UNBLOCKING FINE-GRAINED EVALUATION OF DETAILED CAPTIONS: AN EXPLAINING AUTORATER AND CRITIC-AND-REVISE PIPELINE

Anonymous authors
Paper under double-blind review

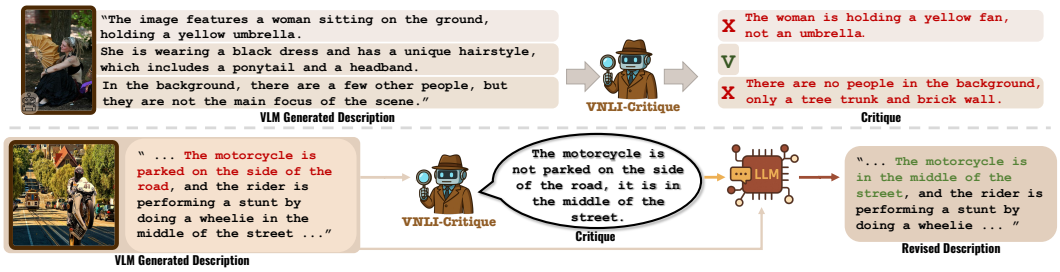

Figure 1: VNLI-Critique in action: operating as a Critic and within the Critic-and-Revise pipeline. (Top) As a Critic, VNLI-Critique evaluates sentence-level factuality in VLM captions and generates error explanations. (Bottom) In the pipeline, its critique of an incorrect sentence guides an LLM to revise it, demonstrating automated evaluation and correction of detailed captions.

ABSTRACT

Large Vision-Language Models (VLMs) now generate highly detailed, paragraph-length image captions, yet evaluating their factual accuracy remains challenging. Current methods often miss fine-grained errors, being designed for shorter texts or lacking datasets with verified inaccuracies. We introduce *DOCCI-Critique*, a benchmark with 1,400 VLM-generated paragraph captions (100 images, 14 VLMs) featuring over 10,216 sentence-level human annotations of factual correctness and explanatory rationales for errors, all within paragraph context. Building on this, we develop *VNLI-Critique*, a model for automated sentence-level factuality classification and critique generation. We highlight three key applications: (1) VNLI-Critique demonstrates robust generalization, validated by state-of-the-art performance on the M-HalDetect benchmark and strong results in CHOCOLATE claim verification. (2) The VNLI-Critique driven AutoRater for DOCCI-Critique provides reliable VLM rankings, showing excellent alignment with human factuality judgments (e.g., 0.98 Spearman). (3) An innovative Critic-and-Revise pipeline, where critiques from VNLI-Critique guide LLM-based corrections, achieves substantial improvements in caption factuality (e.g., a 46% gain on DetailCaps-4870). Our work offers a crucial benchmark alongside practical tools, designed to significantly elevate the standards for fine-grained evaluation and foster the improvement of VLM image understanding.

	Humans	VNLI Critique	GPT-4	Qwen 2.5-VL	Gemini 2.0-Flash
"The image features a person riding a motorcycle down a city street.	V	V	V	V	V
The motorcycle is parked on the side of the road, and the rider is performing a stunt by doing a wheelie in the middle of the street.	X	X	X	X	X
In the background, a cable car can be seen traveling along the street, adding to the lively atmosphere of the scene.	V	V	V	V	V
There are several cars parked along the street, some closer to the motorcycle and others further away.	V	V	V	X	X
Additionally, there are several people visible in the scene, observing the motorcyclist and the cable car passing by."	X	X	V	X	X

Figure 2: Sentence-level factuality assessment by VNLI-Critique. Figure shows an image, VLM-generated caption, and factuality judgments (V Correct / X Incorrect), illustrating VNLI-Critique's fine-grained, human-aligned fact-checking compared to zero-shot VLMs. Errors highlighted in red.

1 Introduction

Automatic descriptive image captioning, a prominent vision-language research area (Stefanini et al., 2023), has evolved from short highlights (Saouabe et al., 2023) to detailed, paragraph-length descriptions, thanks to powerful Large Vision-Language Models (LVLMs) (Chen et al., 2025; Dai et al., 2023; Li et al., 2024; Steiner et al., 2024; Wang et al., 2024b; Ye et al., 2024a). Evaluating these complex captions remains challenging; current metrics, often for short texts, miss subtle, fine-grained details and typically assess sentences in isolation, lacking crucial paragraph context for resolving ambiguities and co-references. While some studies address full paragraphs as Dong et al. (2024), granular sentence-level assessment is still difficult.

Existing VLM factuality benchmarks (e.g., for error/hallucination) often target short sentences or QA, inadequately addressing paragraph-length descriptions. While valuable, datasets like M-HalDetect (Gunjal et al., 2024) and CHOCOLATE (Huang et al., 2024) provide sentence-level annotations but may lack the error diversity of long-form VLM outputs or the consistent, detailed textual explanations for inaccuracies vital for fine-grained automated evaluation. Response-level evaluations like CAPTURE (Dong et al., 2024) score paragraphs, missing sentence-level granularity. A benchmark is critically needed, providing comprehensive, context-aware sentence-level factuality annotations, including explanatory rationales for errors, across diverse VLM-generated paragraphs.

To address this, we introduce *DOCCI-Critique*, a novel benchmark for fine-grained evaluation of detailed image descriptions. It comprises 1,400 paragraph-length captions (14 VLMs, 100 images), with its core value in 10,216 sentence-level human annotations. Each sentence's factuality was judged by five annotators, with detailed textual rationales for every identified error. This multi-perspective annotation offers a new resource for in-depth VLM analysis, providing multiply-verified sentence-level judgments and error explanations for long captions, distinct from existing datasets.

Building on DOCCI-Critique, we developed VNLI-Critique, a model for automated sentence-level factuality classification (using paragraph context) and explanatory critique generation. This dual capability enables a novel Critic-and-Revise pipeline (Fig. 1): VNLI-Critique evaluates and generates a critique for an incorrect VLM-generated sentence, guiding an LLM to revise it. The utility of VNLI-Critique and this pipeline is shown via key results: (1) VNLI-Critique achieves state-of-the-art performance on external benchmarks Gunjal et al. (2024) (0.76 Macro-F1) and competitive results on Huang et al. (2024), demonstrating strong generalization. (2) On our benchmark, the VNLI-Critique powered DOCCI-Critique AutoRater shows VLM rankings with strong correlation to human judgments (Table 3). (3) The Critic-and-Revise pipeline significantly boosts factuality of incorrect sentences (e.g., by 46% on Dong et al. (2024), 51% on Singla et al. (2024)), confirmed by human evaluation. Collectively, these contributions deliver an essential benchmark and powerful methodologies, enabling more precise fine-grained assessment and demonstrably enhancing the factual accuracy of detailed image understanding by VLMs.

2 Related Work

Our work intersects with advancements in Vision-Language Models (VLMs) for detailed captioning, the development of image captioning datasets, and methodologies for evaluating caption quality, especially factual accuracy and fine-grained detail.

Vision-Language Models Recent VLMs (Steiner et al., 2024; OpenAI et al., 2024; Bai et al., 2025; Li et al., 2024; Liu et al., 2024; Wang et al., 2024b; Deitke et al., 2024; Ye et al., 2024b; Chen et al., 2023; Lin et al., 2024; Dai et al., 2023; Zhu et al., 2023; Team et al., 2024) have achieved remarkable SOTA performance in multimodal tasks . Typically, they combine a visual encoder Radford et al. (2021); Zhai et al. (2023); Tschannen et al. (2025); Dosovitskiy et al. (2021) with an LLM (Yang et al., 2024; Chung et al., 2024; Chiang et al., 2023; Touvron et al., 2023), often using a connector module to bridge visual features with textual tokens. Training usually involves pre-training the visual encoder and then fine-tuning the LLM, with objectives like masked language modeling or image-text matching. While end-to-end training is explored, this two-stage approach is common, balancing dataset scale, computational resources, and evaluation.

Image Captioning Datasets Image understanding and captioning datasets are crucial for advancing image captioning techniques. Early datasets like COCO (Lin et al., 2014), Flickr8k (Hodosh et al., 2013), and Flickr30k (Young et al., 2014) offered short-sentence, positive image-text pairs, primarily focusing on main objects and scenes. More recent datasets offer image-text pairs with longer, more detailed descriptions. The DCI dataset (Urbanek et al., 2024), for instance, introduces long, mask-aligned descriptions to evaluate VLM understanding of distinct image regions. PixelProse (Singla et al., 2024) offers 16.9 million synthetically generated captions using Gemini-1.0-Pro-Vision (Team et al., 2024); however, their correctness is not guaranteed. The IIW (Garg et al., 2024) dataset uses a VLM to generate initial captions, then a human-in-the-loop framework to ensure high-quality positive image-text pairs. M-HalDetect (Gunjal et al., 2024) and CHOCOLATE (Huang et al., 2024) use various VLMs to caption images and charts, respectively, then use sentence-level human annotation to assess correctness. DetailCaps (Dong et al., 2024) uses GPT-4 (OpenAI et al., 2024) to assign a quality score (0-5) to existing synthetic captions. The DOCCI dataset (Onoe et al., 2024) is a valuable resource for training and evaluating VLMs, offering high-resolution images of diverse real-life scenes paired with detailed, fully human-written descriptions. We leverage DOCCI as the image foundation for our DOCCI-Critique benchmark, for which we gathered over 10,216 new sentence-level factuality annotations. Furthermore, DOCCI's commercial-friendly license and its use of original, author-taken images without people address significant attribution and privacy challenges common in web-scraped datasets.

Evaluation of Detailed Captions Traditional metrics (e.g., BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), CIDEr (Vedantam et al., 2015)) use n-gram overlap to compare captions to references, often missing semantic nuance crucial for paragraph-length text. Embedding-based methods (e.g., CLIPScore (Hessel et al., 2021), SIGLip (Zhai et al., 2023)) offer better semantic assessment but evaluate sentences in isolation, lacking the paragraph context needed to resolve ambiguities or co-references. QA-based metrics (e.g., VQAScore (Lin et al., 2025), TIFA (Hu et al., 2023), VQ^2 (Yarom et al., 2023), GECKO (Wiles et al., 2025)) assess understanding via QA but face scalability challenges for long narratives. Recent work like Gordon et al. (2024) provides detailed feedback for misalignments, but primarily for text-to-image and shorter textual inputs. Response-level evaluations like Dong et al. (2024) score paragraphs but lack sentence-level factuality details. Our work provides the needed fine-grained, context-aware sentence-level factuality evaluation, including rich, human-verified explanatory feedback for detailed, paragraph-length image captions.

3 The DOCCI-Critique Benchmark: Construction, Annotation, and Analysis

DOCCI-Critique is a novel benchmark for fine-grained factuality assessment of paragraph-level image descriptions. Its primary purpose is twofold: (1) providing a robust platform to assess SOTA captioning models' descriptive capabilities and factual accuracy, and (2) serving as a challenging testbed for automated image understanding and fact-checking systems (auto-raters).

Construction began with 100 diverse, high-resolution images from the DOCCI dataset's 'qual-dev' split (Onoe et al., 2024). For each, 14 SOTA Large Vision-Language Models (Table 2) generated detailed, paragraph-length descriptions, yielding 1,400 model-generated descriptions. This corpus deliberately captures a wide spectrum of stylistic variations, detail levels, and factual accuracies, from concise and factual to more verbose accounts that might introduce subtle inconsistencies (e.g., object misidentification).

Table 1: Illustrative example from the DOCCI-Critique benchmark, detailing sentence-level annotations for fine-grained factuality assessment, including rater judgments and explanatory rationales.

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	-	• The 16 number is not on the side of the racing car, but in the front of it. • The mural does feature a red Formula 1 race car, but the number 16 is painted on the front not the side	• There is no checkered flag visible in the image. • There's no checkered flag in the poster/mural. • The background mural does feature the colors blue, white and red, but there is no checkered flag

The core of DOCCI-Critique is its rich human annotation layer. Following Steiner et al. (2024) (see Appendix for annotation template), five human annotators independently evaluated each sentence within the 1,400 descriptions for factual correctness against the image, assigning labels: 'Entailment' (factually supported), 'Neutral' (not verifiable/contradicted), 'Contradiction' (factually contradicted), or 'Nothing to assess' (e.g., filler). A sentence is classified as 'Non-entailment' if a majority labeled it 'Neutral' or 'Contradiction'. Crucially, annotators provided detailed textual rationales for each non-entailed judgment. Thus, a single non-entailed sentence often has multiple rationales, capturing diverse perspectives on the inaccuracy and offering insights into VLM error types (e.g., object misidentification, attribute/spatial errors, hallucination).

Table 1 illustrates this structure, showing per-sentence annotations: five independent factuality judgments (✅ for Entailment, ❌ for Contradiction/Neutral) and image content reliance. For non-entailed votes, collected textual rationales explain the specific error, as exemplified by the car mural's number placement or the non-existent checkered flag. Multiple rationales for one incorrect sentence reflect diverse annotator viewpoints.

This comprehensive annotation resulted in 10,216 sentence-level judgments. The dataset—with its diverse VLM outputs, fine-grained majority-vote factuality labels, and multiple rich explanatory rationales per error—is an invaluable resource. It enables rigorous VLM evaluation beyond surface-level similarity, allowing deeper probes into model understanding and descriptive fidelity. For more details on the human annotation process, including the task format and inter-rater agreement, please refer to Appendix C.

Table 2 details per-model statistics from DOCCI-Critique, including description and sentence lengths, factual accuracy, and lexical diversity. These internal statistics reveal quantitative and qualitative VLM behavioral differences. For instance, high-accuracy models like GPT-4o contrast with Gemini models that produce more sentences with comparable accuracy, suggesting a verbosity/detail vs. error-risk trade-off. Paragraphs in DOCCI-Critique average 752.7 characters in length. This is considerably longer than in other contemporary datasets used for factual error analysis, such as M-HalDetect (averaging 456.2 characters), CHOCOLATE (577.6), and DetailCaps-4870 (612.9). This emphasis on longer, more intricate descriptions, combined with patterns like common error types (discernible from rationales, though not directly in Table 1), highlights DOCCI-Critique's utility for nuanced comparative studies of VLM generation strategies and visual faithfulness.

Table 2: DOCCI-Critique statistics, detailing paragraph-level metrics and lexical diversity (unique 2-grams) for each LLM's generated image descriptions.

	Description Length avg.	# Sentences Avg	Sentence Length Avg	% Correct Sentence in Description	Uni. 2-gram
MiniGPT-4 (Zhu et al., 2023)	483.5	5.6	84.8	45.6	4,695
mPLUG-Owl2-7B (Ye et al., 2024b)	458.8	4.4	102.1	52.7	4,038
LLaVa-1.5-7B (Liu et al., 2024)	395.4	4.2	91.5	60.0	3,081
InstructBLIP (Dai et al., 2023)	509.8	4.0	195.4	61.3	3,260
PALI-5B (Chen et al., 2023)	1098.9	10.9	69.5	68.0	1,881
VILA (Lin et al., 2024)	870.7	8.6	100.4	78.1	6,841
mPLUG-Owl3-7B (Ye et al., 2024a)	118.0	2.0	65.2	80.4	700
LLaVA-Onevision-7B (Li et al., 2024)	672.0	6.4	107.7	81.8	5,878
Molmo-7B-D (Deitke et al., 2024)	747.6	6.6	111.9	82.7	6,788
LLaVA-Onevision-7B-Chat (Li et al., 2024)	1091.6	9.5	113.1	85.7	8,550
Qwen2-VL-7B-Instruct (Wang et al., 2024a)	1022.6	9.8	102.9	87.6	8,250
Gemini-1.5-Pro (Team et al., 2024)	1326.9	12.0	109.3	95.1	11,705
Gemini-1.5-Flash (Team et al., 2024)	1199.0	11.8	100.0	96.1	10,186
GPT-4o [2024-08-06] (OpenAI et al., 2024)	583.5	6.2	94.2	97.1	6,160
TOTAL	752.7	7.3	103.8	76.5	40,444

4 VNLI-CRITIQUE: DEVELOPMENT AND EVALUATION

4.1 VNLI-CRITIQUE MODEL DEVELOPMENT

We developed *VNLI-Critique* by fine-tuning the 10B parameter PaliGemma-2 architecture (Steiner et al., 2024) (details in Appendix) for automated sentence-level factuality assessment and critique generation. This required a specialized training dataset of VLM-generated captions, distinct from DOCCI-Critique, annotated for factuality and error critiques. To create this diverse training data, we first generated paragraph-length captions using over 70 PaliGemma-2 variants (fine-tuned with varied configurations on DOCCI training data (Onoe et al., 2024)) to capture many generation styles and potential errors. These synthetic captions were then human-annotated per the protocol in Section 3 (majority vote for labels; longest rationale for non-entailed sentences as critique target).

VNLI-Critique was fine-tuned on this curated data for two tasks using specific prompts incorporating paragraph context (<PREFIX>Claim-Prefix</PREFIX>), vital for accurate assessment of potentially ambiguous standalone sentences. For **Factuality Classification**, the prompt was: *"Given the image and the prompt prefix <PREFIX>Claim-Prefix</PREFIX>, does the following text align with the image: <TARGET>Target-Claim</TARGET>?"*, requiring a "Yes"/"No" prediction. For **Critique Generation**, the prompt was: *"Given the image and the prompt prefix <PREFIX>Claim-Prefix</PREFIX>, the text <TARGET>Target-Claim</TARGET> is considered inaccurate. Explain the misalignments and factual inaccuracies that make it inaccurate."*. This dual-task strategy enables VNLI-Critique to identify discrepancies and articulate their reasons.

4.2 FACTUALITY CLASSIFICATION: BENCHMARKING AND GENERALIZATION RESULTS

This section details the performance of VNLI-Critique in its factuality classification task, presenting key results from its application as an automated benchmarking tool on DOCCI-Critique and its generalization capabilities when tested on diverse external datasets.

The DOCCI-Critique AutoRater: Automated VLM Benchmarking Results. A primary application of VNLI-Critique's classification capability is to serve as an AutoRater for establishing an automated leaderboard that ranks Vision-Language Models (VLMs) based on their factual accuracy when describing images from the DOCCI-Critique benchmark. The objective is to provide a scalable and reliable alternative to extensive human evaluation for this task. To assess its viability, we evaluated VNLI-Critique alongside other VLM-based methods as potential automated rankers. We compared how well their automated assessments correlated with human judgments across three distinct factuality criteria: (1) Response-Level Correctness (whether the entire generated paragraph was factually accurate), (2) Percentage of Correct Sentences Overall (total correct sentences across all generated descriptions for a model), and (3) Average Percentage of Correct Sentences per Description. The detailed leaderboards showing the rankings of VLMs for each of these criteria, as

determined by both human evaluation and the automated methods, are provided in Appendix. The correlation results, using Spearman's ρ (Sp ρ) and Kendall's τ (Kd τ), are presented in Table 3. **VNLI-Critique demonstrates exceptional performance as an AutoRater, achieving the highest Spearman correlation with human rankings** on Response-Level Correctness (Sp ρ = 0.981) and Percentage of Correct Sentences Overall (Sp ρ = 0.979), and a very high correlation for Average Percentage of Correct Sentences per Description (Sp ρ = 0.968). Its strong performance across these different evaluation granularities, significantly aligning with human assessments, validates its effectiveness as a reliable tool for automatically benchmarking VLM factuality on the DOCCI-Critique dataset.

Evaluating Broader Applicability on External Benchmarks. To assess VNLI-Critique's capabilities beyond our specific benchmark, we evaluate its performance on two established external datasets: M-HalDetect (Gunjal et al., 2024), a benchmark for detecting hallucinations in VLM descriptions of diverse images, and CHOCOLATE (Huang et al., 2024), which focuses on descriptions of charts and plots. We compare VNLI-Critique against various baselines, including other VLM-based classifiers and embedding-similarity methods, using two standard meta-evaluation metrics: ROC-AUC and Macro-F1. Figure 2 provides a qualitative example of these sentence-level classification comparisons.

The performance metrics reported in Table 4 were generated based on each model's output type. For models trained to classify via specific output tokens (e.g., 'Yes' and 'No') – this includes our VNLI-Critique and other VLM-based classifiers (where scores are derived via a 5-sample strategy) – metrics reflect both confidence and prediction. For all such VLM classifiers, the entailment score for ROC-AUC calculation is obtained by applying a softmax function to the confidence scores associated with the positive and negative classification outputs (yielding a normalized positive probability). The binary classification ('Accurate' or 'Inaccurate') for Macro-F1 is determined by selecting the label with the higher confidence score. In contrast, for methods like CLIPScore (Hessel et al., 2021), SigLIP (Zhai et al., 2023), and TIFA (Hu et al., 2023), which output numerical similarity scores, only ROC-AUC is reported, as it directly applies to such scores without requiring an arbitrary threshold.

As shown in Table 4, **VNLI-Critique achieves state-of-the-art (SOTA) performance on M-HalDetect. Furthermore, its highly competitive performance on the CHOCOLATE dataset demonstrates notable adaptability and robust reasoning capabilities**, even when evaluating descriptions of charts and graphs without specific training on such visual data. These strong findings across different benchmarks underscore the general utility of our fine-tuned model for factual verification tasks.

4.3 Evaluating Critique Generation

Beyond classifying sentences' correctness, a key capability of VNLI-Critique is generating textual critiques that explain why a sentence is factually inaccurate. To assess the quality and correctness of these generated explanations, we conducted a dedicated human evaluation study. The evaluation proceeded as follows: First, we sampled sentences previously identified by human annotators as incorrect from our DOCCI-Critique and M-HalDetect benchmarks. For each, we prompted VNLI-

Table 3: Evaluating Automated Methods as AutoRaters. Correlation (Spearman's ρ, Kendall's τ) between model-based rankings and human judgments of VLM factuality on DOCCI-Critique across three accuracy metrics. **Bold** indicates the best score, and underline indicates the second best.

	% 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: Evaluating VNLI-Critique's factuality classification: Comparison with baselines on in-distribution (DOCCI-CRITIQUE) and external (M-HalDetect, CHOCOLATE) datasets. Key results include SOTA on M-HalDetect and strong generalization to CHOCOLATE.

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

Critique and other competitive VLMs (Table 5) to generate an explanation detailing the specific misalignments, using the prompt format from Section 4.1. Human annotators were then presented with evaluation instances containing: (1) the original image, (2) the incorrect sentence, and (3) the model-generated critique. Annotators judged if the critique accurately and relevantly identified the factual error(s) based on the image. Table 5 reports the percentage of critiques deemed correct and relevant.The results (Table 5) demonstrate the effectiveness of our specialized, open-source VNLI-Critique (a 10B parameter model). Despite its significantly smaller scale, our model achieves critique generation quality comparable to, and in some cases exceeding, large-scale state-of-the-art commercial models. On DOCCI-Critique sentences, VNLI-Critique achieves the highest score (73.39%), slightly surpassing GPT-4o (73.1%). On M-HalDetect, it scores 79.33%, performing comparably to Gemini-2.0-Flash (79.89%) and again surpassing GPT-4o (78.77%). Notably, VNLI-Critique also significantly outperforms other open-source VLMs of a similar size, such as Janus-Pro-7B, Qwen-2.5-VL-Instruct, and LLaVA-OV on both datasets. This highlights VNLI-Critique's ability to generate high-quality, accurate explanations, providing a crucial capability for interpretable feedback and downstream correction that is accessible to the wider research community.

Table 5: Human evaluation of critique quality. Percentage of generated explanations judged as correct and relevant for incorrect sentences sampled from DOCCI-Critique and M-HalDetect.

	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>

5 CRITIC-AND-REVISE

Many vision-language tasks, from image captioning to text-to-image generation, rely heavily on large-scale datasets of image-text pairs, often utilizing VLM-generated captions for training or as part of their data. For example, large synthetic caption datasets like PixelProse (Singla et al., 2024) are used to train captioning models, and datasets pairing images with descriptive text are fundamental for training text-to-image synthesis models (e.g., leveraging datasets like Schuhmann et al. (2022)). However, the factual accuracy and visual alignment of these automatically generated or

web-crawled captions can vary, potentially introducing noise or inaccuracies into downstream model training. Improving the quality and factual alignment of such datasets is therefore crucial for advancing these fields. Leveraging the critique generation capability of VNLI-Critique, we introduce and evaluate a novel *Critic-and-Revise* pipeline. This pipeline is designed not only to correct individual factual inaccuracies in image captions but also offers a pathway to enhance the overall quality of image-text training datasets, thereby potentially improving the performance of models trained on them. This section first outlines the pipeline's methodology (Section 5.1). We then evaluate its applicability in correcting synthetically generated captions from large-scale datasets (Section 5.2).

5.1 Pipeline Methodology

The Critic-and-Revise pipeline, illustrated in Figure 1, operates in two main steps. First, in the *Critic step*, VNLI-Critique analyzes each sentence of a given caption using its classification function; sentences identified as factually inaccurate trigger the generation of a textual critique explaining the specific error based on the image content. Subsequently, in the *Revise step*, the original inaccurate sentence and its corresponding critique from VNLI-Critique are used to instruct a separate Large Language Model (LLM) to fix the inaccurate description. For our experiments, we utilized Gemini-2.0-Flash[1] as the revision LLM. This revision LLM is prompted to rewrite the original sentence, specifically addressing the factual errors highlighted in the critique, while aiming to preserve relevant information and maintain stylistic coherence. The full Critic-and-Revise cycle—factuality classification by VNLI-Critique for all sentences, followed by critique-guided revision for those flagged as inaccurate—produces a revised caption with enhanced factual alignment to the image.

5.2 Correcting Synthetically Generated Captions

To demonstrate the downstream utility of our proposed Critic-and-Revise pipeline, we applied it to captions from two large-scale datasets known for detailed yet potentially unverified descriptions: PixelProse (Singla et al., 2024), featuring 16.9M synthetic caption pairs, and DetailCaps-4870 (Dong et al., 2024), a subset with 4,870 images each accompanied by three detailed synthetic captions. We conducted a human study to evaluate the pipeline's effectiveness: after VNLI-Critique identified and critiqued inaccurate sentences, and the revision LLM corrected them, human evaluators assessed the factual correctness of both the original flagged sentences (for critic precision) and the revised sentences (for pipeline effectiveness).

The results, summarized in Table 6, highlight significant improvements. For DetailCaps-4870, while VNLI-Critique's initial flagging showed a 15% false positive rate (original sentences deemed correct by humans), **the pipeline successfully corrected a large portion of the genuinely inaccurate sentences,** with human judges confirming 61% of the revised sentences as factually accurate. This represents a 46% gain in accuracy for the set of sentences initially considered incorrect by the critic. VNLI-Critique's own re-evaluation classified 64% of these revised sentences as accurate, indicating strong self-consistency. Similar positive trends were observed for PixelProse, where human judges found 75% of revised sentences to be accurate (a 51% gain), demonstrating the pipeline's capability to enhance factual accuracy in detailed image captions at scale. Qualitative examples illustrating the Critic-and-Revise process, including original incorrect sentences, critiques from VNLI-Critique, and the LLM-revised sentences, are provided in Appendix.

Table 6: Critic-and-Revise Pipeline factuality: Human and VNLI-Critique judgments on original vs. revised claims. Δ = accuracy increase post-revision. **The pipeline markedly improves claim accuracy (human-confirmed to 61% on DetailCaps, 75% on PixelProse for fixed claims, from low initial values)**. VNLI-Critique's judgments align, showing high self-consistency.

	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%

[1]Accessed via the Vertex AI API: `https://cloud.google.com/vertex-ai`

6 Limitations and Future Work

The DOCCI-Critique benchmark, while richly annotated with 1,400 VLM-generated captions and over 10,000 sentence judgments, is constructed from a base set of 100 unique images. While these images provide diversity and the caption variations are extensive, expanding the number of base images could further enhance the benchmark's statistical power and coverage. Nevertheless, our experiments demonstrate strong generalization capabilities. Specifically, VNLI-Critique, when trained for factuality classification on DOCCI-Critique, performs well on external, unseen claim verification datasets like M-HalDetect and CHOCOLATE (Section 4.2). Furthermore, our Critic-and-Revise pipeline, leveraging critiques from VNLI-Critique, effectively corrects captions on entirely different datasets such as DetailCaps-4870 and PixelProse (Section 5.2). This collective evidence of generalization across different tasks and datasets suggests the current DOCCI-Critique size is effective for developing robust and transferable evaluation models and correction methodologies.

Additionally, while VNLI-Critique achieves strong results in several settings, its performance is not perfect. Our pipeline evaluation (Section 5) indicates that its factuality classification can result in false positives and false negatives (e.g., a 15% false positive rate on DetailCaps-4870). The quality of its generated critiques, while generally high as shown by human evaluations in Table 5, can also exhibit variability. Future work could enhanceVNLI-Critique's performance by further leveraging our rich annotations. For example, one could investigate methods to merge multiple rationales provided by different annotators into a single, more comprehensive and concise explanation, instead of solely using the longest rationale as the training target for critique generation. Furthermore, our annotation protocol captures whether a sentence's factuality is dependent on the image content or relies on world knowledge (e.g., distinguishing "The cat is on the mat" which requires the image, from "A cat is a mammal" which does not). This currently unused label could enable a two-stage verification process: first classifying if image grounding is needed, and then applying either the visual fact-checker (VNLI-Critique) or a knowledge-based verifier accordingly, potentially improving overall accuracy and efficiency.

Regarding the Critic-and-Revise pipeline, its current design involves two distinct steps: critique generation by VNLI-Critique followed by revision using a separate LLM. While effective, this contrasts with a hypothetical end-to-end model that might directly output a corrected sentence. However, we argue that the two-step approach offers significant advantages in interpretability. Generating an explicit critique allows for a clear understanding of why a sentence was flagged and what specific error is being addressed. This insight into error types and sources is valuable for analyzing and improving the underlying captioning models, a benefit potentially lost in a direct, black-box correction approach. Therefore, while future work might explore direct revision models, the explanatory power of the intermediate critique remains a key strength of our pipeline.

Addressing these limitations and exploring the suggested avenues for leveraging the full extent of the dataset annotations offers exciting directions for future research in robust and interpretable evaluation of detailed image understanding.

7 Conclusion

This work tackled the critical challenge of evaluating and improving the factual accuracy of detailed, paragraph-length VLM-generated image captions. We introduced *DOCCI-Critique*, a novel benchmark featuring 1,400 VLM captions with over 10,216 sentence-level human annotations for factuality, including explanatory rationales for errors, providing a vital resource for fine-grained VLM assessment. Building on this, we developed VNLI-Critique, a model proficient in automated factuality classification and critique generation. VNLI-Critique demonstrated strong generalization with state-of-the-art results on external datasets like M-HalDetect, and its use in the DOCCI-Critique showed high correlation with human judgments (0.98 Spearman). Furthermore, we presented a novel Critic-and-Revise pipeline where VNLI-Critique's critiques guide an LLM to automatically correct factual errors, significantly improving caption accuracy as confirmed by human evaluation. Collectively, DOCCI-Critique, VNLI-Critique, and the Critic-and-Revise pipeline offer essential tools and methodologies for advancing VLMs towards generating more detailed, fluent, and factually reliable image descriptions. Future work, as outlined in Section 6, will explore expanding the benchmark and further enhancing the pipeline's capabilities.

REFERENCES

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.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909/.

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. URL https://openreview.net/forum?id=mWVoBz4W0u.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

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, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

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), January 2024. ISSN 1532-4435.

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*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

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. URL https://doi.org/10.48550/arXiv.2409.17146.

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.

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. URL https://openreview.net/forum?id=YicbFdNTTy.

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 Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 93–127, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.6. URL https://aclanthology.org/2024.emnlp-main.6/.

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

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*, pp. 310–328, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72997-3. doi: 10.1007/978-3-031-72998-0_18. URL https://doi.org/10.1007/978-3-031-72998-0_18.

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'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i16.29771. URL https://doi.org/10.1609/aaai.v38i16.29771.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.595. URL https://aclanthology.org/2021.emnlp-main.595/.

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, May 2013. ISSN 1076-9757.

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)*, pp. 20349–20360, Los Alamitos, CA, USA, October 2023. IEEE Computer Society. doi: 10.1109/ICCV51070.2023.01866. URL https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.01866.

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 Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 730–749, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.41. URL https://aclanthology.org/2024.findings-acl.41/.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *ICLR (Poster)*, 2015. URL http://dblp.uni-trier.de/db/conf/iclr/iclr2015.html#KingmaB14.

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.

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

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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 366–384, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72673-6.

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)*, pp. 26296–26306, June 2024.

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*, pp. 291–309, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73026-9. doi: 10.1007/978-3-031-73027-6_17. URL https://doi.org/10.1007/978-3-031-73027-6_17.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, 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. URL https://arxiv.org/abs/2410.21276.

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*, ACL '02, pp. 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

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 Parrish, 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. URL https://doi.org/10.48550/arXiv.2408.00118.

Abdelkrim Saouabe, Said Tkatek, Merouane Mazar, and Imad Mourtaji. Evolution of image captioning models: An overview. In *2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM)*, pp. 1–5, 2023. doi: 10.1109/WINCOM59760.2023.10322923.

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*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

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. URL https://arxiv.org/abs/2406.10328.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From Show to Tell: A Survey on Deep Learning-Based Image Captioning . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 45(01):539–559, January 2023. ISSN 1939-3539. doi: 10.1109/TPAMI.2022.3148210. URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2022.3148210.

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.

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 Rakićević, 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 Kepa, 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. URL https://arxiv.org/abs/2403.05530.

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. URL https://arxiv.org/abs/2307.09288.

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. URL https://arxiv.org/abs/2502.14786.

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)*, pp. 26700–26709, June 2024.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pp. 4566–4575. IEEE Computer Society, 2015. ISBN 978-1-4673-6964-0. URL http://dblp.uni-trier.de/db/conf/cvpr/cvpr2015.html#VedantamZP15.

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*, 2024a.

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*, 2024b.

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. URL https://openreview.net/forum?id=Im2neAMlre.

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. URL https://arxiv.org/abs/2407.10671.

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 A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 1601–1619. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/056e8e9c8ca9929cb6cf198952bf1dbb-Paper-Conference.pdf.

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, 2024a. URL https://arxiv.org/abs/2408.04840.

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)*, pp. 13040–13051, June 2024b.

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. doi: 10.1162/tacl_a_00166. URL https://aclanthology.org/Q14-1006/.

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)*, pp. 11975–11986, October 2023.

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.