34.2	33.4	35.7	36.3	35.1	34.7	18.1
来自Qwen2.5-VL-3B [7]		**32.8**	24.2	**33.2**	**29.7**	**34.6**	**44.2**	**33.0**	**33.1**	**31.5**	**10.5**
HATCH-3B [97]§	3B	28.2	25.5	31.0	27.3	33.1	41.6	21.1	29.7	28.5	10.1
MindCube-3B [132]§	3B	29.0	**25.9**	28.8	28.6	27.9	34.4	27.4	28.9	28.1	9.6
来自Qwen2.5-VL-7B [7]		45.8	38.3	40.0	42.0	36.1	40.3	42.2	40.7	40.2	27.2
Visual-Jigsaw-7B [143]§	7B	44.0	40.6	40.2	40.7	36.1	40.3	40.4	40.3	39.9	23.6
VLAA-Thinker-7B [14]	7B	44.5	39.0	40.3	42.7	36.6	40.3	41.9	40.8	40.5	25.7
NoisyRollout-7B [78]	7B	46.4	40.3	40.4	42.8	36.2	39.6	42.0	41.1	40.7	27.3
VPPO-7B [45]	7B	47.1	41.3	41.9	42.5	36.3	39.0	42.4	41.5	41.3	25.3
R1-ShareVL-7B [153]	7B	47.5	42.6	42.3	42.9	36.8	39.6	41.4	41.9	41.6	24.2
PAPO-G-H-7B [137]	7B	47.1	41.5	42.0	43.2	36.3	40.9	**42.7**	42.0	41.5	25.3
PeRL-7B [165]§	7B	46.6	44.0	41.7	40.8	36.8	44.2	40.7	42.1	41.0	27.0
PAPO-D-7B [137]	7B	45.8	42.3	43.1	42.6	36.4	42.9	42.0	42.2	41.7	27.7
CICA-Qwen2.5-VL-7B（我们的方法）	7B	**52.1**	**47.1**	**44.3**	**47.2**	**44.0**	**50.0**	42.4	**46.7**	**45.5**	**29.9**
来自GLM-4.1V-9B-Thinking [43]		64.3	58.2	66.4	51.8	44.7	70.1	55.3	58.7	56.3	37.7
CICA-GLM-4.1V-9B-Thinking（我们的）	9B	**69.3**	**61.1**	**67.1**	**52.3**	**46.4**	**77.3**	**55.4**	**61.3**	**57.6**	**40.2**
来自Qwen3.5-9B [105]		77.9	64.3	71.9	55.3	51.2	77.9	61.3	65.7	62.1	43.3
CICA-Qwen3.5-9B（我们的）	9B	**80.0**	**68.3**	**80.7**	**60.8**	**58.4**	**81.8**	**64.0**	**70.6**	**68.0**	**47.0**

表 S10。**相较于基础模型的提升**。 基于公开基础模型后训练的开源多模态大语言模型及其基础模型的 ASTERISM 平均值（平均值）：针对多图像任务训练的模型（上）；针对推理、视频、空间或感知任务进行后训练的模型（大多采用强化学习）（中）；以及 CICA（下）。每对模型使用相同的硬件和运行器，但各对模型分别运行，因此同一基础模型在不同配对中的得分可能略有差异。 Δ 是舍入前的差值，† 表示其 95% 区间不包含零的 Δ 。

模型	基础模型	Base	模型	Δ
Migician-7B [72]	Qwen2-VL-7B [128]	37.3	40.3	+3.1†
UniVG-R1-70亿参数 [8]	Qwen2-VL-7B	37.3	35.1	−2.1†
Visual-Jigsaw-7B [143]	Qwen2.5-VL-7B [7]	40.5	40.3	−0.2
PeRL-7B [165]	Qwen2.5-VL-7B	40.5	42.1	+1.6†
HATCH-3B [97]	Qwen2.5-VL-3B [7]	33.1	29.7	−3.4†
MindCube-3B [132]	Qwen2.5-VL-3B	33.1	28.9	−4.2†
Mantis-80亿参数-Idefics2 [49]	Idefics2-8B [61]	26.5	27.3	+0.8
Video-R1-7B [29]	Qwen2.5-VL-7B	40.6	37.4	−3.1†
VideoRFT-7B [130]	Qwen2.5-VL-7B	40.6	38.2	−2.3†
Open-o3-Video-7B [89]	Qwen2.5-VL-7B	40.6	39.8	−0.8
Spatial-SSRL-7B [82]	Qwen2.5-VL-7B	40.6	40.8	+0.3
Time-R1-7B [136]	Qwen2.5-VL-7B	40.6	41.0	+0.4
ArrowRL-7B [150]	Qwen2.5-VL-7B	40.6	41.2	+0.6
ThinkLite-VL-7B [135]	Qwen2.5-VL-7B	40.6	40.7	+0.2
NoisyRollout-7B [78]	Qwen2.5-VL-7B	40.6	41.1	+0.5†
Vision-SR1-70亿参数 [73]	Qwen2.5-VL-7B	40.6	38.4	−2.2†
Vision-Zero-7B [131]	Qwen2.5-VL-7B	40.5	40.4	−0.2
R1-ShareVL-7B [153]	Qwen2.5-VL-7B	40.5	41.9	+1.3†
SophiaVL-R1-7B [27]	Qwen2.5-VL-7B	40.5	40.9	+0.4
Perception-R1-7B [146]	Qwen2.5-VL-7B	40.5	40.7	+0.2
VisionReasoner-7B [81]	Qwen2.5-VL-7B	40.5	40.6	+0.1
TON-7B-Math [127]	Qwen2.5-VL-7B	40.5	38.1	−2.4†
SpatialReasoner-7B [86]	Qwen2.5-VL-7B	40.5	34.4	−6.2†
VLAA-Thinker-7B [14]	Qwen2.5-VL-7B	40.5	40.8	+0.2
SpatialLadder-3B [69]	Qwen2.5-VL-3B	33.2	29.1	−4.1†
CapRL-3B [149]	Qwen2.5-VL-3B	33.2	32.8	−0.4
Embodied-R1-3B [159]	Qwen2.5-VL-3B	33.2	33.4	+0.2
GRIT-3B [28]	Qwen2.5-VL-3B	33.2	31.7	−1.5†
Mulberry-7B [152]	Qwen2-VL-7B	37.3	34.0	−3.3†
TVC-7B [118]	Qwen2-VL-7B	37.3	31.8	−5.4†
CICA-Qwen2.5-VL-7B（我们的）	Qwen2.5-VL-7B	40.7	46.7	+6.0†
CICA-Qwen3.5-9B（我们的）	Qwen3.5-9B [105]	65.7	70.6	+4.9†
CICA-GLM-4.1V-9B-Thinking（我们的方法）	GLM-4.1V-9B-Thinking [43]	58.7	61.3	+2.6†

9B-Thinking。

H.3. 使用 CICA 进行训练

表 S9给出了各骨干网络及其使用 CICA 训练后的版本在各任务类别上的准确率。在 ASTERISM 上，配对自助法 [26] 95% 项目区间不包含零，表明 CICA 与表中其他每一行之间的差异4。

表S16 列出了图7 中三种骨干模型相对于基础模型和五种强化学习基线在各掩码条件下的差异；图7中的误差线显示了基于项目进行配对自助法计算得出的 95% 区间。对于 Qwen2.5-VL-7B 相对于基础模型、GRPO 和 DAPO 的结果，表S17列出了在模型内部隐藏其他图像（如 CICA 所做的那样）（a）或将其替换为来自其他项目 (b)。

H.4. 单图基准测试

在十个涉及幻觉、真实世界图像、高分辨率图像、光学字符识别和图表的单图像基准测试中，表S18 比较了各骨干网络与使用 CICA 训练的同一骨干网络。CICA 提高了每种骨干网络在这十个基准测试上的平均得分（平均值）。

H.5. MMIU

表S19 给出了 Qwen2.5-VL-7B 和 Qwen3.5-9B 在 MMIU 上的 [88] 得分，以及表4 中的五种强化学习基线和使用 CICA 训练的两种骨干模型。MMIU 基于现有数据集构建了 52 项多图像任务，并将其发布为 60 个任务划分。在两种骨干模型上，CICA 的得分均高于基础模型

表 S11。**组合图像的项目属于最难的项目之一**。对于以下标签类别（每类至少包含 200 个项目），准确率相对于每个模型自身平均值的百分点差异，按 58 个开源多模态大语言模型的平均值计算（Δ），以及其 95% 区间低于（更难）或高于（更容易）零的模型所占比例。标签采用补充材料 B.4中的标签。

Axis	类别	项目	Δ	更难	更容易
跨图像操作：合并或组合		434	−15.5	95%	0%
	对齐或配准	731	−11.7	95%	0%
	转移或应用	939	−12.3	84%	10%
	排列序列	1,015	−12.8	97%	0%
	查找差异	1,888	−8.5	86%	3%
	比较	3,498	+6.2	0%	100%
关系	空间的	3,104	−7.3	97%	0%
	逻辑或形式的	3,565	−6.1	84%	9%
	语义	7,800	+7.2	0%	100%
每项操作次数	三个或更多	1,756	−3.8	71%	3%
	two	8,386	−0.7	38%	7%
	one	8,471	+1.5	5%	66%
推理跳数	四个或更多	7,351	−4.1	98%	0%
	two	1,779	+8.6	2%	90%
时间尺度	分钟到小时	648	−16.7	98%	0%

和所有五种强化学习基线。

H.6. 训练后的捷径

图 S19 在图 2.2 中，对每个使用 GRPO 或 CICA 训练的骨干网络，在图 7的灰色遮罩条件下，重复了错误答案测试。经过 CICA 训练后，各骨干网络中重复单图像答案的错误答案占比均低于 GRPO 训练后的占比，GLM-4.1V-9B-Thinking 上最多低 7.7 个百分点 [43]。

H.7. 增益的可靠性

在 18,696 个项目上，模型的 ASTER- ISM 得分的标准误最多为 0.7 个百分点。将两个模型在相同项目上的答案配对后，对于 CICA 与基础模型以及与表 4中最强的强化学习基线之间的得分差，标准误降至 0.4–0.6 个百分点。在每个骨干网络上，按任务类别对项目进行重采样（2,000 次重采样）后，两项差值的配对自助法 95% 置信区间均至少高于零 1.3 个百分点。微平均（表† 中的平均值S9），即对所有项目赋予相同权重，以及不包含最小类别文本/网络的宏平均，也都显示在每个骨干网络上，CICA 的得分高于基础模型和最强的强化学习基线。剔除 1,406 个改写条目后，在 Qwen2.5-VL-7B、GLM-4.1V-9B-Thinking 和 Qwen3.5-9B 上，CICA 的得分仍分别比基础模型高 2.8 至 5.8 个百分点，比最强的强化学习基线高 2.3 至 3.6 个百分点，配对区间也都高于零。在除检索之外的六个任务类别上取平均后，每个骨干网络上的 CICA 得分也高于基础模型和全部五个强化学习基线。

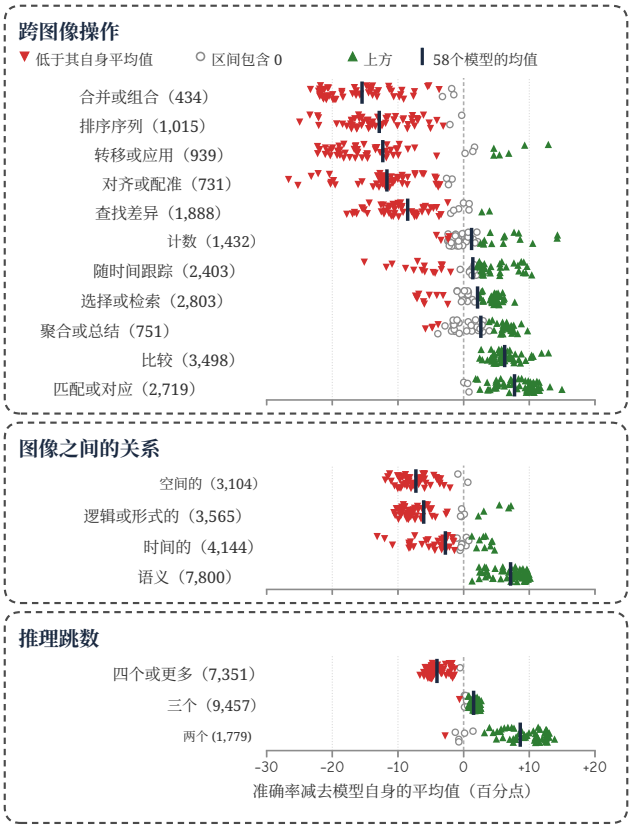


图 S16. **开源多模态大语言模型最难的条目类型**。对于表 S9 和 S10中的 58 个开源多模态大语言模型，在至少包含 200 个条目的跨图像操作、图像关系和推理跳数类别中，计算各标签类别的准确率与模型在所有已标注 ASTERISM 条目上的准确率之差（括号内为条目数）。向下的红色三角形表示模型的 95% 区间低于零，向上的绿色三角形表示模型的区间高于零，灰色圆圈表示模型的区间包含零，横线表示模型的平均值。

I. 局限性与未来工作

ASTERISM 的图像和大部分问题取自 19个来源，因此其领域和任务构成也随之而定。仅其 1,406 个改写以及基于这些图像新增的人工核验问题，规模就超过了大多数来源（表 S2）。后续版本可以纳入其他领域的来源。

ASTERISM 得分对七类任务取宏平均，避免规模最大的任务类别占据主导（第 3.2 节），从而让规模最小的文本/网络获得相同权重。我们还报告了微平均，即为所有条目赋予相同权重（表 S9）。按此平均方式计算，或去掉文本/网络后，开源多模态大语言模型的排序几乎不变，而且在每种骨干网络上，CICA 的表现都优于基础模型和最强的强化学习基线（补充材料 H.1 和 H.7）。来源标识符还允许用户删除任意来源或调整其权重（补充材料 J.2）。

对超过 32,000 个源条目进行审计（表S2）时，Claude Fable 5 和 GPT-5.6-Sol 这两个来自

表 S12. **单图恢复率 ρ 按图像数量划分。** 图 3 中已移除项目与 ASTERISM 项目的对比 在灰色遮罩下，显示 95% 置信区间；在图像数量行中， n 括号内为数值（五图项和六图项已合并）。最后几行涵盖所有图像数量（全部）、按 ASTERISM 图像数量分布重新加权的已移除项目（五张和六张图像分别计数，K 标准化），以及答案不是图像索引的项目（无索引答案）。当一张图像及其灰色伴随图像的得分低于纯文本输入时，数值为负；当其得分高于所有图像时，数值大于 1。

	Qwen3.5-9B [105]		Qwen2.5-VL-7B [7]	
图像	已移除	ASTERISM	已移除	ASTERISM
2	0.55 [0.46, 0.63] (298)	0.38 [0.22, 0.52] (297)	0.75 [0.64, 0.88] (298)	−0.08 [−0.70, 0.22] (297)
3	0.17 [−0.39, 0.45] (94)	−0.13 [−1.11, 0.26] (58)	0.91 [0.73, 1.18] (94)	0.22 [−0.36, 0.55] (58)
4	0.58 [0.25, 0.93] (83)	−0.04 [−1.22, 0.31] (107)	0.87 [0.52, 1.34] (83)	0.14 [−0.29、0.40] (107)
5–6	0.69 [0.60, 0.79] (283)	0.04 [−0.12, 0.16] (210)	0.88 [0.78, 1.00] (283)	0.21 [−0.06, 0.44] (210)
all	0.58 [0.52, 0.64]	0.17 [0.06, 0.27]	0.84 [0.77, 0.91]	0.11 [−0.06, 0.25]
K 标准化	0.64	0.17	0.88	0.11
无索引答案	0.62	0.20	0.85	0.15

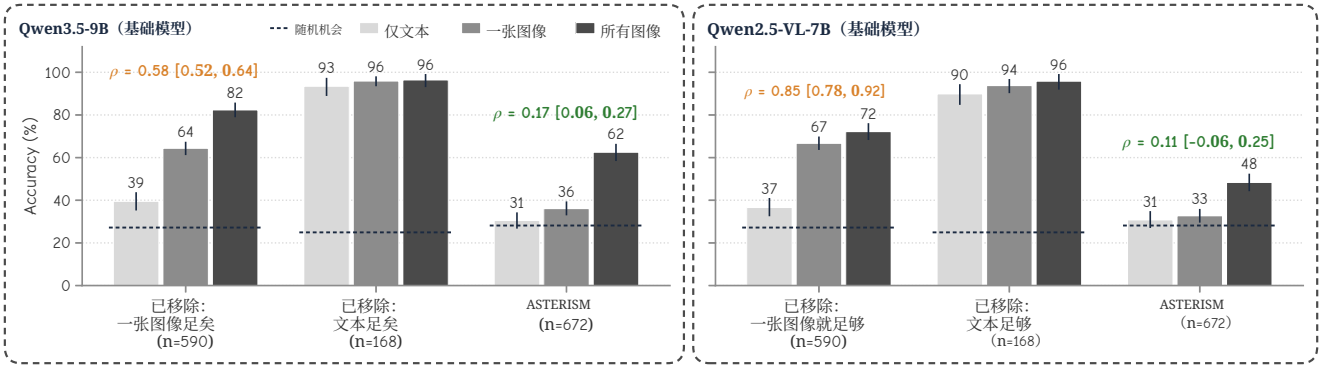


图 S17。 **审计移除的条目可以根据一张图像或文本作答；ASTERISM 条目则不行。** 对于图 3 中的多项选择题，图示比较了 Qwen3.5-9B 和 Qwen2.5-VL-7B 在将所有图像替换为灰色（仅文本）、保留一张图像（对保留的图像取平均）以及使用所有图像时的准确率。 3。误差线表示 95% 置信区间，虚线表示均匀猜测的准确率。恢复比率 ρ 表示在因一张图像足以作答而被移除的条目及 ASTERISM 条目上，一张图像所恢复的全图增益占比，并给出其 95% 置信区间；表 S12 给出了 ρ 在所有移除项上的结果。

不同模型系列且当时实力最强的多模态大语言模型给出判定结果，人工裁定者则解决双方意见不一致的情况，并处理升级提交的判定结果（补充材料B.1）。因此，对于未经人工审核的条目，其判定结果仍取决于这些模型；而在一个分层样本中，审阅者对审计结果的认同程度与对另一位审阅者意见的认同程度相同，同时人工会核查每个修正后的标签和重写内容（补充材料B.3）。单图筛选器同样依赖由四个开源多模态大语言模型组成的固定委员会；对于不在其中的 23 个多模态大语言模型，在 ASTERISM 条目上，一张图像所恢复的全图增益远少于在移除项上的表现（第2.2节）。

我们在三个骨干模型上训练 CICA，每个模型进行一次训练运行，结果见表 4 和5。在每个骨干模型上，CICA 的得分都高于所有强化学习基线，且领先幅度超过各强化学习基线之间的差距；与仅可见一张图像相比，所有图像均可见时，CICA 相对每个基线的提升更大（见表 4 和S16）。在 Qwen2.5-VL-7B 上，表 5和S22 中所有保留 σ 与图像绑定的六次 CICA 训练运行，其得分均高于所有强化学习基线。在 MMIU 上，CICA 的得分也高于所有强化学习基线，并且保留或提高了每一项单图得分（见表 S18 和 S19 ）。五个强化学习基线使用与 CICA 相同的 ASTERISM-Train 项目进行训练，并剔除了与 ASTERISM 图像近重复项（补充材料 G.2 和 G.3）。三个骨干模型在架构和预训练方面各不相同（补充材料 G.1），其中两个此前已通过强化学习进行后训练 [43, 105]。特别是，GLM-4.1V-9B-Thinking 经历了基于 STEM 问题到图形用户界面智能体的大规模强化学习 [43], , 且强化学习基线对其得分的影响小于对另外两个骨干模型的影响（表 4）。受计算资源限制，将 CICA 应用于更大的多模态大语言模型进行测试留待未来工作。

CICA 将功劳转移给需要多张图像的词元，但并不针对检索项；在这类条目中，任务是找到包含答案的那张图像或页面。它使三种骨干模型的检索准确率均有所提高：Qwen3.5-9B 提高了 2.7 个百分点，另外两种模型也略有提升（表 S9），而如何为搜索本身计功仍有待解决。

J. 数据卡

J.1. 许可证与发布

ASTERISM 遵循每个源基准测试的许可证（表 S20）。对于许可证允许再分发和商业用途的来源，我们会发布包含图像的条目，以及修正后的标签和答案格式、改写条目与逐条标签。对于

表 S13。**当其他条目的图像替换灰色图像时，所有多模态大语言模型的单图像恢复差距依然存在。** 单图像恢复比率 ρ 图中 23 个开源多模态大语言模型的 S18 在图 3的多项选择题上的表现，其中其他图像为灰色（Gray）或替换为其他条目的图像（互换）。

	Gray		互换	
模型	已移除 ASTERISM		已移除 ASTERISM	
Gemma-3-4B [34]	0.81	−0.04	0.73	0.07
Gemma-3-12B [34]	0.75	0.28	0.58	0.14
GLM-4.1V-9B-Thinking [43]	0.59	0.14	0.53	0.12
HATCH-3B [97]	0.87	0.09	0.82	−0.06
InternVL3-2B [170]	0.79	0.27	0.67	0.36
InternVL3-8B [170]	0.77	0.19	0.56	0.17
LLaVA-Interleave-7B [67]	0.77	0.31	0.62	0.05
LLaVA-OneVision-7B [65]	0.84	−0.11	0.64	0.15
Migician-7B [72]	0.76	0.21	0.65	0.19
MindCube-3B [132]	0.76	0.28	0.67	0.30
Pixtral-12B [3]	0.73	0.23	0.63	0.24
Qwen2-VL-7B [128]	0.82	0.16	0.63	0.11
Qwen2.5-VL-3B [7]	0.78	0.47	0.68	−0.06
Qwen2.5-VL-7B [7]	0.84	0.11	0.68	0.10
Qwen2.5-VL-32B [7]	0.73	0.18	0.64	0.11
Qwen3-VL-2B [104]	0.97	0.27	0.77	0.06
Qwen3-VL-4B [104]	0.80	0.09	0.70	0.15
Qwen3-VL-4B-Thinking [104]	0.57	0.12	0.55	0.11
Qwen3-VL-8B [104]	0.82	0.14	0.74	0.16
Qwen3.5-9B [105]	0.58	0.17	0.53	0.09
Qwen3.6-27B [106]	0.56	0.25	0.48	0.19
UniVG-R1-70亿参数 [8]	0.97	0.11	0.72	0.16
Visual-Jigsaw-70亿参数 [143]	0.79	0.22	0.71	0.17
中位数	0.78	0.18	0.65	0.14

对于许可条款限制商业用途或未注明许可证的来源，我们会发布条目标识符、我们的修正内容、改写后的问题和选项，以及用于从各来源公开发布内容中组装条目的构建脚本。如果某个来源的数据集卡片与代码仓库或论文中注明的许可证不同，我们会遵循限制更严格的许可证。与 ASTERISM 一同发布的还有 CICA-Qwen2.5-VL-7B、CICA-Qwen3.5-9B 和 CICA-GLM-4.1V-9B-Thinking 的权重，以及相应的配置、分词器和处理器文件。

J.2. 组成与预期用途

ASTERISM包含来自 19 个公开基准测试的 18,696 个条目，归为七类任务（表 S21），且每个条目都包含两张或更多图像。源条目只有通过两阶段单图筛选器，并经由两个多模态大语言模型和一名人工裁定者审计后，才会被纳入。审计还会更正标签、答案格式和文本。改写保留源条目的图像，且只有在任何单张图像都无法回答改写内容时才会纳入；人工审阅者会检查每个更正后的标签和每项改写（补充材料 B.1 至 B.3）。每个条目无论是以发布形式提供，还是由构建脚本汇编而成，都会记录其来源基准及来源条目标识符、图像排序、提示及任何选项、答案及其答案格式、任务类别、是否为改写，以及（如有）逐条标签（补充材料 B.4）。ASTERISM旨在用于评估多模态大语言模型；ASTERISM-Train则移除了与ASTERISM图像近重复的图像，是专为训练而设的数据集（第3.4节）。在 ASTERISM 上训练

表 S14。**错误答案会重复模型自己的单图答案。**对于包含两至六张图像的多项选择 ASTER-ISM 条目，针对 31 个开源多模态大语言模型，统计选择某个选项且存在匹配条目 (n) 的错误答案中，重复该模型某个单图答案的比例（匹配）；重复另一条目单图答案的比例，该条目与当前条目选项数相同、正确选项字母相同且图像数量相同（匹配机会水平）；二者之差及其 95% 区间（差值）；以及对其他模型取平均后，重复这些模型单图答案的比例（其他模型）。不列出具有 $n < 100$ 的模型。

模型	n	匹配	匹配机会水平	差异。		其他
Gemma-3-4B [34]	412	90.5	47.6	+42.9	[39.2, 46.5]	54.2
Gemma-3-12B [34]	399	83.5	48.1	+35.4	[31.3, 39.2]	50.3
Gemma-4-12B [35]	336	74.4	46.5	+27.9	[23.3, 32.7]	56.8
GLM-4.1V-9B-Thinking [43]	262	80.9	53.3	+27.6	[22.6, 32.3]	53.2
HATCH-3B [97]	386	86.3	41.0	+45.3	[41.3, 48.9]	54.5
InternVL3-2B [170]	160	83.8	27.9	+55.9	[49.3, 62.1]	51.0
InternVL3-8B [170]	320	84.1	47.2	+36.9	[32.8, 40.9]	52.5
LLaVA-Interleave-7B [67]	361	82.5	42.0	+40.5	[36.2, 44.6]	47.6
LLaVA-OneVision-7B [65]	393	87.8	43.7	+44.1	[40.4, 47.5]	56.8
LLaVA-OneVision-2-80亿参数 [4]	342	79.2	42.4	+36.8	[31.9, 41.6]	55.2
Mantis-80亿参数-Idefics2 [49]	320	80.6	34.8	+45.8	[41.3, 50.4]	54.3
Migician-7B [72]	259	86.5	39.6	+46.9	[42.2, 51.6]	59.1
MiMo-VL-7B [147]	269	74.0	47.9	+26.1	[20.8, 31.5]	51.3
mPLUG-Owl3-7B [154]	316	89.9	33.4	+56.5	[52.4, 60.2]	56.9
NoisyRollout-7B [78]	316	82.6	50.4	+32.2	[27.8, 36.6]	57.1
Pixtral-12B [3]	396	88.6	46.2	+42.4	[38.8, 45.8]	56.1
Qwen2-VL-7B [128]	187	81.8	28.8	+53.0	[48.0, 58.0]	60.6
Qwen2.5-VL-3B [7]	303	80.5	36.3	+44.2	[39.7, 48.8]	57.2
Qwen2.5-VL-7B [7]	321	84.4	49.5	+34.9	[30.8, 39.1]	57.9
Qwen2.5-VL-32B [7]	267	75.7	47.0	+28.7	[22.7, 34.4]	50.3
Qwen3-VL-2B [104]	330	84.5	49.1	+35.4	[31.0, 39.3]	55.3
Qwen3-VL-4B [104]	291	84.5	50.5	+34.0	[29.5, 38.5]	56.6
Qwen3-VL-4B-思考 [104]	199	67.8	43.1	+24.7	[18.2, 31.4]	54.5
Qwen3-VL-8B [104]	311	83.6	50.2	+33.4	[28.8, 37.6]	57.4
Qwen3-VL-30B-A3B [104]	307	79.8	46.7	+33.1	[27.9, 38.2]	53.0
Qwen3-VL-32B [104]	274	75.5	47.8	+27.7	[22.4, 32.9]	53.8
Qwen3.5-9B [105]	205	75.1	47.4	+27.7	[21.5, 33.6]	47.9
Qwen3.6-27B [106]	229	69.4	47.0	+22.4	[16.4, 28.6]	55.2
R1-ShareVL-7B [153]	306	81.0	48.2	+32.8	[28.4, 37.2]	54.4
UniVG-R1-7B [8]	100	79.0	19.1	+59.9	[51.2, 68.7]	55.6
Visual-Jigsaw-7B [143]	316	81.6	54.4	+27.2	[22.7, 31.7]	55.6

表 S15。**CICA 的单图像输入同样展现出这一差距。** 单图恢复率 ρ 当其他图像像 CICA 那样隐藏在模型内部时（第 4.1 节）。准确率是指在具有固定选项集的条目中，正确选项在有效答案中的精确概率（775 个 ASTERISM 条目和 796 个移除项；后者将图 3 中多项选择题条目之外的 是/否 和 真/假 条目纳入其中。3）。

模型	ASTERISM 条目	移除项	差异
Qwen2.5-VL-7B [7]	0.22	0.74	+0.52
Qwen3.5-9B [105]	0.11	0.66	+0.55

不在范围内的包括其中图像所展示人物的识别或画像，以及对来源带有非商业条款的项目进行商业用途（表S20）。ASTERISM 不收集新图像，而是继承这些图像以及其来源中的项目；除 1,406 条改写内容和我们修正了标签、答案格式或文本的 569 个项目外，来源中的项目可能包含可识别的人物和敏感内容。来源标识符和逐条标签使用户能够排除任何来源、类别或项目。我们通过版本化发布内容维护 ASTERISM，并在公开问题列表中跟踪报告的错误和内容问题。

表 S16。**CICA 在不同掩码条件下的增益**。图中 ASTERISM 条目的准确率差异（CICA 减去基线，以百分点计）。7以及四舍五入前所有图像与一张图像的增益之差。

骨干网络	vs.	仅文本	一张图像	所有图像	全部 − 单个
Qwen2.5-VL-7B [7]	base	+0.5	+1.1	+6.5	+5.5
	GRPO	−0.9	+0.0	+5.0	+5.0
	DAPO	+0.1	+1.0	+4.2	+3.2
	APPO	+0.3	+0.9	+3.5	+2.6
	VGPO	−0.1	+1.0	+4.4	+3.4
	CFPO	−1.3	−0.9	+3.5	+4.4
GLM-4.1V-9B-Thinking [43]	base	+1.1	+1.6	+2.9	+1.3
	GRPO	+2.0	+1.6	+3.7	+2.1
	DAPO	+0.9	+1.8	+3.7	+1.9
	APPO	+2.0	+0.3	+2.9	+2.6
	VGPO	+1.1	−0.5	+2.9	+3.4
	CFPO	+1.1	+0.6	+2.5	+1.9
Qwen3.5-9B [105]	base	−1.0	+0.9	+5.7	+4.8
	GRPO	−0.7	+0.6	+3.7	+3.1
	DAPO	+1.6	+2.3	+4.7	+2.4
	APPO	−0.5	+2.4	+4.6	+2.2
	VGPO	+0.5	+1.1	+4.4	+3.3
	CFPO	+1.4	+4.3	+8.1	+3.8

表 S17. **当其他图像在模型内部被隐藏（a）或替换为来自其他条目的图像（b）时，CICA 按条件划分的增益**。（a）与表 S16相同，但文本条件和单图像条件像 CICA 一样在模型内部隐藏图像（第 4.1 节），而非将其灰显；准确率是表 S15 中 775 个 ASTERISM 项目的有效答案里，正确选项的精确概率。（b）与表 S16相同，但第一种条件（无真实图像）将每张图像替换为来自其他条目的图像，第二种条件（一张真实图像）保留一张真实图像并替换其余图像（对保留的图像取平均），共使用另外 1,000 个 ASTERISM 项目。 † 表示 95% 区间不包含零的差异。

		(a) 隐藏在模型内部				(b) 来自其他条目的图像			
骨干网络	vs.	仅文本	一张图像	所有图像	全部 − 一个	没有真实图像	一张真实图像	所有图像	全部 − 一个
Qwen2.5-VL-7B [7]	base	+0.6	+0.6	+8.3†	+7.7†	−1.2	+1.6	+6.7†	+5.1†
	GRPO	+0.5	+0.3	+5.7†	+5.4†	−0.5	0.0	+5.9†	+5.9†
	DAPO	+0.2	−0.2	+4.0†	+4.3†	−1.5	+0.6	+3.5†	+2.9†

表 S18。**单图像基准**。 各骨干网络以及采用 CICA 训练的相同骨干网络在十项单图像基准测试上的得分。这些基准测试分为幻觉（HallusionBench [38], POPE [71]）、真实世界图像（RealWorldQA [145],MME-RealWorld-Lite [166], R-Bench-Dis [66]）、高分辨率图像（V*Bench [142], HR-Bench 4K 和 8K [134]）以及 OCR 与图表（OCRBench [79], AI2D [54]），并使用 VLMEvalKit [25] 及其默认答案提取方式进行评估。POPE 报告 F1 值。R-Bench-Dis 会对图像施加真实世界扰动。Avg 是十列的平均值，每对数值中较高的值以粗体显示。

	幻觉		真实世界			高分辨率			OCR/图表		
模型	Hallusion 基准	POPE	真实世界 QA	MME-RW Lite	R-Bench Dis	V* 基准	HR-Bench 4K	HR-Bench 8K	OCR AI2D 基准测试		Avg
Qwen2.5-VL-7B [7]	50.1	86.5	69.4	44.5	65.3	75.9	68.8	65.0	88.1	83.7	69.7
⌞ + CICA（我们的）	**51.0**	**89.2**	**71.1**	**48.7**	**68.5**	**77.0**	**72.6**	**69.5**	**88.3**	**83.9**	**72.0**
GLM-4.1V-9B-Thinking [43]	**63.1**	87.2	72.5	46.1	69.9	**78.0**	74.1	66.4	84.9	87.5	73.0
⌞ + CICA（我们的方法）	**63.1**	**88.4**	**75.2**	**47.6**	**72.7**	**78.0**	**75.0**	**66.5**	**85.7**	**87.8**	**74.0**
Qwen3.5-9B [105]	65.7	88.3	81.3	59.5	75.4	88.5	88.8	83.4	89.4	87.7	80.8
⌞ + CICA（我们的方法）	**67.3**	**88.5**	**81.7**	**60.8**	**76.6**	**89.5**	**90.0**	**85.7**	**90.1**	**90.5**	**82.1**

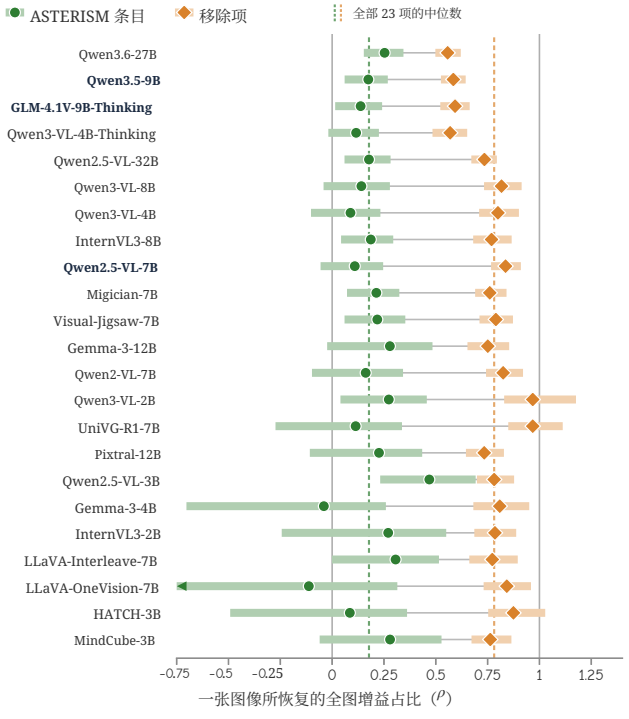


图 S18. **单图像恢复 ρ 所有 23 个多模态大语言模型的恢复情况。** 恢复比率 ρ 图 3 中各条目在灰色遮罩下的恢复比率。 3 针对第 2.2 节中的 23 个开源多模态大语言模型， 2.2 按 ASTERISM 得分排序（表 S9）。色带表示 95% 置信区间，虚线表示 23 个模型的中位数，粗体名称表示第 5.2 节训练的骨干模型。 5.2 箭头表示区间超出坐标轴范围。

表 S19. **MMIU**。 MMIU [88] Qwen2.5-VL-7B 和 Qwen3.5-9B、表 4 中五个强化学习基线，以及使用 CICA 训练的各骨干模型的准确率。评估采用 MMIU 官方提示和答案提取方式，以及补充材料 E.3 中的设置；得分按官方评估脚本的方式，对官方发布内容中的 60 个任务划分取平均。4 和每个使用 CICA 训练的骨干模型。评估采用 MMIU 官方提示及答案提取方式和补充材料 E.3 中的设置E.3，得分按官方评估脚本的方式，对官方发布内容中的 60 个任务划分取平均。

骨干网络	基础模型	GRPO	DAPO	APPO	VGPO	CFPO	CICA
Qwen2.5-VL-7B [7]	51.9	53.0	52.9	53.0	52.8	52.7	**54.4**
Qwen3.5-9B [105]	64.5	65.2	63.0	64.5	66.1	60.8	**68.5**

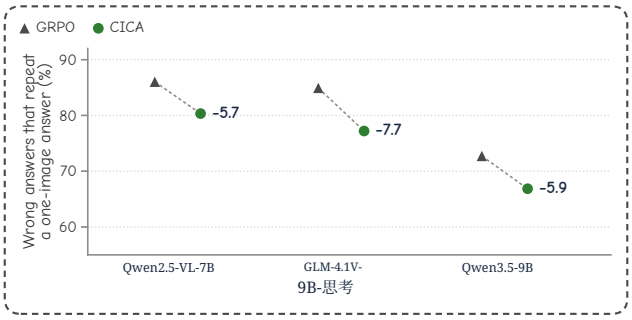


图 S19。**使用 CICA 训练后，重复全图像答案的错误答案少于使用 GRPO 训练后。**对于使用 GRPO 或 CICA 训练的每个骨干网络，错误的全图像答案中，与同一模型在仅显示一张图像且将其他图像显示为灰色时给出的答案相同的比例。条目是 1,000 个 ASTERISM 项目中符合以下条件的多项选择题：基础模型和两个训练模型在每种掩码条件下都给出了答案。数字表示 CICA 相对于 GRPO 的变化，单位为百分点。

表 S20。**来源基准的许可证。**每个来源在其数据集卡片或代码仓库中声明的许可证，以及 ASTERISM 为其发布的内容：要么是包含图像的条目（项目），要么是包含条目标识符以及用于根据公开发布内容组装条目的构建脚本的标识符（标识符）。 [a]数据集卡片与代码仓库或论文声明了不同的许可证，此处列出限制更严格的许可证。 [b]数据集卡片中列出了名称，但其许可证文件为空。图像保留其来源数据集的许可证。

来源	许可证	发布内容
MileBench [116]	CC BY-NC 4.0[a]	标识符
MMRB [19]	知识共享署名 4.0 国际许可协议	项目
VLM2-Bench [162]	CC BY-NC 4.0[a]	标识符
MMR-Life [70]	知识共享署名 4.0 国际许可协议	项目
ReMI [52]	知识共享署名 4.0[b]	项目
MuirBench [125]	知识共享署名 4.0 国际许可协议	项目
MRAG-Bench [44]	知识共享署名 4.0 国际许可协议	项目
M4Bench [155]	未注明	标识符
MV-MATH [129]	MIT	项目
BLINK [31]	Apache 2.0	项目
MMLongBench-Doc [87]	CC BY-NC 4.0	标识符
DUDE [60]	知识共享署名 4.0 国际许可协议	项目
MaRVL [76]	知识共享署名 4.0 国际许可协议（文本），CC BY-NC-SA 4.0（图像）	标识符
MMSI-Bench [151]	知识共享署名 4.0 国际许可协议	项目
Q-Bench2 [167]	S-Lab 许可证 1.0（非商业用途）[a]	标识符
MIRB [168]	MIT	项目
DiCoBench [68]	MIT	项目
EMMA [42]	未说明	标识符
Mantis-Eval [49]	Apache 2.0	项目

表 S21. **ASTERISM 统计数据**。条目数、来源基准（Src）、每个条目的图像平均数量和最大数量、答案为选项字母（Opt）、是/否或真/假（Y/N）、数字（Num）或短文本（Str）的条目百分比，以及标签、答案格式、问题或选项与源发布内容不同（选项排序除外）的源条目（Fix）。

类别	条目数	来源	图像		选项	是/否	数字	短文本	修订
			平均值	最大值					
聚合	1,055	13	3.4	15	56	0	37	7	19
比较/变化	2,262	14	2.9	35	99	1	0	0	106
推理/验证	4,793	13	3.0	35	35	36	22	7	86
空间对应	4,618	13	3.7	35	96	0	0	4	91
时间/动作	3,463	9	12.6	109	99	0	0	1	172
文本/网络	154	7	3.4	10	100	0	0	0	8
检索	2,351	8	19.1	64	39	2	17	42	87
ASTERISM	18,696	19	7.0	109	72	10	10	9	569

表 S22. **将 β 降至 0.5 或将 τ 加倍，可使 CICA 的 ASTERISM 得分高于表4 中 Qwen2.5-VL-7B 上的所有强化学习基线**。各训练目标的 ASTERISM 得分。这两种调整都会使权重更接近 1，得分也更接近 GRPO（$\beta = 0$）。

训练目标	ASTERISM
GRPO（$\beta = 0$）	41.7
CICA，$\beta = 0.5$	44.2
CICA，$\tau = 1.0$	44.5
CICA（$\beta = 0.95$，$\tau = 0.5$）	46.7