解除细粒度详细字幕评价的阻碍：一个解释性自动评分器和批评与修订流程

Brian Gordon*[1,2], **Yonatan Bitton***[2], **Andreea Marzoca**[2], **Yasumasa Onoe**[2],
Xiao Wang[2], **Daniel Cohen-Or**[1], **Idan Szpektor**[2],
[1]Tel Aviv University, [2]Google Research
https://google.github.io/unblocking-detail-caption

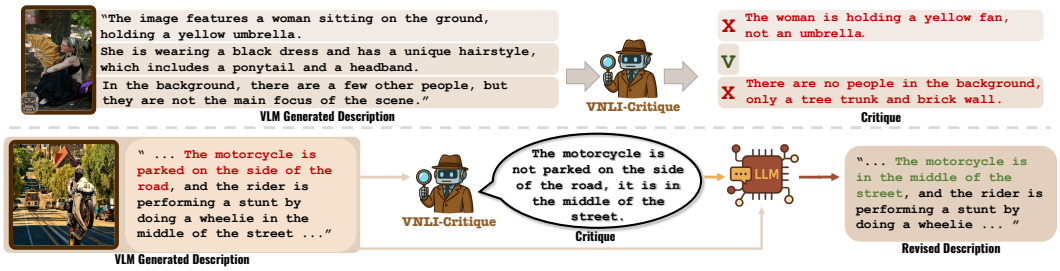

Figure 1: VNLI-Critique 的应用：作为评论者以及在评论和修正流程中操作。（顶部）作为评论者，VNLI-Critique 评估 VLM 标注中句子的事实性，并生成错误解释。（底部）在流程中，它对不正确句子的评论指导 LLM 进行修正，展示了详细标注的自动化评估和修正。

自动描述性图像字幕生成是一个显著的视觉-语言研究领域，已经从简短的亮点发展到详细的段落长度的描述，这要归功于强大的大型视觉-语言模型。评估这些复杂的字幕仍然具有挑战性；现有的度量标准通常适用于短文本，常常忽略细致入微的细节，并且通常将句子单独评估，缺乏解决歧义和共指所需的重要段落上下文。虽然一些研究关注完整段落，但在细粒度的句子级别评估上仍然困难。

评估 VLM 生成的文本真实性已导致专门的基准测试。然而，许多现有的错误/幻觉检测基准针对的是短句或 QA 任务，无法充分解决段落长度描述的问题。尽管诸如 M-HalDetect [13] 和 CHOCOLATE [17] 的数据集提供了有价值的句子级真实性注释，但它们可能缺乏来自长形式 VLM 输出的完整错误多样性或对不准确性的连贯、详细文本解释——这对于开发有效的、细粒度的自动化评估至关重要。响应级别的评估（例如，CAPTURE [8]）对整个段落进行评分，但缺乏句子级的细粒度。因此，迫切需要一个具有全面、上下文感知的句子级真实性注释的基准，包括对各种 VLM 生成段落中错误的解释理由。

为了解决这个问题，我们引入了 DOCCI-Critique，这是一种用于精细评估详细图像描述的新基准。它包含了 1,400 个段落长度的描述（14 个 VLMs，100 张图像），其核心价值在于 10,216 条句子级别的人工注释。每个句子的真实性由五位注释员进行评判，并对每个识别出的错误提供详细的文字理由。这种多视角的注释为深入分析 VLM 提供了一个新资源，提供了经过多方验证的句子级判定和长描述错误解释，与现有数据集不同。

基于 DOCCI-Critique，我们开发了 VNLI-Critique，它是一个用于自动句子级别事实性分类（使用段落上下文）和解释性批判生成的模型。这种双重能力使一种新颖的批判与修订流程成为可能（图 1）：VNLI-Critique 对不正确的 VLM 生成的句子进行评估并生成批判，引导 LLM 对其进行修订。VNLI-Critique 和此流程的效用通过关键结果体现出来：（1）VNLI-Critique 在外部基准 M-HalDetect (Macro-F1 0.76) 上取得了最先进的性能，并在 CHOCOLATE [17] 上取得了有竞争力的结果，展示了很强的泛化能力。（2）在我们的基准测试中，VNLI-Critique 驱动的 DOCCI-Critique AutoRater 显示出 VLM 排名与人类判断之间的高度相关性（Table 3

* Equal contribution

Preprint. Under review.

）。（3）批判与修订流程显著提升了不正确句子的事实性（例如，DetailCaps-4870 [8] 上提升了 46 %，PixelProse [30] 上提升了 51 %），这一点得到了人类评估的确认。总体而言，这些贡献提供了一个重要的基准和强大的方法，使得更精确的细粒度评估成为可能，并显著增强了 VLM 对详细图像理解的事实准确性。我们的工作与视觉语言模型（VLM）在详细描述、图像描述数据集的开发以及评估描述质量的方法（尤其是事实准确性和细粒度细节）方面的进展有交集。

最近的 VLMs [1, 3, 6, 7, 19, 20, 23, 25, 31, 32, 38, 43, 46] 在多模态任务中取得了显著的 SOTA 性能。通常，它们将视觉编码器 [9, 27, 34, 45] 与 LLM [4, 5, 33, 40] 结合在一起，常使用连接模块将视觉特征与文本标记相连接。训练通常涉及预训练视觉编码器，然后对 LLM 进行微调，目标如掩蔽语言建模或图文匹配。尽管已经探索了端到端的训练，但这种两阶段方法很常见，平衡了数据集规模、计算资源和评估。

图像描述数据集 图像理解和描述数据集对于支持当前图像描述技术的进步至关重要。早期数据集提供了带有简短句子的正面图像-文本对，主要关注主要物体和场景，例如 COCO [21] 、Flickr8k [15] 和 Flickr30k [44] 。最近的图像描述数据集则提供了带有更长、更详细描述的图像-文本对。例如，Densely Captioned Images (DCI) 数据集 [35] 引入了长的、与遮罩对齐的描述，专门用来评估 VLM 对不同图像区域的理解。PixelProse [30] 提供了一个大规模数据集，包含使用 Gemini-1.0-Pro-Vision [32] 合成生成的 1690 万个描述；然而，这些描述的正确性并不能得到保证。IIW [10] 数据集利用 VLM 生成初始描述，然后采用人工参与流程以确保高质量的正面图像-文本对。M-HalDetect [13] 和 CHOCOLATE [17] 分别使用各种 VLM 来描述图像和图表，并采用句子级的人为注释来评估生成描述的正确性。DetailCaps [8] 利用 GPT-4 [25] 为现有数据集的合成生成描述分配质量分数（范围从 0 到 5）。DOCCI 数据集 [24] 是研究的一个特别强有力选择，这要归功于其多样化现实场景的高分辨率图像，以及详细的、完全由人工撰写的描述，提供了一个有价值的资源，用于训练和评估 VLMs。它提供了多样化现实场景的高分辨率图像，每个图像都配有精心制作的、冗长的人为注释描述。

详细说明的评估 传统的度量方法（例如，BLEU [26] ，METEOR [2] ，CIDEr [36] ）通过 n-gram 重叠将生成的标题与参考进行比较，常常忽略段落长度文本中至关重要的语义细微差别。基于嵌入的方法，如 CLIPScore [14] 和 SIGLip [45] 提供了更好的语义评估，但通常将句子孤立评估，缺乏段落上下文需要以解决详细描述中的歧义或共同引用。基于 QA 的方法（例如，VQAScore [22] ，TIFA [16] ，VQ^2 [41] ，GECKO [39] ）通过问答评估理解，但在面对长篇叙述时面临可扩展性挑战。最近的工作如 Mismatch Quest [12] 专注于通过自动生成解释来为错位提供详细的文本和视觉反馈，主要在文本到图像领域，通常针对于较短的、直接的文本输入。虽然诸如 Dong et al. [8] 这类响应级别的评估对整段文字进行评分，但它们并不提供描述性标题的句子级事实详情。我们的工作解决了对详尽的段落长度图像标题进行细致的、上下文感知的句子级事实性评估的需求，并具有丰富的、经过人为验证的解释性反馈。

1 DOCCI-Critique 基准：构建、注释与分析

DOCCI-Critique 是一个新颖的基准，用于对段落级图像描述进行细粒度事实性评估。其主要目的有两方面：（1）提供一个强大的平台评估最先进（SOTA）字幕模型的描述能力和事实准确性，以及（2）作为自动图像理解和事实核查系统（自动评分者）的一个具有挑战性的测试平台。

构建过程始于从 DOCCI 数据集的 ‘qual-dev’ 分割中选取的 100 张多样化、高分辨率图像 [24] 。对于每张图像，14 个 SOTA 大型视觉语言模型（表 2 ）生成了详细的段落长描述，总共产生了 1400 个模型生成的描述。这个语料库有意捕捉了从简明、事实性陈述到较冗长的叙述（可能引入细微的不一致性，例如对象误识别）的广泛风格变化、细节层次和事实准确性。

DOCCI-Critique 的核心是其丰富的人类注释层。根据 Steiner et al. [31] （参见附录中的注释模板），五名人工注释者独立评价了 1400 个描述中的每一个句子在图像中的事实正确性，并赋予标签：“蕴涵”（事实支持）、“中性”（不可验证/矛盾）、“矛盾”（事实矛盾），或“无评估内容”（例如，填充）。如果大多数人将一个句子标记为“中性”或“矛盾”，则该句子被分类为“非蕴涵”。重要的是，注释者为每一个非蕴涵的判断提供了详细的文字理由。因此，一个非蕴涵的句子通常有多个理由，捕捉对不准确的多元观点，并提供对 VLM 错误类型（例如，物体识别错误、属性/空间错误、幻觉）的见解。

Table 1: 来自 DOCCI-Critique 基准的示例，详细说明用于细粒度事实性评估的句子级注释，包括评估者的判断和解释理由。

Image			
Description Sentence	"... Behind the car, there is a large mural or poster on the wall ..."	"... The mural features a Formula 1 racing car, also red, with the number 16 prominently displayed on the side. ..."	"... The background of the mural includes a racing track with the colors of the French flag (blue, white, and red) and a checkered flag, indicating a racing theme ..."
Does the sentence include a claim about the image? (Answers from 5 raters)	✅, ✅, ✅, ✅, ✅	✅, ✅, ✅, ✅, ✅	✅, ✅, ✅, ✅, ✅
Is the sentence factual? (Answers from 5 raters)	✅, ✅, ✅, ✅, ✅	✅, ✅, ❌, ❌, ✅	❌, ✅, ❌, ❌, ✅
Rationales	-	• 数字 16 不在赛车的侧面，而是在赛车的前面。 • 壁画确实有一辆红色的一级方程式赛车，但号码 16 是画在前面而不是侧面	• 图像中没有可见的方格旗。 • 海报/壁画中没有方格旗。 • 背景壁画确实包含蓝色、白色和红色，但没有方格旗

Table 1 展示了这一结构，显示了每个句子的注释：五个独立的事实性判断（✅ 对于推断，❌ 对于矛盾/中立）和图像内容依赖性。对于非推断的投票，收集到的文字推理解释了具体错误，例如汽车壁画的号码位置或不存在的格子旗。一个错误句子的多个理由反映了不同标注者的观点。

这一综合注释产生了 10,216 个句子级别的判断。该数据集拥有多样的 VLM 输出、精细的多数投票真实性标签和每个错误的多个丰富解释依据，是一项无价的资源。它不仅能在表面相似性上进行严格的 VLM 评估，还可以更深入地探查模型理解和描述的忠实性。

表格 2 详细介绍了来自 DOCCI-Critique 的每个模型的统计数据，包括描述和句子长度、事实准确性和词汇多样性。这些内部统计数据揭示了 VLM 在数量和质量上的行为差异。例如，像 GPT-4o 这样的高准确性模型与 Gemini 模型形成对比，后者生成更多句子且准确性相近，这表明存在详尽性/细节与错误风险之间的权衡。DOCCI-Critique 中的段落平均长度为 752.7 个字符。这明显长于用于事实错误分析的其他当代数据集，比如 M-HalDetect（平均 456.2 个字符）、CHOCOLATE（577.6）和 DetailCaps-4870（612.9）。这种对更长、更复杂描述的强调，结合常见错误类型的模式（从推理中可察觉，尽管在表格 1 中未直接显示），突显出 DOCCI-Critique 对于 VLM 生成策略和视觉真实性细致对比研究的实用性，强调其在评估详细文本生成中多样化 VLM 行为中的作用。

2 VNLI-Critique：开发和评估

2.1 VNLI-Critique 模型开发

我们通过微调 10B 参数的 PaliGemma-2 架构 [31]（详见附录），开发了 VNLI-Critique，用于自动化句子级准确性评估和批评生成。这需要一个专门的训练数据集，该数据集由生成的 VLM 字幕组成，与 DOCCI-Critique 不同，并对其准确性和错误评价进行了标注。为了创建这个多样化的训练数据，我们首先使用超过 70 个 PaliGemma-2 变体（在 DOCCI 训练数据 [24] 上以不同配置微调）生成段落长度的字幕，以捕获广泛的生成风格和潜在错误。然后，按照 Section 1 中的协议（标签多数投票；将最长的理由作为未蕴涵句子的批评目标），对这些合成字幕进行人工标注。

VNLI-Critique 在这组精心整理的数据上被微调用于两个任务，通过具体的提示结合段落上下文（<PREFIX> Claim-Prefix </PREFIX>），对于潜在模糊的独立句子的准确评估至关重要。对于事实性分类，提示是："给定图像和提示前缀 <PREFIX> Claim-Prefix </PREFIX>，以下文本是否与图像一致：<TARGET> Target-Claim </TARGET> ?"，需要一个"是"/"否"的预测。对于批判生成，提示是："给定图像和提示前缀 <PREFIX> Claim-Prefix </PREFIX>，文

Table 2: DOCCI-Critique 统计，详细描述每个 LLM 所生成图像描述的段落级别指标和词汇多样性（独特的 2-grams）。

	Description Length avg.	# Sentences Avg	Sentence Length Avg	% Correct Sentence in Description	Uni. 2-gram
MiniGPT-4 [46]	483.5	5.6	84.8	45.6	4,695
mPLUG-Owl2-7B [43]	458.8	4.4	102.1	52.7	4,038
LLaVa-1.5-7B [23]	395.4	4.2	91.5	60.0	3,081
InstructBLIP [6]	509.8	4.0	195.4	61.3	3,260
PALI-5B [3]	1098.9	10.9	69.5	68.0	1,881
VILA [20]	870.7	8.6	100.4	78.1	6,841
mPLUG-Owl3-7B [42]	118.0	2.0	65.2	80.4	700
LLaVA-Onevision-7B [19]	672.0	6.4	107.7	81.8	5,878
Molmo-7B-D [7]	747.6	6.6	111.9	82.7	6,788
LLaVA-Onevision-7B-Chat [19]	1091.6	9.5	113.1	85.7	8,550
Qwen2-VL-7B-Instruct [37]	1022.6	9.8	102.9	87.6	8,250
Gemini-1.5-Pro [32]	1326.9	12.0	109.3	95.1	11,705
Gemini-1.5-Flash [32]	1199.0	11.8	100.0	96.1	10,186
GPT-4o [2024-08-06] [25]	583.5	6.2	94.2	97.1	6,160
TOTAL	752.7	7.3	103.8	76.5	40,444

本 <TARGET> Target-Claim </TARGET> 被认为不准确。请解释导致其不准确的对齐和事实不符之处。”。这种双任务战略使 VNLI-Critique 能够识别差异并说明其原因。

2.2 事实性分类：基准测试和泛化结果

本节详细介绍了 VNLI-Critique 在事实性分类任务中的表现，展示了作为自动化基准测试工具在 DOCCI-Critique 上的关键结果以及在测试各种外部数据集时的泛化能力。

DOCCI-Critique AutoRater：自动化 VLM 基准测试结果。 的分类能力的一个主要应用是作为一个自动评分器，用于建立一个自动化的排行榜，根据对来自 DOCCI-Critique 基准的图像描述的事实准确性对视觉-语言模型（VLMs）进行排名。其目标是为这一任务提供一个可扩展且可靠的替代方案，以替代大量的人力评估。为了评估其可行性，我们评估了 VNLI-Critique 以及其他基于 VLM 的方法，作为潜在的自动排名工具。我们比较了它们的自动评估与人类判断在三个不同的事实性标准上的相关性：(1) 响应级别的正确性（整个生成的段落是否在事实上准确），(2) 整体正确句子的百分比（针对一个模型所有生成描述的总正确句子数），以及 (3) 每个描述的平均正确句子百分比。附录中提供了展示每一标准下由人类评估和自动化方法确定的 VLM 排名的详细排行榜。使用 Spearman 的 ρ (Sp ρ) 和 Kendall's τ (Kd τ) 的相关性结果在表 3 中呈现。VNLI-Critique 作为一个自动评分器表现出卓越的性能，与人类排名在响应级正确性（Sp $\rho = 0.981$）和整体正确句子的百分比（Sp $\rho = 0.979$）上达到最高的 Spearman 相关性，并且在每个描述的正确句子平均百分比上有非常高的相关性（Sp $\rho = 0.968$）。其在不同评价粒度上的强劲表现，与人类评估显著一致，验证了其作为一个可靠工具在 DOCCI-Critique 数据集上自动基准测试 VLM 事实性的有效性。

Table 3: 评估自动化方法作为自动评级器。模型排名与人类对 DOCCI-Critique 上 VLM 事实性判断的相关性（Spearman 的 ρ，Kendall 的 τ），涉及三个准确性指标。加粗表示最佳分数，underline 表示第二优。

	% Response Correct		% Sentences Overall		% Sentences per Description	
Ranking Method (Model)	Sp ρ	Kd τ	Sp ρ	Kd τ	Sp ρ	Kd τ
Emu3-Chat	-0.192	-0.167	0.059	0.011	0.007	-0.055
InstructBLIP [Vicuna-7B]	-0.059	-0.046	0.367	0.187	0.354	0.143
Qwen2.5-VL-7B-Instruct	0.290	0.249	0.692	0.516	0.697	0.495
Janus-Pro-7B	0.294	0.211	0.521	0.341	0.578	0.407
mPLUG-Owl3-7B	0.734	0.573	0.741	0.582	0.798	0.648
LLaVa-OneVision[Qwen2-7B]	0.889	0.760	0.855	0.758	0.851	0.736
GPT-4o	0.920	0.818	0.975	0.911	0.987	0.934
Gemini-2.0-Flash	0.972	0.884	0.976	0.911	0.956	0.890
VNLI-Critique (Ours)	0.981	0.928	0.979	0.912	0.968	0.906

Table 4: 评估 VNLI-Critique 的事实性分类：与基线在分布内（DOCCI-CRITIQUE ）和外部（M-HalDetect、CHOCOLATE）数据集上的比较。主要结果包括在 M-HalDetect 上的 SOTA 表现以及对 CHOCOLATE 的强泛化能力。

Model	DOCCI-Critique		M-HalDetect		CHOCOLATE	
	ROC-AUC	Macro-F1	ROC-AUC	Macro-F1	ROC-AUC	Macro-F1
CLIPScore	0.48	-	0.59	-	0.56	-
VQAScore [CLIP-FlanT5]	0.73	-	0.79	-	0.71	-
VQAScore [GPT-4o]	0.88	-	0.85	-	0.84	-
SigLIP	0.50	-	0.63	-	0.56	-
TIFA	0.61	-	0.70	-	0.57	-
PaliGemma2 [9B-448res]	0.51	0.23	0.61	0.39	0.53	0.00
Qwen2.5-VL-7B-Instruct	0.65	0.36	0.81	0.75	0.81	0.74
InstructBLIP [Vicuna-7B]	0.50	0.45	0.45	0.40	0.53	0.37
Emu3-Chat	0.51	0.50	0.52	0.42	0.50	0.37
Janus-Pro-7B	0.67	0.58	0.72	0.59	0.65	0.47
LLaVa-OneVision[Qwen2-7B]	0.76	0.58	0.82	0.60	0.75	0.44
mPLUG-Owl3-7B	0.73	0.65	0.76	0.68	0.68	0.54
Gemini-2.0-Flash	0.73	0.74	0.74	0.74	0.81	0.79
GPT-4o	-	0.74	-	0.69	-	0.70
VNLI-Critique (Ours)	0.93	0.83	0.86	0.76	0.73	0.68

为了评估 VNLI-Critique 在我们特定基准之外的能力，我们在两个已建立的外部数据集上评估其性能：M-HalDetect [13]，这是一个用于检测不同图像描述中的幻觉现象的基准，以及 CHOCOLATE [17]，专注于图表和曲线图的描述。我们将 VNLI-Critique 与各种基线进行比较，包括其他基于 VLM 的分类器和嵌入相似性的方法，使用两个标准的元评估指标：ROC-AUC 和 Macro-F1。 **??** 提供了这些句子级分类比较的定性例子。

表 4 中报告的性能指标是基于每个模型的输出类型生成的。对于通过特定输出标记进行分类的模型（例如，“是” 和 “否”）——这包括我们的 VNLI-Critique 和其他基于 VLM 的分类器（分数是通过 5 样本策略得出的）——指标反映了信心和预测。对于所有这样的 VLM 分类器，ROC-AUC 计算的蕴涵分数是通过对与正负分类输出相关的置信度分数应用 softmax 函数获得的（生成归一化的正概率）。用于 Macro-F1 的二元分类（“准确” 或 “不准确”）是通过选择具有更高置信度分数的标签来确定的。相比之下，对于像 CLIPScore [14] 、SigLIP [45] 和 TIFA [16] 这样的方法，它们输出数值相似性分数，只有 ROC-AUC 被报告，因为它直接适用于这种分数而不需要任意阈值。

如表 4 所示，VNLI-Critique 在 M-HalDetect 上实现了最先进的（SOTA）性能。此外，它在 CHOCOLATE 数据集上的高度竞争性能展示了显著的适应性和强大的推理能力，即使在没有对这些视觉数据进行特定训练的情况下，评估图表描述时也是如此。这些跨不同基准的强大结果强调了我们微调模型在事实验证任务中的普遍实用性。

除了对句子正确性进行分类外，VNLI-Critique 的一个关键能力是生成文本批评，以解释为什么一个句子在事实层面上是不准确的。为了评估这些生成解释的质量和正确性，我们进行了一项专门的人类评估研究。评估过程如下：首先，我们从 DOCCI-Critique 基准和 M-HalDetect 数据集中抽取了一组先前由人类标注者识别为事实错误的句子。对于每个抽取的错误句子，我们提示 VNLI-Critique 以及几种具有竞争力的 VLM（列在 Table 5 中）生成一个解释，详细说明具体的事实不准确或不对齐，使用 Section 2.1 中描述的批评生成提示格式。然后，人类标注者会被展示评估实例，每个实例包含：(1) 原始图像，(2) 具体的事实错误句子，以及 (3) 被评估模型生成的批评。标注者的任务是判断批评本身的质量——具体而言，是看它是否准确且相关地识别了句子中存在的事实错误，并与图像中的视觉证据进行比较。Table 5 显示了此人类评估的结果，报告了生成的评论中被注释者认为正确和相关的百分比。结果表明，VNLI-Critique 在生成有用的评论方面非常有效。在针对 DOCCI-Critique 句子的评论中，VNLI-Critique 取得了最高分（73.39 %），略微超过了 GPT-4o（73.1 %）。在针对 M-HalDetect 句子的评论中，VNLI-Critique 表现非常强劲（79.33 %），仅次于 Gemini-2.0-Flash（79.89 %），但领先于 GPT-4o（78.77 %）。值得注意的是，VNLI-Critique 显著超越了几款其他能力较强的 VLMs，如 Janus-Pro-7B、Qwen-2.5-VL-Instruct 和 LLaVA-OV 在这两组数据集上。这表明 VNLI-Critique 能够一致地生成高质量、准确的事实错误解释，这对于提供可解释的字幕质量反馈和启用后续纠正任务是关键能力。

Table 5: 对批判质量的人类评估。从 DOCCI-Critique 和 M-HalDetect 中采样的错误句子中，生成的解释被判断为正确和相关的百分比。

	DOCCI-Critique	M-HalDetect
LLaVA-OV	35.96	48.04
Qwen-2.5-VL-Instruct	45.03	58.1
Janus-Pro-7B	44.15	62.57
Gemini-2.0-Flash	64.91	79.89
GPT-4o	<u>73.1</u>	78.77
VNLI-Critique (Ours)	73.39	<u>79.33</u>

3 批判与修正

许多视觉语言任务，从图像描述到文本生成图像，严重依赖于大型图文对数据集，通常利用 VLM 生成的描述进行训练或作为其数据的一部分。例如，大型合成描述数据集如 PixelProse [30] 用于训练描述模型，而将图像与描述性文本配对的数据集是训练文本到图像合成模型的基础（例如，利用如 LAION [29] 之类的数据集）。然而，这些自动生成或网络抓取的描述在事实准确性和视觉匹配上可能有所不同，可能会在后续模型训练中引入噪声或不准确。因此，提高这些数据集的质量和事实匹配性对于推进这些领域至关重要。利用 VNLI-Critique 的批评生成能力，我们引入并评估了一种新颖的批评-修订流程。该流程的设计不仅旨在修正图像描述中的个别事实不准确性，还提供了一种途径来提升图文训练数据集的整体质量，从而可能提高以其为训练基础的模型的性能。本节首先概述了该流程的方法学 (Section 3.1)。然后，我们评估其在修正大规模数据集中合成生成的描述中的适用性 (**??**)。

3.1 流水线方法论

批评与修订流程，如 Figure 1 所示，包括两个主要步骤。首先，在批评步骤中，VNLI-Critique 使用其分类功能分析给定标题的每个句子；被识别为事实错误的句子会触发文本评论的生成，解释基于图像内容的具体错误。随后，在修订步骤中，来自 VNLI-Critique 的原始不准确句子及其相应评论用于指导一个独立的大型语言模型（LLM）修正不准确的描述。在我们的实验中，我们使用了 Gemini-2.0-Flash [1] 作为修订 LLM。此修订 LLM 会被提示重写原始句子，特别是针对评论中强调的事实错误，同时努力保持相关信息并维护风格上的一致性。完整的批评与修订循环——由 VNLI-Critique 针对所有句子的事实性分类，然后为标记为不准确的句子进行评论指导的修订——生成一个在事实与图像更一致的修订标题。

为了展示我们提出的 Critic-and-Revise 流程在下游的效用，我们将其应用于两个以详细但潜在未经验证描述而闻名的大规模数据集的字幕: PixelProse [30] ，具有约 16.9M 的合成字幕对，以及 DetailCaps-4870 [8] ，一个包含 4870 张图像的子集，每张图像都伴随着三条详细的合成字幕。我们进行了一项人类研究来评估这一流程的有效性：在 VNLI-Critique 识别并批评不准确的句子后，修订 LLM 对其进行纠正，人类评估者评估了原标记句子的事实正确性（用于批评精确度）以及修订后句子的正确性（用于流程的有效性）。

结果总结在 Table 6 中，显示了显著的改进。对于 DetailCaps-4870，虽然 VNLI-Critique 的初始标记显示了 15 % 的误报率（被人类认为正确的原始句子），但管道成功地纠正了大量真正不准确的句子，人类评委确认 61 % 的修正句子在事实上是准确的。这代表了在最初被评论者认为错误的句子集中 46 % 的准确度提升。VNLI-Critique 自己的重新评估将这些修正后的句子中 64 % 的句子分类为准确，显示出强大的自我一致性。在 PixelProse 中观察到类似的积极趋势，人类评委发现 75 % 的修正句子是准确的（提升了 51 %），展示了管道在大规模增强详细图像标题事实准确性的能力。附录中提供了说明批评和修订过程的定性例子，包括原始错误句子、来自 VNLI-Critique 的批评以及经过 LLM 修订的句子。

[1] 通过 Vertex AI API 访问：`https://cloud.google.com/vertex-ai`

Table 6: 批评与修订流程的事实性：原始声明与修订声明的人类和 VNLI-Critique 评判。Δ = 修订后准确性增加。该流程显著提高了声明的准确性（人类确认 DetailCaps 的固定声明准确性为 61 %，PixelProse 为 75 %，起始值较低）。VNLI-Critique 的评判一致，显示出高度自我一致性。

	DetailCaps-4870			PixelProse		
Judge Type	Original	Fixed	Δ	Original	Fixed	Δ
Human Judge	15 %	61 %	+46 %	24 %	75 %	+51 %
VNLI-Critique as Judge	0 %	64 %	+64 %	0	61 %	+61 %

4 限制和未来工作

尽管 DOCCI-Critique 基准测试包含了由 VLM 生成的 1,400 个字幕和超过 10,000 个句子判断的丰富注释，它是从一个由 100 个独特图像构成的基础集构建的。虽然这些图像提供了多样性且字幕的变化十分广泛，增加基础图像的数量可以进一步提升基准测试的统计能力和覆盖范围。然而，我们的实验展示了强大的泛化能力。具体来说，VNLI-Critique ，当在 DOCCI-Critique 上训练进行事实性分类时，在像 M-HalDetect 和 CHOCOLATE（Section 2.2 ）这样的外部未见过的声明验证数据集上表现良好。此外，我们的批判与修正流程利用来自 VNLI-Critique 的评论，能够有效地在完全不同的数据集如 DetailCaps-4870 和 PixelProse（**??** ）上修正字幕。这种跨不同任务和数据集的泛化集体证据表明当前 DOCCI-Critique 的规模对于开发稳健且可转移的评价模型和修正方法是有效的。

此外，尽管 VNLI-Critique 在多个环境中取得了强劲的成绩，其表现并不完美。我们的流程评估（Section 3 ）表明，其真实性分类可能会导致误判和漏判（例如，在 DetailCaps-4870 上的虚假阳性率为 15 %）。尽管其生成的评论质量通常很高，如在人类评估中 Table 5 所示，但也可能表现出不一致性。未来的工作可以通过进一步利用我们丰富的注释来提高 VNLI-Critique 的表现。例如，可以研究将不同标注者提供的多个理由合并为一个更全面和简洁的解释的方法，而不是仅仅使用最长的理由作为评论生成的训练目标。此外，我们的标注协议记录了一个句子的真实性是否依赖于图像内容或依赖于世界知识（例如，区分需要图像的“猫在垫子上”和不需要图像的“猫是哺乳动物”）。目前未使用的标签可以实现两阶段验证过程：首先分类是否需要图像关联，然后相应地应用视觉事实检查器（VNLI-Critique）或基于知识的验证器，从而可能提高整体准确性和效率。

关于批评与修订流程，其当前设计包含两个不同的步骤：VNLI-Critique 进行批评生成，随后使用一个单独的 LLM 进行修订。虽然有效，但这种方法与一个可假设的端到端模型可能直接输出一个已修正的句子的做法形成对比。然而，我们认为两步方法在可解释性方面提供了显著的优势。生成明确的批评可以清楚地理解为何某个句子被标记以及正在解决的具体错误是什么。对错误类型和来源的这种洞察对分析和改进基础的描述模型是有价值的，这种好处可能在直接的黑箱修正方法中丧失。因此，虽然未来的工作可能探索直接修订模型，但中间批评的解释性仍然是我们流程的一个关键强项。

解决这些限制并探索所建议的方法，以充分利用数据集注释，提供了激动人心的未来研究方向，在细致图像理解的稳健和可解释性评估方面。

5 结论

本研究解决了评估和改进由 VLM 生成的详细段落长度的图像说明的事实准确性这一关键挑战。我们引入了 DOCCI-Critique，这是一个新颖的基准，具有 1,400 个 VLM 说明和超过 10,216 个面向事实性的人类注释，包含错误的解释性理由，为细粒度 VLM 评估提供了重要资源。在此基础上，我们开发了 VNLI-Critique ，这是一种擅长自动事实性分类和批评生成的模型。VNLI-Critique 在像 M-HalDetect 这样的外部数据集上展示了强大的泛化能力，并且其在 DOCCI-Critique 中的使用显示出与人类判断的高度相关性（0.98 Spearman）。此外，我们提出了一种新颖的批评和修正管道，在该管道中，VNLI-Critique 的批评指导 LLM 自动纠正事实错误，从而显著提高说明的准确性，这已通过人类评估得到证实。总的来说，DOCCI-Critique、VNLI-Critique 和批评与修正管道为推进 VLM 生成更详细、流利和事实可靠的图像描述提供了重要工具和方法。未来的工作，如 Section 4 所述，将探索扩展基准并进一步增强该管道的能力。

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[2] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, 2005. Association for Computational Linguistics.

[3] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*, 2023.

[4] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023.

[5] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tai, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25(1), 2024.

[6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2023. Curran Associates Inc.

[7] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, Yen-Sung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Jen Dumas, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross B. Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *CoRR*, abs/2409.17146, 2024.

[8] Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. Benchmarking and improving detail image caption. *arXiv preprint arXiv:2405.19092*, 2024.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[10] Roopal Garg, Andrea Burns, Burcu Karagol Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Michael Baldridge, and Radu Soricut. ImageInWords: Unlocking hyper-detailed image descriptions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 93–127, Miami, Florida, USA, 2024. Association for Computational Linguistics.

[11] Google Cloud. Introduction to Cloud TPU. `https://cloud.google.com/tpu/docs/intro-to-tpu`, 20xx. Accessed: 2024-07-04.

[12] Brian Gordon, Yonatan Bitton, Yonatan Shafir, Roopal Garg, Xi Chen, Dani Lischinski, Daniel Cohen-Or, and Idan Szpektor. Mismatch quest: Visual and textual feedback for image-text misalignment. In *Computer Vision –ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LVII*, page 310–328, Berlin, Heidelberg, 2024. Springer-Verlag.

[13] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2024.

[14] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.

[15] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: data, models and evaluation metrics. *J. Artif. Int. Res.*, 47(1):853–899, 2013.

[16] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering . In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20349–20360, Los Alamitos, CA, USA, 2023. IEEE Computer Society.

[17] Kung-Hsiang Huang, Mingyang Zhou, Hou Pong Chan, Yi Fung, Zhenhailong Wang, Lingyu Zhang, Shih-Fu Chang, and Heng Ji. Do LVLMs understand charts? analyzing and correcting factual errors in chart captioning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 730–749, Bangkok, Thailand, 2024. Association for Computational Linguistics.

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

[19] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[20] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26689–26699, 2024.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[22] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation withăimage-to-text generation. In *Computer Vision – ECCV 2024*, pages 366–384, Cham, 2025. Springer Nature Switzerland.

[23] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306, 2024.

[24] Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, Su Wang, and Jason Baldridge. Docci: Descriptions of connected and contrasting images. In *Computer Vision –ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LX*, page 291–309, Berlin, Heidelberg, 2024. Springer-Verlag.

[25] OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mdry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern,

Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024.

[26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, page 311–318, USA, 2002. Association for Computational Linguistics.

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[28] Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Par-

rish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozinska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucinska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjösund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, and Lilly McNealus. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118, 2024.

[29] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: an open large-scale dataset for training next generation image-text models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2022. Curran Associates Inc.

[30] Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. From pixels to prose: A large dataset of dense image captions, 2024.

[31] Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*, 2024.

[32] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe

Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakievi, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang,

Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce

Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kpa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024.

[33] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[34] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025.

[35] Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26700–26709, 2024.

[36] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575. IEEE Computer Society, 2015.

[37] 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, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[38] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

[39] Olivia Wiles, Chuhan Zhang, Isabela Albuquerque, Ivana Kajic, Su Wang, Emanuele Bugliarello, Yasumasa Onoe, Pinelopi Papalampidi, Ira Ktena, Christopher Knutsen, Cyrus Rashtchian, Anant Nawalgaria, Jordi Pont-Tuset, and Aida Nematzadeh. Revisiting text-to-image evaluation with gecko: on metrics, prompts, and human rating. In *The Thirteenth International Conference on Learning Representations*, 2025.

[40] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024.

[41] Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. What you see is what you read? improving text-image alignment evaluation. In *Advances in Neural Information Processing Systems*, pages 1601–1619. Curran Associates, Inc., 2023.

[42] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models, 2024.

[43] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13040–13051, 2024.

[44] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[45] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986, 2023.

[46] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A DOCCI-Critique AutoRater 排行榜

本附录展示了完整的排行榜，详细说明各种视觉语言模型（VLMs）作为自动排序器（AutoRaters）在 DOCCI-Critique 基准测试上的表现。这些表格补充了在 Section 2.2 和 Table 3 中找到的汇总相关性指标（Spearman's ρ 和 Kendall's τ ）。我们的目标是评估包括 VNLI-Critique 在内的自动化方法在事实准确性上，将生成标题的 VLMs 与人为派生的地面真实排名相比的排名效果。

提供了三个排行榜，每个排行榜对应一个不同的事实性标准：

1. 响应级别的正确性：完全事实准确段落的百分比（表 7 ）。
2. 整体正确句子：所有描述中正确句子的总百分比（表格 9 ）。
3. 每个描述中的正确句子数：每个描述中正确句子的平均百分比（表格 8 ）。

在每个排行榜（表 7 、9 和 8 ）中，行列出了来自 DOCCI-Critique 的 14 个生成标题的 VLM（详情见 Table 2 ）。列表示自动排名方法（例如，“我们的（VNLI-Critique）”、“GPT-4o”）。单元格显示该列的方法给予该行的 VLM 的排名，标上上标表示原始指标分数。最后两行报告了 Spearman 的 ρ （带有 p 值上标）和 Kendall 的 τ 与该标准的真实值的相关性，提供了对每个 AutoRater 性能的细致视图。

Table 7: 在 DOCCI-Critique 上 VLM AutoRater 关于响应级准确性的排名。每个单元格显示 $Rank^{Metric-Score}$ 。最后两行：与人类的 Spearman 的 $\rho^{p-value}$ 和 Kendall 的 $\tau^{p-value}$ 相关性。

Captioner VLM \ Ranking Method	Human	Ours	Gemini 2.0-Flash	GPT-4o	InstructBLIP	LLaVa-OV	Janus-Pro-7B	Qwen2.5-VL	mPLUG-Owl3-7B	Emu3-Chat
MiniGPT-4	$14^{0.04}$	$14^{0.06}$	$14^{0.09}$	$14^{0.12}$	$7^{0.96}$	$14^{0.42}$	$13^{0.35}$	$4^{0.00}$	$13^{0.11}$	$3^{0.93}$
MPlugOwl-2	$13^{0.11}$	$12^{0.12}$	$13^{0.28}$	$12^{0.18}$	$2^{0.98}$	$12^{0.56}$	$10^{0.56}$	$4^{0.00}$	$12^{0.17}$	$9^{0.55}$
LLaVA	$11^{0.19}$	$11^{0.17}$	$9^{0.40}$	$11^{0.26}$	$3^{0.97}$	$10^{0.71}$	$5^{0.68}$	$4^{0.00}$	$8^{0.21}$	$12^{0.48}$
PALI-5B	$12^{0.11}$	$13^{0.09}$	$12^{0.31}$	$13^{0.17}$	$9^{0.96}$	$11^{0.58}$	$14^{0.19}$	$4^{0.00}$	$14^{0.07}$	$2^{0.99}$
VILA	$10^{0.21}$	$10^{0.20}$	$10^{0.39}$	$9^{0.33}$	$13^{0.91}$	$9^{0.85}$	$1^{0.79}$	$4^{0.00}$	$8^{0.21}$	$4^{0.78}$
InstructBLIP	$9^{0.26}$	$9^{0.21}$	$11^{0.38}$	$10^{0.31}$	$9^{0.96}$	$13^{0.54}$	$9^{0.60}$	$2^{0.01}$	$6^{0.26}$	$10^{0.54}$
Molmo-7B-D	$8^{0.31}$	$8^{0.27}$	$7^{0.68}$	$8^{0.58}$	$1^{0.99}$	$5^{0.93}$	$4^{0.70}$	$4^{0.00}$	$8^{0.21}$	$4^{0.78}$
LLaVA-OV-7B-Chat	$7^{0.36}$	$6^{0.35}$	$8^{0.63}$	$5^{0.64}$	$12^{0.93}$	$8^{0.90}$	$8^{0.66}$	$4^{0.00}$	$4^{0.30}$	$8^{0.58}$
Qwen2-VL-7B-Instruct	$5^{0.41}$	$4^{0.45}$	$5^{0.75}$	$4^{0.65}$	$3^{0.97}$	$6^{0.91}$	$2^{0.78}$	$4^{0.00}$	$5^{0.29}$	$11^{0.49}$
LLaVA-OV-7B	$5^{0.41}$	$7^{0.34}$	$6^{0.72}$	$7^{0.63}$	$14^{0.90}$	$2^{0.97}$	$6^{0.67}$	$2^{0.01}$	$2^{0.41}$	$7^{0.61}$
Gemini-1.5-Pro	$4^{0.64}$	$5^{0.43}$	$4^{0.84}$	$3^{0.67}$	$3^{0.97}$	$6^{0.91}$	$12^{0.49}$	$4^{0.00}$	$7^{0.24}$	$14^{0.31}$
Gemini-1.5-Flash	$3^{0.68}$	$3^{0.52}$	$3^{0.88}$	$5^{0.64}$	$11^{0.94}$	$4^{0.94}$	$11^{0.51}$	$4^{0.00}$	$11^{0.18}$	$13^{0.40}$
mPLUG-Owl3-7B	$2^{0.71}$	$2^{0.69}$	$2^{0.93}$	$2^{0.77}$	$7^{0.96}$	$1^{0.98}$	$3^{0.73}$	$1^{0.02}$	$1^{0.75}$	$1^{1.00}$
GPT-4o[2024-08-06]	$1^{0.83}$	$1^{0.73}$	$1^{0.94}$	$1^{0.89}$	$3^{0.97}$	$2^{0.97}$	$6^{0.67}$	$4^{0.00}$	$3^{0.37}$	$6^{0.74}$
Spearman's Rank ρ	–	0.98^{5e-10}	$\underline{0.97}^{6e-9}$	0.92^{3e-6}	-0.06^{8e-1}	0.88^{2e-5}	0.30^{3e-1}	0.29^{3e-1}	0.73^{3e-3}	-0.2^{5e-1}
Kendall Tau τ	–	0.93^{4e-6}	$\underline{0.89}^{1e-5}$	0.82^{5e-5}	-0.05^{8e-1}	0.76^{2e-4}	0.21^{3e-1}	0.25^{3e-1}	0.27^{5e-3}	-0.17^{4e-1}

Table 8: VLM AutoRater 在 DOCCI-Critique 上正确句子平均百分比的排名。每个单元格显示 $Rank^{Metric-Score}$ 。最后两行：Spearman 的 $\rho^{p-value}$ 和 Kendall 的 $\tau^{p-value}$ 与人类的相关性。

Captioner VLM \ Ranking Method	Human	Ours	Gemini 2.0-Flash	GPT-4o	InstructBLIP	LLaVa-OV	Janus-Pro-7B	Qwen2.5-VL	mPLUG-Owl3-7B	Emu3-Chat
MiniGPT-4	$14^{0.46}$	$14^{0.48}$	$14^{0.53}$	$14^{0.42}$	$8^{0.99}$	$14^{0.84}$	$13^{0.77}$	$13^{0.05}$	$14^{0.51}$	$3^{0.99}$
MPlugOwl-2	$13^{0.53}$	$13^{0.54}$	$13^{0.66}$	$13^{0.56}$	$7^{0.99}$	$12^{0.89}$	$12^{0.86}$	$11^{0.05}$	$13^{0.53}$	$12^{0.86}$
LLaVA	$12^{0.60}$	$11^{0.59}$	$11^{0.75}$	$12^{0.59}$	$9^{0.99}$	$11^{0.91}$	$9^{0.90}$	$9^{0.08}$	$11^{0.58}$	$14^{0.83}$
InstructBLIP	$11^{0.61}$	$12^{0.58}$	$12^{0.71}$	$11^{0.62}$	$12^{0.98}$	$13^{0.86}$	$10^{0.89}$	$10^{0.05}$	$12^{0.56}$	$13^{0.84}$
PALI-5B	$10^{0.67}$	$10^{0.68}$	$10^{0.79}$	$10^{0.67}$	$3^{1.00}$	$10^{0.91}$	$14^{0.73}$	$12^{0.05}$	$10^{0.61}$	$2^{1.00}$
VILA	$9^{0.78}$	$8^{0.78}$	$9^{0.86}$	$9^{0.81}$	$11^{0.99}$	$9^{0.98}$	$1^{0.97}$	$4^{0.22}$	$8^{0.76}$	$4^{0.96}$
mPLUG-Owl3-7B	$8^{0.80}$	$6^{0.80}$	$4^{0.98}$	$8^{0.87}$	$13^{0.98}$	$5^{0.99}$	$11^{0.87}$	$14^{0.05}$	$2^{0.85}$	$1^{1.00}$
LLaVA-OV-7B	$7^{0.82}$	$6^{0.80}$	$7^{0.94}$	$6^{0.91}$	$14^{0.97}$	$1^{1.00}$	$5^{0.94}$	$6^{0.20}$	$7^{0.81}$	$8^{0.92}$
Molmo-7B-D	$6^{0.83}$	$9^{0.78}$	$6^{0.95}$	$7^{0.90}$	$1^{1.00}$	$7^{0.99}$	$4^{0.94}$	$8^{0.10}$	$9^{0.73}$	$5^{0.95}$
LLaVA-OV-7B-Chat	$5^{0.86}$	$5^{0.85}$	$8^{0.94}$	$4^{0.94}$	$10^{0.99}$	$8^{0.99}$	$3^{0.94}$	$1^{0.28}$	$1^{0.85}$	$7^{0.93}$
Qwen2-VL-7B-Instruct	$4^{0.88}$	$4^{0.88}$	$5^{0.97}$	$5^{0.93}$	$4^{1.00}$	$6^{0.99}$	$2^{0.97}$	$5^{0.22}$	$5^{0.83}$	$9^{0.91}$
Gemini-1.5-Pro	$3^{0.95}$	$3^{0.93}$	$3^{0.99}$	$2^{0.97}$	$2^{1.00}$	$4^{0.99}$	$8^{0.93}$	$3^{0.25}$	$4^{0.83}$	$11^{0.89}$
Gemini-1.5-Flash	$2^{0.96}$	$2^{0.94}$	$2^{0.99}$	$3^{0.95}$	$5^{0.99}$	$3^{0.99}$	$7^{0.93}$	$2^{0.27}$	$3^{0.84}$	$10^{0.90}$
GPT-4o[2024-08-06]	$1^{0.97}$	$1^{0.95}$	$1^{0.99}$	$1^{0.98}$	$6^{0.99}$	$2^{1.00}$	$6^{0.93}$	$7^{0.17}$	$6^{0.82}$	$6^{0.95}$
Spearman's Rank ρ	–	$\underline{0.97}^{1e-8}$	0.96^{9e-8}	0.99^{7e-11}	0.35^{2e-1}	0.85^{1e-4}	0.58^{3e-2}	0.70^{5e-3}	0.80^{6e-4}	0.00^{1e-0}
Kendall Tau τ	–	$\underline{0.91}^{7e-6}$	0.90^{2e-7}	0.93^{1e-8}	0.14^{5e-1}	0.74^{7e-5}	0.40^{4e-2}	0.50^{1e-2}	0.65^{7e-4}	-0.05^{8e-1}

B VNLI-Critique 模型开发细节

本节进一步详细介绍了用于开发我们的 VNLI-Critique 模型的架构、微调过程和计算资源，这些内容在主体文章的 Section 2 中介绍。

Table 9: DOCCI-Critique 上整体正确句子百分比的 VLM AutoRater 排名。每个单元格显示 $Rank^{Metric-Score}$ 。最后两行：与人为对比的 Spearman's $\rho^{p-value}$ 和 Kendall's $\tau^{p-value}$ 相关性。

Captioner VLM \ Ranking Method	Human	Ours	Gemini 2.0-Flash	GPT-4o	InstructBLIP	LLaVa-OV	Janus-Pro-7B	Qwen2.5-VL	mPLUG-Owl3-7B	Emu3-Chat
MiniGPT-4	$14^{0.48}$	$14^{0.49}$	$14^{0.55}$	$14^{0.44}$	$7^{0.99}$	$14^{0.83}$	$13^{0.77}$	$11^{0.06}$	$13^{0.52}$	$3^{0.99}$
MPlugOwl-2	$13^{0.52}$	$13^{0.53}$	$13^{0.64}$	$13^{0.56}$	$7^{0.99}$	$12^{0.87}$	$11^{0.85}$	$13^{0.05}$	$14^{0.51}$	$12^{0.86}$
InstructBLIP	$12^{0.57}$	$12^{0.57}$	$12^{0.68}$	$11^{0.59}$	$11^{0.99}$	$13^{0.84}$	$10^{0.87}$	$12^{0.05}$	$12^{0.54}$	$14^{0.82}$
LLaVA	$11^{0.59}$	$11^{0.59}$	$10^{0.74}$	$12^{0.59}$	$7^{0.99}$	$10^{0.90}$	$9^{0.89}$	$9^{0.08}$	$11^{0.58}$	$13^{0.83}$
PALI-5B	$10^{0.67}$	$10^{0.66}$	$10^{0.74}$	$10^{0.62}$	$4^{1.00}$	$11^{0.88}$	$14^{0.73}$	$10^{0.07}$	$10^{0.59}$	$2^{1.00}$
VILA	$9^{0.79}$	$8^{0.79}$	$9^{0.86}$	$9^{0.81}$	$11^{0.99}$	$9^{0.98}$	$1^{0.97}$	$4^{0.22}$	$8^{0.77}$	$4^{0.96}$
LLaVA-OV-7B	$8^{0.82}$	$7^{0.81}$	$8^{0.95}$	$6^{0.91}$	$13^{0.98}$	$1^{1.00}$	$5^{0.93}$	$6^{0.21}$	$2^{0.85}$	$9^{0.91}$
mPLUG-Owl3-7B	$7^{0.82}$	$6^{0.82}$	$4^{0.96}$	$8^{0.84}$	$14^{0.98}$	$8^{0.99}$	$12^{0.81}$	$14^{0.05}$	$3^{0.84}$	$1^{1.00}$
Molmo-7B-D	$6^{0.83}$	$9^{0.79}$	$6^{0.95}$	$7^{0.90}$	$1^{1.00}$	$6^{0.99}$	$4^{0.94}$	$8^{0.10}$	$9^{0.72}$	$5^{0.96}$
Qwen2-VL-7B-Instruct	$5^{0.87}$	$4^{0.88}$	$4^{0.96}$	$5^{0.93}$	$2^{1.00}$	$5^{0.99}$	$2^{0.97}$	$5^{0.22}$	$6^{0.83}$	$8^{0.91}$
LLaVA-OV-7B-Chat	$4^{0.87}$	$5^{0.88}$	$6^{0.95}$	$4^{0.94}$	$10^{0.99}$	$7^{0.99}$	$3^{0.96}$	$1^{0.31}$	$1^{0.87}$	$7^{0.92}$
Gemini-1.5-Pro	$3^{0.95}$	$3^{0.93}$	$3^{0.99}$	$2^{0.97}$	$2^{1.00}$	$4^{0.99}$	$8^{0.93}$	$3^{0.25}$	$5^{0.84}$	$11^{0.89}$
Gemini-1.5-Flash	$2^{0.96}$	$2^{0.94}$	$1^{0.99}$	$3^{0.95}$	$5^{1.00}$	$1^{1.00}$	$5^{0.93}$	$2^{0.27}$	$4^{0.84}$	$9^{0.91}$
GPT-4o[2024-08-06]	$1^{0.97}$	$1^{0.95}$	$1^{0.99}$	$1^{0.98}$	$5^{1.00}$	$1^{1.00}$	$5^{0.93}$	$7^{0.16}$	$7^{0.81}$	$6^{0.95}$
Spearman's Rank ρ	–	0.98^{1e-9}	$\underline{0.98}^{2e-9}$	0.97^{2e-9}	0.37^{2e-1}	0.86^{10e-5}	0.52^{6e-2}	0.69^{6e-3}	0.74^{2e-3}	0.06^{8e-1}
Kendall Tau τ	–	0.91^{5e-8}	$\underline{0.91}^{5e-8}$	0.91^{5e-8}	0.19^{4e-1}	0.76^{4e-5}	0.34^{1e-1}	0.52^{10e-3}	0.58^{3e-3}	0.01^{1e+0}

B.1 模型架构

VNLI-Critique 是通过微调 PaliGemma 10B 架构 [31] 开发的。该架构将 Gemma2-9B 大型语言模型 (LLM) [28] 作为其文本骨干，并使用 SigLIP 模型 [45] 作为其视觉编码器。在视觉处理方面，输入图像被标准化为 $448px^2$ 像素的分辨率。在此分辨率下，SigLIP 视觉编码器将每个图像处理为 1024 个视觉标记的序列，然后将其输入 LLM 组件以进行多模态理解和生成任务。

B.2 微调过程

我们对 PaliGemma 10B 模型进行了全面的微调以开发 VNLI-Critique。微调过程进行了 5 个周期。使用了批量大小为 128 的训练，并应用了 0.1 的 dropout 率以帮助正则化。在训练过程中没有使用权值衰减。使用默认超参数的 Adam 优化器 [18] 进行优化，并在整个微调过程中保持一个恒定的学习率 1×10^{-6} 。

B.3 计算资源

VNLI-Critique 模型的训练在 Google Cloud TPUv5e [11] 加速器上进行。具体而言，采用了 128 个 TPUv5e 芯片的配置来完成微调任务。5 个训练轮次的总耗时大约为 1 小时 30 分钟。基于每个芯片每小时 $ 1.20 的估算成本，训练 VNLI-Critique 的总计算成本约为 $ 230.40。

C 人工标注细节

DOCCI-Critique 基准的创建以及我们模型输出的评估，包括批判生成和批判与修订流程，依赖于全面的人类注释。我们通过 Prolific [2] 聘请了第三方人类注释者。每一个需要进行人工评价的数据条目，无论是 DOCCI-Critique 中的句子级事实性，还是生成的批判的质量评估，都是由五个不同的注释者独立评估的。这种多注释者的方法有助于确保收集到的判断的稳健性，并减轻个体偏见。注释者的工作报酬为每小时 $ 20。

以下小节对为两项主要人工评估任务设计的标注界面进行了说明性概述：评估 VLM 生成描述句子的真实性（第 C.1 节）以及评估生成批评意见的质量（第 C.2 节）。

C.1 描述句子注释界面

在对 VLM 生成的段落描述进行句子级别的事实性标注任务中（如第 3 节所述的 DOCCI-Critique 基准测试），标注者需要使用一个界面，该界面会显示源图像、完整段落上下文以及正在评估的具体句子。图 2 展示了这个标注界面的一个代表性例子。标注者需要判断该句子是否准确描述了图像内容，并给出标签，如“蕴含”、“中立”或“矛盾”，并为任何非蕴含的判断提供文本理由。

[2] https://www.prolific.com/

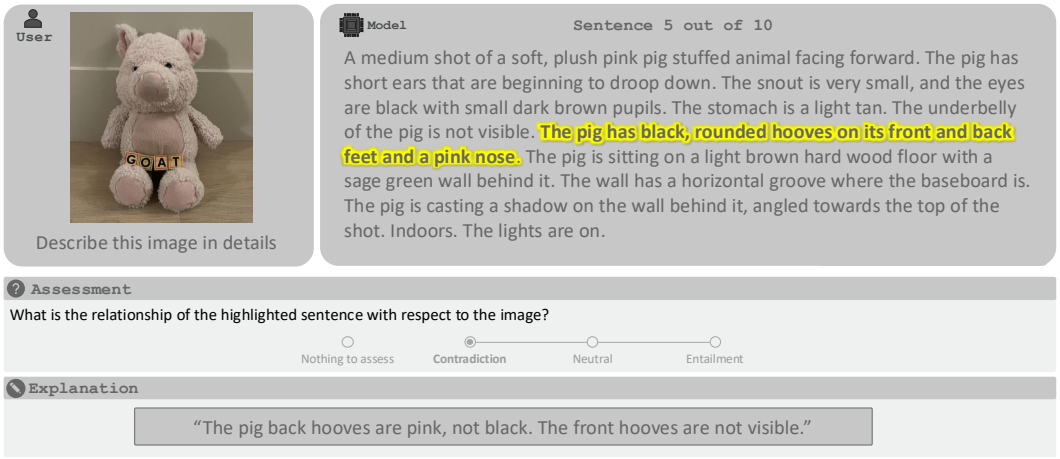

Figure 2: 描述句子注释界面的示例。注释者会看到图像、完整的 VLM 生成段落以及被高亮的句子。他们通过选择一个标签（这里为“矛盾”）并提供对于观察到的不准确之处的文字解释来评估其真实性。

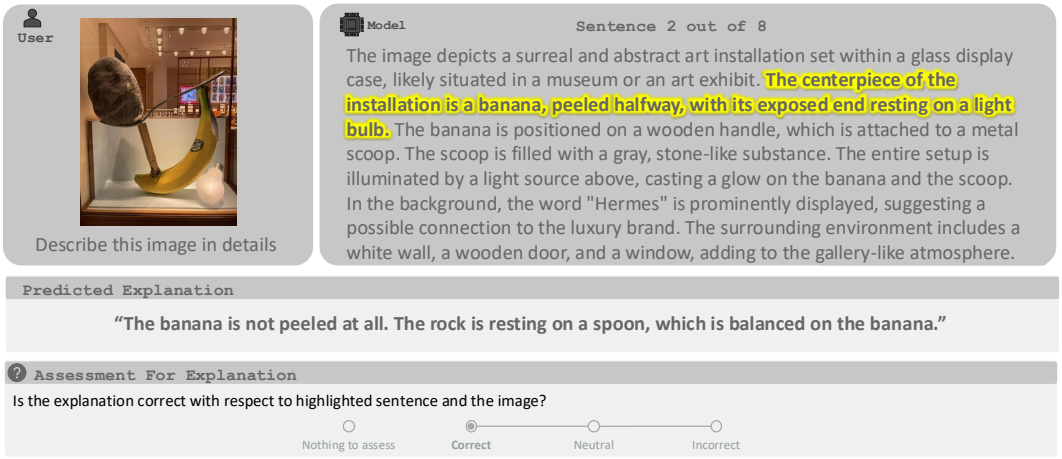

Figure 3: 评论注释界面的示例。注释者评估“预测的解释”是否正确识别出 VLM 中相对于图像的高亮句子的错误。

C.2 批注界面

为了评估 VNLI-Critique 和其他基线模型（如第 4.3 节所述）生成的评论的质量，使用了不同的界面。该界面向人工标注者展示了原始图像、被评论的事实性错误句子，以及模型生成的评论。图 3 显示了该界面的一个示例。标注者的任务是判断所提供的评论是否准确且相关地识别了与图像中的视觉证据相比，原句中存在的事实性错误。

D 定性例子

为了进一步说明我们工作的核心组件和输出，本节提供了额外的定性示例，以补充主论文中提出的讨论和汇总结果。

表格 10 展示了 DOCCI-Critique 基准的另一个详细条目。该示例突出了我们句子级别注释的细致性，包括多位标注者对句子是否对图像提出主张的判断、其基于视觉证据的事实正确性，以及标注者为任何识别到的不准确之处提供的多样化人写推理。这些示例强调了该基准在评估细微理解和错误分析方面的丰富性。

此外，表格 11 提供了关于我们的批评和修订流程的逐步演示，该流程应用于来自 Pixel-Prose [30] 数据集的图像描述。示例展示了：（1）原始的 VLM 生成的描述中包含事实错

Table 10: 额外的 DOCCI-Critique 基准注释示例（每个评估有 5 个评分者）。详细说明句子级别的主张、事实性和错误的多样性人类推理，展示不同的观点。

Image			
Description Sentence	"... Looking closely, we can see eight flamingos lined up. ..."	"... They are standing in a body of water, their reflection is seen in the water, and there are trees in the background. ..."	"... Flamingos primarily eat brine shrimp, blue-green algae, small insects, mollusks, and crustaceans ..."
Does the sentence include a claim about the image? (Answers from 5 raters)	✅, ✅, ✅, ✅, ✅	✅, ✅, ✅, ✅, ✅	❌, ❌, ❌, ❌, ❌
Is the sentence factual? (Answers from 5 raters)	❌, ❌, ❌, ❌, ❌	❌, ✅, ✅, ❌, ✅	✅, ✅, ✅, ✅, ✅
Rationales	• 八只火烈鸟的数量是不正确的；我能看到至少十只。 • 图中有 11 只火烈鸟。 • 指出的火烈鸟数量不正确，似乎实际数量更多。 • 我看到十一只火烈鸟，不是八只。 • 十一只火烈鸟排成一列，而不是八只。	• 我在背景中没有看到任何显著的树木，主要只是遥远、模糊的树叶或土地。 • 背景看起来更像是远处的海岸线或低矮的植被，而不是明显的树木。	-

误，(2) 由 VNLI-Critique 检测到的具体不实句子，(3) 由 VNLI-Critique 生成的相应批评，(4) 根据这些批评由 LLM 进行的各个句子的修订，以及 (5) 最终更为准确的修订版描述。这说明了我们的流程在自动纠正详细图像标题中的错误方面的实际应用。

Table 11: 表 11：通过 PixelProse 数据集中的一个样本，逐步说明 Critic-and-Revise 流程的实际操作。‘原始描述’ 包含若干不准确之处。‘ VNLI-Critique 检测到的不真实句子’ 强调了这些错误（例如，关于手的位置、光源、文字位置）。‘ VNLI-Critique 预测的批评’ 为这些错误提供了解释。‘Critic-and-Revise 输出’ 展示了 Large Language Model 在批评的指导下所纠正的个别句子。最后，‘修订后的描述’ 将这些更正整合成一个更加实际的段落。

Image			
Original Description	A young man with short brown hair and dark brown eyes. He is wearing a black jacket and a white shirt. He has a serious expression on his face. He is looking at the viewer with his left hand on his chin and the other holding his jacket. There is a dark background with some light coming from the left side of the image. There is text at the top of the image that says “The right to use my friends as a weapon, that is the sinful crown I shall adorn - Shu Ouma”. The text is in a white font. The image is in an anime style.		
Detected Unfactual Sentences by VNLI-Critique	He is looking at the viewer with his left hand on his chin and the other holding his jacket.	There is a dark background with some light coming from the left side of the image.	There is text at the top of the image that says “The right to use my friends as a weapon, that is the sinful crown I shall adorn - Shu Ouma”.
Predicted Critiques by VNLI-Critique	He is looking at the viewer but his hands are not visible	The light is coming from the right side of the image, not the left.	The text is at the bottom of the image and not the top.
Critic-and-Revise output	He is looking at the viewer.	There is a dark background with some light coming from the right side of the image.	There is text at the bottom of the image that says “The right to use my friends as a weapon, that is the sinful crown I shall adorn - Shu Ouma”.
Revised Description	A young man with short brown hair and dark brown eyes. He is wearing a black jacket and a white shirt. He has a serious expression on his face. He is looking at the viewer. There is a dark background with some light coming from the right side of the image. There is text at the bottom of the image that says “The right to use my friends as a weapon, that is the sinful crown I shall adorn - Shu Ouma”. The text is in a white font. The image is in an anime style.		