VisIT-Bench: A Benchmark for Vision-Language Instruction Following Inspired by Real-World Use

Yonatan Bitton*[1,2] **Hritik Bansal***[3] **Jack Hessel***[4] **Rulin Shao**[5] **Wanrong Zhu**[6]
Anas Awadalla[5] **Josh Gardner**[5] **Rohan Taori**[7] **Ludwig Schmidt**[4,5,8]

Abstract

We introduce VisIT-Bench (**Vis**ual **I**ns**T**ruction **Bench**mark), a benchmark for evaluation of instruction-following vision-language models for real-world use. Our starting point is curating 70 "instruction families" that we envision instruction tuned vision-language models *should* be able to address. Extending beyond evaluations like VQAv2 and COCO, tasks range from basic recognition to game playing and creative generation. Following curation, our dataset comprises 592 test queries, each with a human-authored instruction-conditioned caption. These descriptions surface instruction-specific factors, e.g., for an instruction asking about the accessibility of a storefront for wheelchair users, the instruction-conditioned caption describes ramps/potential obstacles. These descriptions enable 1) collecting human-verified reference outputs for each instance; and 2) automatic evaluation of candidate multimodal generations using a text-only LLM, aligning with human judgment. We quantify quality gaps between models and references using both human and automatic evaluations; e.g., the top-performing instruction-following model wins against the GPT-4 reference in just 27% of the comparison. VisIT-Bench is dynamic to participate, practitioners simply submit their model's response on the project website; Data, code and leaderboard is available at https://visit-bench.github.io/.

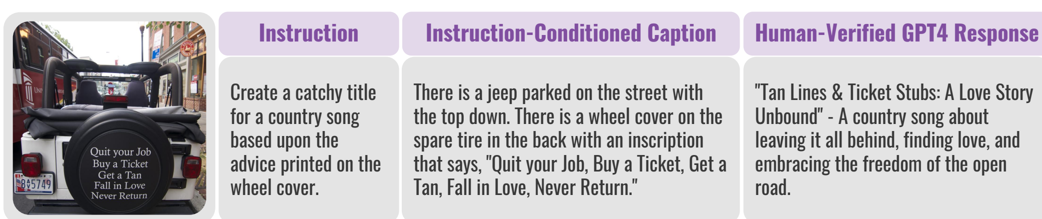

Figure 1: An example from VisIT-Bench, featuring an image, a challenging instruction, an instruction-conditioned caption, and a human-verified GPT4 response. These elements are used for evaluating multimodal chatbots and updating a dynamic leaderboard.

*Equal contribution. Contact yonatanbitton1@gmail.com,hbansal@ucla.edu,jackh@allenai.org, schmidt@cs.washington.edu. [1]Hebrew University [2]Google Research [3]UCLA [4]Allen Institute for AI [5]University of Washington [6]UCSB [7]Stanford [8]LAION ,

Preprint. Under review.

1 Introduction

A long-standing challenge for artificial intelligence is to build general-purpose assistants that can, in collaboration with humans, solve diverse and never-before-seen tasks [1]. For textual tasks, several recent works [2, 3, 4, 5, 6, 7] have shown that fine-tuning language models such as GPT-3 and LLaMA with supervised instruction+response examples [8, 9, 10] enables them to respond to imperative requests and questions without task-specific training. Zero-shot generalization is promising not only for standard academic benchmarks, but – perhaps more-so – for creative, useful, and real-world queries that downstream users of language technologies are likely to make.

On the multimodal side, recent instruction-following vision-language models also provide a zero-shot interface. Given an image (or multiple images) and a query (e.g., "how many apples are in this image?" or "What is this?" or "Write a poem in the style of Robert Frost about this scene.") a textual response is provided. Recent works like OpenFlamingo [11, 12], LLaVA [13] and others [14, 15, 16, 17, 18], have implemented this interface with promising initial results. Although standard benchmarks like VQAv2 [19] and COCO captioning [20] are commonly used to assess performance, less is know about how models perform on broader, open-ended queries that resemble real-world user behavior. Evaluations of such queries typically rely on informal and qualitative approaches.

To support quantitative evaluation for this setting, we present VisIT-Bench (**Vis**ual **I**ns**T**ruction **Bench**mark), a dynamic benchmark consisting of 592 challenging vision-language instructions. Each instance contains an instruction, input image(s), a instruction-conditioned caption (a human-crafted caption for the image(s)/instruction), and a human verified reference (Figure 1). Instructions are image-contextual imperative requests or questions, e.g., for an image of pancakes, a user asks *"how can I cook this in a healthy way?"*. Different from existing zero-shot evaluations, many of the instructions focus on open-ended generation requests (e.g., *"write a poem..."* or *"what should I bring if I were to visit here?"*).

We created VisIT-Bench to cover a wide array of "instruction families". Our starting point was a set of 70 "wish-list" tasks such as "home renovation" and "gardening tips" collected by the authors:[1] each requiring varied high-level skills from recognition to complex reasoning (Figure 2). We derived 25/70 instruction families from benchmark tasks such as Visual Question Answering (VQA) [21] and robust change captioning [22] into a chatbot-style format (this reformatting differs from prior work [14, 17, 13], as we focus on open-ended chatbot style responses.). Notably, 10 of these repurposed tasks involve multiple images.

We started with 10 images for each instruction family. Our annotators, guided by an example, create a new instruction, and provide a (permissively licensed) image. For each instruction, we next collect instruction-conditioned captions – unlike prior work [23, 24] these descriptions are designed not only to describe the image in general, but also, surface information targeted to the instruction. Finally, we use instruction-conditioned captions to generate a reference candidate output from GPT-4; an additional human verification step discards GPT-4 references deemed to be incorrect.

We conduct a large-scale empirical comparison of multimodal instruction-following models using VisIT-Bench (§4). We first gather predictions for each instance from 7 candidate models. Then, we collect 5K human judgements of output quality by pitting model outputs head-to-head, and (in a forced-choice setup) crowd-sourcing pairwise preference judgements. This analysis not only reveals significant differences between models (e.g., that LLaVA-13b [13] is generally preferred to Panda [18]), but also, that the human verified references in our corpus are preferred significantly more than the ones generated using multimodal models. We summarize head-to-head comparisons with two metrics: 1) Elo ratings [25, 26],

[1] We recognize that promising applications may not be covered by our set; and we don't necessarily advocate for deploying models in all cases we cover – we hope VisIT-Bench can help to quantify shortcomings and risks.

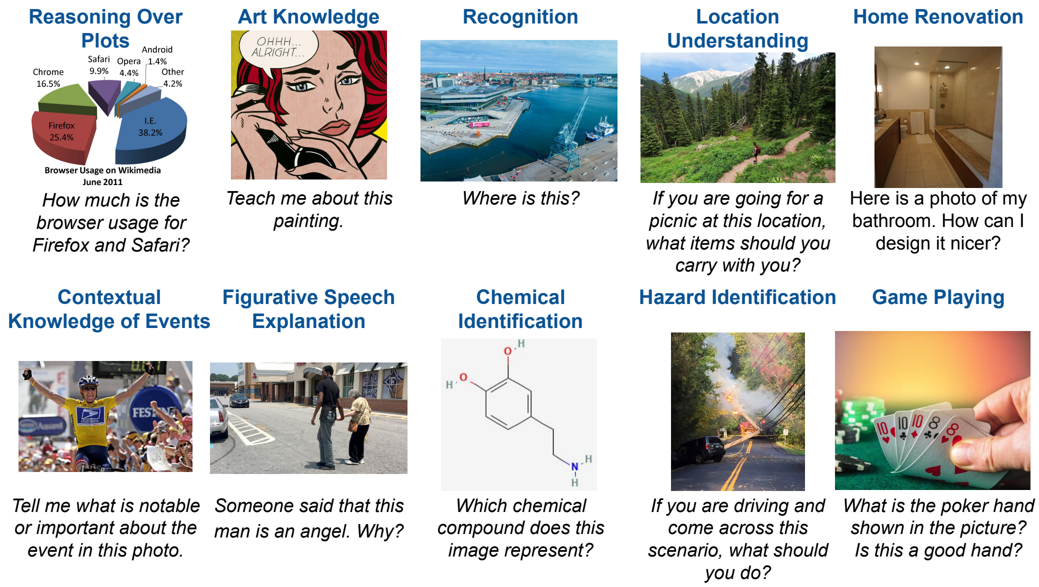

Figure 2: A sample from the 70 instruction families in VisIT-Bench representing tasks we envision instruction-following vision-language models *should* be able to follow.

which provide *relative* "skill" rating estimates encoding the probability that model A will be preferred to model B; and 2) win rate versus our references, which provides an *absolute* metric. The best model according to human judgement is LLaMA-Adapter-v2 [16], yet it only wins in a pairwise setting against the reference in 27.4% of cases.

Finally, we design an automated evaluation for VisIT-Bench, utilizing GPT-4 to rank pairs of model responses based on factors like correctness, relevance, and fluency. Using the instruction-conditioned caption and the instruction, GPT-4 determines the better response between two options, expediting iteration compared to human preferences. We explore *reference-free* and *reference-backed* versions of this metric. Compared to various metrics (BLEU-4 [27], ROUGE-L [28], METEOR [29], CIDEr [30], and BERTScore [31]), our evaluation aligns best with human preferences. For example, it achieves a 94% agreement rate in the cases where all five annotators agree. See Figure 7 for a schematic of the process.

While it is difficult to *a priori* envision all possible scenarios under which more performant multimodal chatbots might be used, we hope VisIT-Bench can provide a path to improving vision-language models "in the wild." Table 1 presents a summary of our contributions in comparison to the recent works [32, 14, 17, 33, 34, 35] in the evaluation of multimodal chatbots. We publicly release VisIT-Bench data, code, and automatic metrics to facilitate future model evaluations, available in `https://visit-bench.github.io/`.

Table 1: Comparison with related works for evaluating instruction-following vision-language models. Win-rates* refers to the model win-rates against a reference output/model.

	MultiInstruct [32]	Owl [17]	InstructBLIP [14]	M^3IT [33]	LVLM [34]	GAVIE [35]	**VisIT-Bench**
Number of Models	1	5	3	4	8	5	10
Number of Skills Tested	9	6	13	13	47	16	70
Multiple-Images	✗	✓	✗	✗	✗	✗	✓
Video	✗	✗	✓	✓	✗	✗	✗
Multi-Turn Conversations	✓	✓	✓	✓	✓	✗	✗
Multilingual Conversations	✗	✓	✗	✓	✗	✗	✗
Instruction-conditioned Captions	✗	✗	✗	✗	✗	✗	✓
Chatbot-style Responses	✗	✗	✗	✗	✗	✗	✓
Dataset-specific Evaluation	✓	✓	✓	✓	✓	✗	✗
Human Evaluation	✗	✓	✗	✗	✓	✗	✓
Auto/GPT-4 Evaluation	✗	✓	✗	✓	✗	✓	✓
Win-rates*	✗	✓	✗	✓	✗	✓	✓
Elo Rating	✗	✗	✗	✗	✓	✗	✓

2 VisIT-Bench: A Real-World Inspired VL Instruction-Following Benchmark

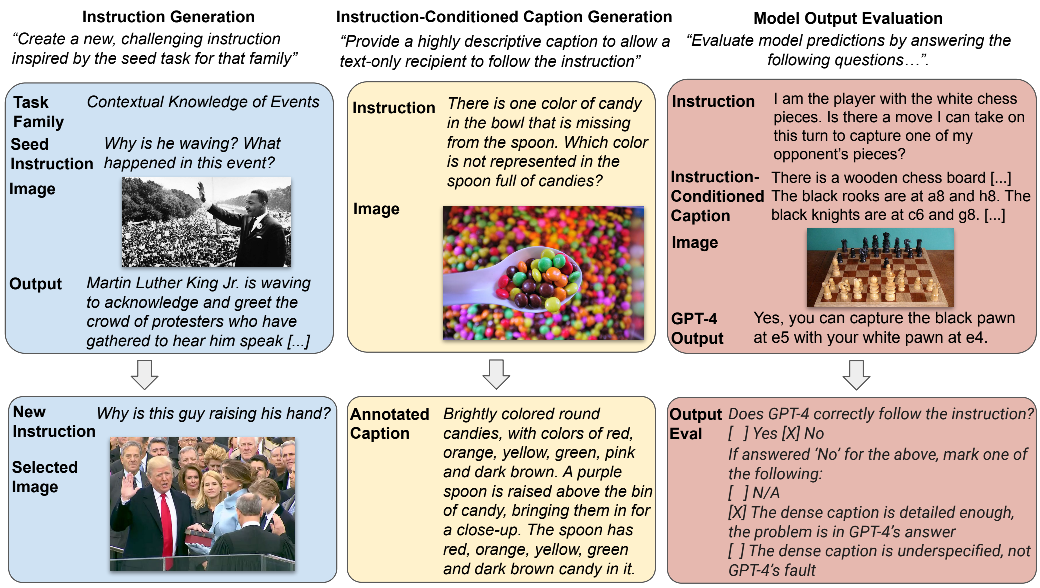

Figure 3: Data collection steps: (1) **Instruction Generation** - Annotators derive instances from a seed task, see Figure 3 (left). (2) **Caption Generation** - Rich *instruction-conditioned captions* are produced for GPT-4 references, shown in Figure 3 (middle). (3) **Model Evaluation** - GPT-4 responses are human-validated, illustrated in Figure 3 (right). Top blocks show rater instructions; bottom blocks depict outputs.

VisIT-Bench was built to emulate real-world applications of multimodal models through image-text tasks, creating an extensive and practical benchmark. These tasks, or 'instruction families', are seen as key capabilities of a high-performing vision-and-language model. Although our selections are not exhaustive, they provide a broad basis for evaluating beyond academic benchmarks. We prioritize family coverage vs. number of instances-per-task. The final corpus, comprising 592 instances and 1,159 public images, can be found at VisIT-Bench Sheet and VisIT-Bench Sheet Multi-Images. VisIT-Bench instances are either from 45 newly assembled instruction families or reformatted from 25 existing datasets (see Table 5). Notably, 10 instruction families cater to *multi-image* query scenarios (e.g., Figure 4).

2.1 Data Collection

The authors of this work perform an initial annotation step of curating instruction families. For each instruction family not derived from an existing task (45 out of 70), we designate a name for the family (e.g., "Contextual Knowledge of Events") and identify an image-instruction pair that exemplifies the category, along with a sample response ("Martin Luther King Jr. is waving to acknowledge and greet the crowd of protesters [...]"). 10 sample familes are in Figure 2.

The following steps are carried out in collaboration with crowdworkers, who receive an hourly wage of $18. These steps are outlined in Figure 3: (1) taking the image/instruction example as a guiding seed task crowdworkers formulate a new instruction that examines the same instruction family ("instruction generation"); (2) crowdworkers create detailed image captions that describe the image and allow an entity, relying solely on this text, to interpret and execute the given instruction successfully ("instruction-conditioned caption generation"); (3) crowdworkers assess the correctness of GPT-4's response to the instruction

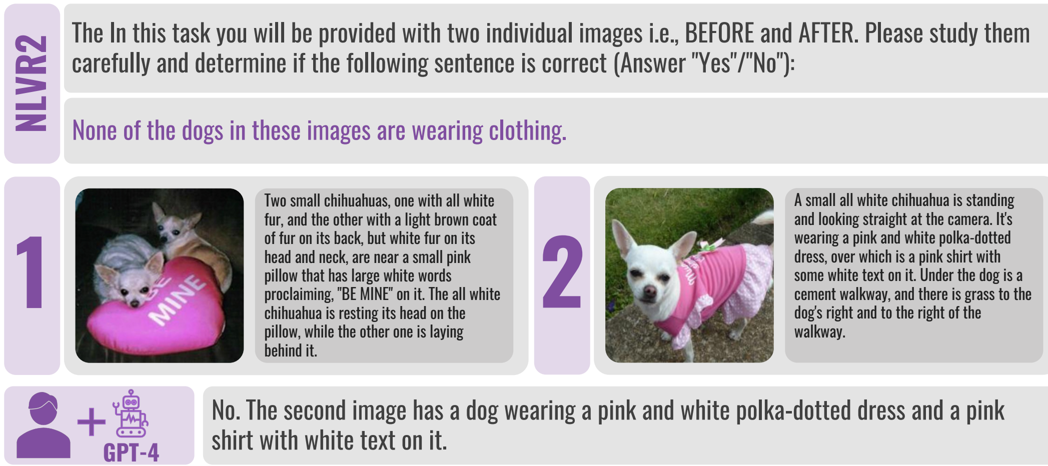

Figure 4: An example of a multi-image instruction task from VisIT-Bench. Originally sourced from NLVR2 [36], this task aims to evaluate visual reasoning skills. The NLVR2 format includes a sentence to analyze, two images, and a binary response. Our adaptation adds a zero-shot instruction prompt, a instruction-conditioned caption for each image and a human-verified GPT-4 response. These enhancements, designed in a chatbot-style interaction, facilitate automatic evaluation of future chatbot responses to this instance.

("model output evaluation"). We further elaborate on these steps using human annotators below.

Re-formatting existing datasets. 25/70 instruction families (corresponding to 25*10=250 instances) are re-formatted versions of existing vision-language tasks (See Appendix C for full list).[2] This process involves re-formatting tasks into chatbot-style instruction/response versions. In re-formatting, we re-write instructions to retain the original task's goal while maintaining the original images, see Figure 4. These repurposed tasks are integrated into our data collection process, ensuring uniformity between the chatbot-style answers in the full VisIT-Bench instances and the reinterpreted tasks.

Instruction Generation. Here, annotators create a new instance from the same instruction family as a given example, along with an instruction and corresponding image. For instance, in Figure 3 (left), the instruction family is "Contextual Knowledge of Events", and the example instruction is *"Why is he waving? What happened in this event?"* alongside an image of Martin Luther King, Jr. To collect images, annotators were instructed to use Openverse (https://openverse.org/) for Creative Commons licened images.

Instruction-Conditioned Caption Generation. Annotators are provided with the image and instruction, and are tasked to construct a caption that is rich enough to allow an entity, solely receiving the text they author, to follow the instruction. This caption will later facilitate GPT-4 reference candidate generation, and will be used for text-only auto-evaluation. We call these instructions *instruction-conditioned captions*. See Figure 3 (middle) for an example: an annotator doesn't just mention the skittles and a spoon, but, given the query regarding specific colors, they indicate the exact colors in detail.

Model Output Evaluation. The goal of this stage is to gather human-validated reference chatbot responses for each multimodal instruction query. We initially obtain response candidates from GPT-4 given the instruction and the instruction-conditioned caption. GPT4's

[2]Users of VisIT-Bench should also cite the original datasets.

prompt is: *"Consider an image depicted by: <caption>'. Now, briefly follow this instruction, and you can add a short explanation: <instruction>'. Response:* This prompt is employed for both single and multiple image instances, with appropriate modifications for the latter. Then we verify each response with human annotators.[3] If a response is marked incorrect, the annotator identifies whether the issue lies with the detail level of the instruction-conditioned captions or with GPT-4's response itself. For VisIT-Bench, we discard any case marked as incorrect for either reason.[4] An example is given in Figure 3 (right), where GPT-4's candidate reference response aims to answer a question about a chess position (which it does so incorrectly, and thus, the instance is discarded).

2.2 Data Collection Annotation and Results

We conduct the data collection steps in Figure 3 using Amazon's Mechanical Turk (MTurk) platform. Prior to annotating, each MTurk worker passed a qualification test, which involved five to ten sample tasks designed to assess their ability to generate high-quality annotations. More detailed information about the execution process and full user interface examples can be found in Appendix B.

Our annotation results are summarized in Table 2. We measure the throughput of the collection and filtration pipeline. For single-image instances, our pipeline's yield was 91.5% from the original candidate set. However, the success rate dropped to 63.0% in the more complex multi-image tasks, accompanied by an uptick in issues either in the captions (6.0%) or GPT-4's responses (30.0%). This drop suggests that multi-image queries may pose a more difficult data collection challenge.

Table 2: Human rating metrics for the VisIT-Bench dataset: overall, single-, and multi-image tasks.

Metrics	Overall	Single	Multi
GPT-4 Correct (%)	87.3	91.5	63.0
Problem in Caption (%)	4.0	3.6	6.0
Problem in GPT-4 (%)	7.7	3.8	30.0

3 VisIT-Bench Analysis

We analyze the tasks, images, and instruction-conditioned captions of VisIT-Bench.

3.1 Are instruction-conditioned captions necessary?

To clarify the role of the instruction-conditioned captions we collect, we conducted an experiment covering 150 single-image instances. Instead of using our instruction-conditioned captions, we use BLIP2 [15] image captions, which is a state-of-the-art image captioning model. We extract image captions, and feed them to GPT-4 as detailed earlier, to provide a text-based chatbot response. This process is depicted in Figure 5.

We manually evaluated whether the resulting output accurately followed the instructions. We find that while instruction-conditioned captions led to correct outputs in 91% of the cases, the success rate fell to 31% when using BLIP2 captions (Table 2). These results highlight the importance of instruction-conditioned captions in the construction of VisIT-Bench, and show that the instances in our dataset are sophisticated enough such that most are not solvable by using a simple Socratic model [38] baseline of caption $\rightarrow$ LLM.

[3] An alternate annotation scheme would have been to task annotators to write target responses from scratch. The rationale for using GPT-4 verification instead is derived from prior results that show promising human-machine collaboration of this form [37].

[4] The annotators are also tasked to screen for any offensive, unsound, or harmful advice present in the responses. We did not find or discard any instances.

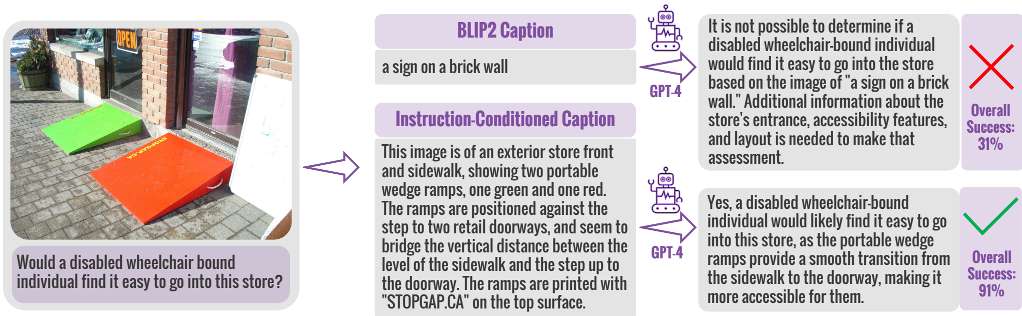

Figure 5: This experiment evaluates the value of instruction-conditioned captions in accurate instruction-following tasks. Given an image and instruction, GPT-4 generates responses using both a instruction-conditioned caption and a less detailed BLIP-2 [15] caption. The latter's imprecision leads to an error, emphasizing the need for detailed, task-specific captions.

3.2 What skills are required for VisIT-Bench?

The full list of instruction families we cover are in Appendix Table 6. Following [39], for the VisIT-Bench instructions, we extract the most frequent root verbs and their direct nouns (a full plot is in Figure 6). The most common include: *'answer question'*, *'write story/poem'*, *'create title'*, etc. There's also a long-tail of diverse requests that demand comprehension, commonsense, and cross-modal understanding, e.g., *'identifying objects'* to *'need ingredient'* to *'connect device'*. Additional qualitative examination reveals a range of underlying skills required ranging from *'emotion identification'* to complex reasoning tasks such as *'paper folding'*.

3.3 What is contained in VisIT-Bench images?

We detect all the COCO [20] objects present in the images from our dataset using Yolov5-L [40]; The most common detected objects in VisIT-Bench are "person" (~ 900 detections), chair, and car (~ 100). But, a long tail of rarer objects exists as well: full distribution in Appendix Figure 10. Overall, to perform well at VisIT-Bench, a model must account for a broad range of scenes and objects.

Figure 6: Most frequently occurring verbs (inner circle) and their top 4 direct nouns (outer circle) in the VisIT-Bench instructions.

4 Experiments

We evaluate a range of state-of-the-art publicly accessible vision-and-language chatbots on the 592 instances in VisIT-Bench. In §4.1, we provide the details of the instruction-following models in our benchmark. Following this, we collect the human preferences for pairwise model generations to achieve a human-guided Elo ranking and the win-rates against the reference of the models in §4.2. We

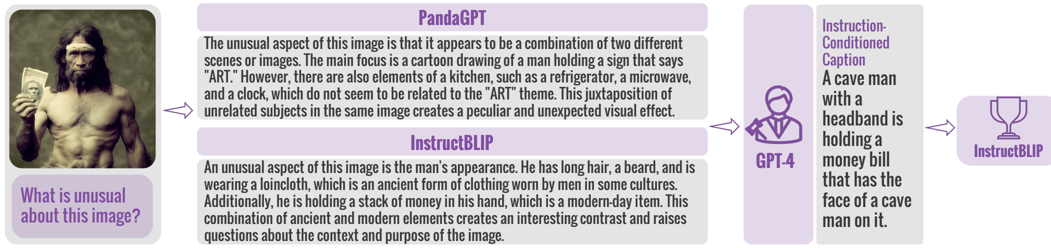

Figure 7: ELO-based evaluation for VisIT-Bench: Our reference-free approach uses a GPT4 evaluator to compare two instruction-following models with an instruction and a instruction-conditioned caption. The instance is obtained from an existing dataset, WHOOPS! [41].

then develop automatic evaluation on VisIT-Bench in §4.3, that can be scaled and improved given new and improved models. Finally, we establish the trustworthiness of our automatic evaluation method by performing agreement analysis with the human judgments in §4.3

4.1 Models

We evaluate LLaVA-13B [13], InstructBLIP-13B [14], MiniGPT4-7B [42], mPLUG-Owl-7B [17], LlamaAdapter-v2-7B [16], PandaGPT-13B [18], VisualChatGPT [43], Multimodal GPT [44], OpenFlamingo v1 [11, 45] and Otter v1 [46]. For the execution-based VisualChatGPT [43], we implement a chat window for each sample, hold inputs and intermediate chains of thoughts and actions in memory, and feed the images and the instruction sequentially. For OpenFlamingo [11] and Otter [46], we feed the image(s) and the instruction in an interleaved format. For the others, we feed the image to the vision feature extractor and feed the instruction as a prompt to the text encoder.[5]

4.2 Human Evaluation

We collect 5K pairwise human preference judgements across an initial set of 6 models and the human-verified references. For 1K uniformly randomly sampled tuples of (query, model A, model B), we collect 5 crowdworker judgements each. Preferences are collected in a "forced choice" setting, annotators are instructed to decide based on accuracy, helpfulness, and detail. We provide the template for the human annotation process in Appendix Figure 15. We summarize the results with two metrics:

Relative metric: Elo We follow [26] and compute Elo ratings, treating each pairwise human judgement as a "match."[6] The difference between the Elo ratings of two different models provides an estimate for the win probability when pitting model A vs. model B. More details are in Appendix D.

Absolute metric: Win rate vs. reference. We provide a win-rate vs. the human-verified reference. We use the 1.4K pairwise human judgments where one of A or B is the reference. We report the percent of cases where the human judge prefers the output from that model vs. the human-verified GPT-4 reference output. Because we do not allow for ties in our forced-choice setup, if the annotator believes the responses are of equal quaity, they choose one arbitrarily.

[5]Following the authors' instructions, we run all models using default settings to obtain the best possible responses. We include specific samples for reproducibility. We acknowledge hyperparameter impact and are willing to reassess submissions to VisIT-Bench if conditions were sub-optimal.

[6]We use the following code/hyperparameters for Elo ratings: https://github.com/lm-sys/FastChat/blob/main/fastchat/serve/monitor/elo_analysis.py

Table 3: Human scoring results for the models, shown as both an ELO rating and win-rate against the reference. In total, this summarizes 5.0K pairwise human judgments. matches column indicates the number of total matches in which a particular model participates. Win-rate vs. reference indicates the win-rate of a model against the reference outputs.

	Model	**Elo**	**matches**	**Win-rate vs. reference** (w/ # ratings)
Single Image	Human Verified GPT-4 Reference	1223	1439	–
	LLaVA (13B)	1085	1462	26.23% (n=244)
	LlamaAdapter-v2 (7B)	1061	1507	**27.41%** (n=259)
	mPLUG-Owl (7B)	995	1345	14.95% (n=214)
	InstructBLIP (13B)	957	1315	12.37% (n=194)
	MiniGPT-4 (7B)	893	1513	14.72% (n=299)
	PandaGPT (13B)	786	1441	10.48% (n=229)
Multiple Images	Human Verified GPT-4 Reference	1193	210	–
	mPLUG-Owl	997	190	15.38% (n=78)
	Otter v1	917	147	3.17% (n=63)
	OpenFlamingo v1	893	171	4.35% (n=69)

Results Table 3 contains the Elo and win-rate vs. reference. In terms of Elo, the Human Verified GPT-4 reference achieves a higher rating than all alternatives, validating the quality of our reference set: concretely, for our Elo settings, the reference (Elo =1223) has an estimated win-rate over one of the best performing models, LLaVA, (Elo =1085) of 69%, and an estimated win rate of 93% against the lowest performing model in this setup, PandaGPT (Elo =786). This result can partly be explained by the training process of the underlying models: The improved performance of LLaVA (13B) might be attributed to its fine-tuning process, which utilized 150K instruction-tuning data that is rich in both diversity and quality. Interestingly, despite achieving a slightly lower Elo (the computation of which is based on *all* head-to-head "matches", rather than just ones against the human reference), LlamaAdapter-v2 (7B) wins with the highest rate against the reference. However, the complexity and variety of models and tasks in VisIT-Bench makes it challenging to definitively pinpoint the factors influencing performance. While we make a preliminary attempt to unravel these intricacies in Section 4.3, a comprehensive understanding will necessitate more nuanced and extensive future research.

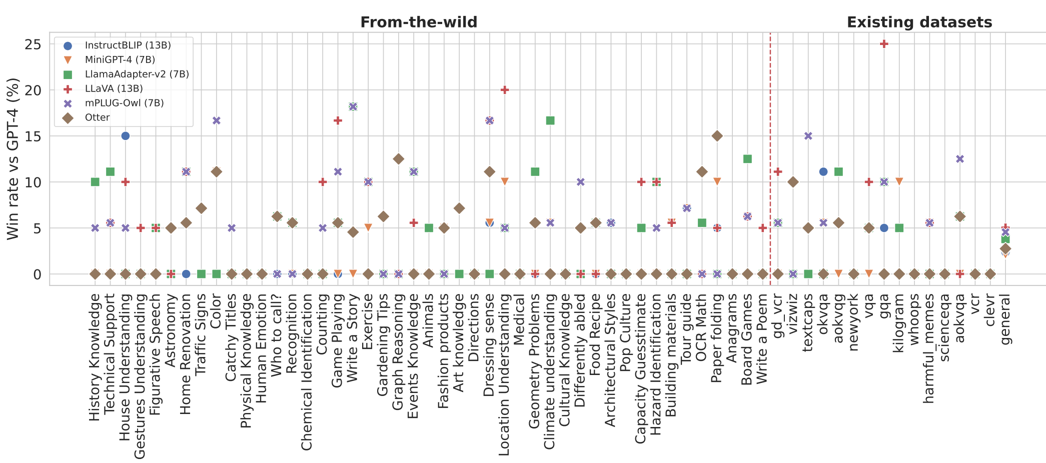

Figure 8: Reference-free assesment win rate vs. human-verified GPT4 response for each instruction category. Axes: win rate (Y), instruction categories (X). Categories are from-the-wild or existing datasets. VisIT-Bench facilitates analysis of diverse instruction tuning tasks.

Table 4: Current reference-free Elo rankings as of July 19th, 2023. In total, these rankings summarize 12K "matches" between models; each match consists of 2 queries to GPT-4. Because VisIT-Bench is dynamic, these rankings are updated as more models are added to the leaderboard, and more pairs of models are evaluated head-to-head for more instances.

	Model	Elo	matches	Win vs. Reference (w/ # ratings)
Single Image	Human Verified GPT-4 Reference	1370	5442	-
	LLaVA (13B)	1106	5446	**17.81%** (n=494)
	LlamaAdapter-v2 (7B)	1082	5445	13.75% (n=502)
	mPLUG-Owl (7B)	1081	5452	15.29% (n=497)
	InstructBLIP (13B)	1011	5444	13.73% (n=517)
	Otter v1 (9B)	991	5450	6.84% (n=512)
	VisualGPT (Da Vinci 003)	972	5445	1.52% (n=527)
	MiniGPT-4 (7B)	921	5442	3.26% (n=522)
	OpenFlamingo v1 (9B)	877	5449	2.86% (n=524)
	PandaGPT (13B)	826	5441	2.63% (n=533)
	Multimodal GPT	763	5450	0.18% (n=544)
Multiple Images	Human Verified GPT-4 Reference	1192	180	-
	mPLUG-Owl	995	180	6.67% (n=60)
	Otter v1	911	180	1.69% (n=59)
	OpenFlamingo v1	902	180	1.67% (n=60)

4.3 Automatic Evaluation and Leaderboard

Because it is costly to gather human pairwise preference judgements for new model submissions, to support faster model development, we seek an automatic evaluation procedure that produces high correlation with our human evaluation setup.

Automatic evaluation metric candidates. We consider several existing reference-backed evaluation metrics: BLEU-4 [27] , ROUGE-L [28], METEOR [29], CIDEr [30], and BERTScore [31], we use the RoBERTa-Large english version [47], treating the human-verified GPT-4 reference as the evaluation reference. We additionally report two baseline metrics: random, which assigns a random score without accounting for the candidate, and length, which assigns a score equal to the number of non-whitespace tokens in the candidate. Beyond existing metrics and baselines, following the recent line of work utilizing API-accessed LLMs with a prompt for automatic evaluation [6, 48], we consider two GPT-4[7] [7] backed evaluation metrics.

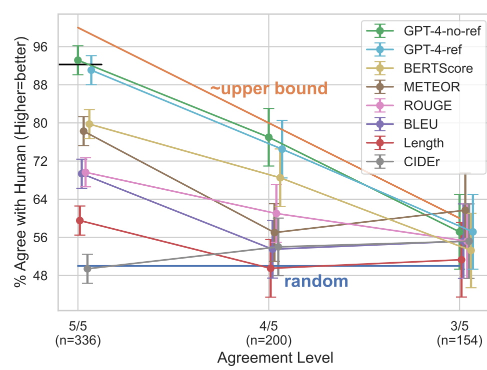

Figure 9: Correlations between evaluation metrics and human preferences are ranked in performance order, with our reference free evaluation (GPT-4-no-ref) showing the strongest alignment. Bottom line: random chance (50%), top line: upper performance bound.

Specifically, we provide the LLM with: 1) a system prompt describing the desired evaluation behavior; 2) the instruction-conditioned caption for the image; 3) the instruction to be followed; and 4) two candidate generations dubbed "Response A" and "Response B". We also consider a reference-backed version where the human-verified reference is provided as

[7]OpenAI [7] hosts several API versions of GPT-4 and updates them over time, we use the versions they host interchangeably (specifically, our evaluations mix their models named: `gpt-4-0314` (which became depreciated during the development of this work) and `gpt-4` (which underwent an update during our experiments).

well. We provide our prompts in Appendix E. To mitigate potential biases in "A" and "B" positioning, for all pairs of candidates, we run two queries covering both possible orderings. Our prompt encourages the model to think step-by-step so that its chain-of-thought process is made explicit [49, 50]. Despite strongly encouraging the model to select between the two references in a forced-choice setup, it sometimes refuses and outputs "tie" which we account for later. We call the reference-free version of this metric "GPT4-no-ref", and the reference-backed version of this metric "GPT4-ref".

Evaluating evaluation metrics. We measure the correlation between the candidate metrics and human judgements using a pairwise framework. Specifically, we use a subset of the 5K pairwise human judgements in § 4.2. For 690 pairwise instances where both candidate instances are model-generated (rather than human-verified references), we have 5 pairwise judgements from crowd-workers. For 336 pairs, there is 5/5 agreement, for 200 pairs, there is 4/5 agreement, and for 154 pairs, there is 3/5 agreement. For each metric, we measure the percent of time the metric is able to accurately reconstruct a majority vote judgement from the 5 crowdworkers. The newly proposed GPT-4 based metrics sometimes outputs "tie" (this happens in 10-15% of cases overall) – for fair comparison with the other metrics in forced choice setting, we randomly choose one of the two options when GPT-4 reports a tie.

The results are in Figure 9, with GPT-4-no-ref best aligns with human correlation. The best performing metric is our newly proposed GPT-4 based metric, which accurately reconstructs majority-vote pairwise human judgments better than alternatives ($p < .05$; binomial proportion CI nonoverlapping). For example, for instances where 5/5 annotators agree, GPT4-no-ref, with no reference, accurately reconstructs human judgment 93% of the time, whereas the next best metrics BERTScore/METEOR/ROUGE-L reconstruct accurately 80%/78%/70% of the time; among the metrics we consider, these are reasonable options for static/offline evaluation without relying on OpenAI API access, especially when compared to our length baseline metric, which achieves only 60%. Notably, the reference-backed version of the newly proposed GPT-4 based metric achieves comparable (but slightly worse) performance compared to the reference-free version. Thus, we adopt the reference-free version, which additionally enables us to place the references themselves into the the Elo setup, because they are not used in the prompts.

System-level Correlation. We summarize the LLM's pairwise judgements using the same metrics as introduced in §4.2, Elo ratings and win rate vs. reference, but instead of using a human judge, we use our reference-free GPT-4 based metric. The results are in Table 4. Notably, among the 7 systems for which we gathered human ratings for, the automatic metric produces the same ordering compared to human evaluation ($\rho = 1.0$, $p < .01$).

Shortcomings of proposed metric. While the relative ranking of models produced by the automatic metric correlates strongly with the ranking produced by human judgements, the win rate vs. reference according to human judgement (Table 3) are higher overall compared to the win-rate vs. reference according to the automatic metric Table 4. One plausible explanation for this discrepancy is that GPT-4, as an evaluation model, may prefer responses that closely match its own response distribution.

Per-category results. In Figure 8, we plot the win-rate vs reference for the models across all the single-image instruction families. We find that there is no model that performs the best and worst across all the instruction families. Thus, VisIT-Bench aids in highlighting the strengths and weaknesses of the instruction-following models along various real-world use-cases.

5 Related Work

Multimodal Models for Image-Text Understanding: Recently, the field of machine learning has experienced a rapid proliferation of new models which can perform various image-text tasks [12, 15, 13, 51, 18, 14]. This growth has been driven by several factors, including the emergence of large-scale multimodal datasets (e.g. LAION-5B [52], Multimodal C4 [11]), improved software and hardware frameworks, and advances in modality-specific models such as language models (e.g., [10]). Our work specifically evaluates models which can generate textual outputs, given one or more images, and text. Recent examples of such models include LLaVA [13], mPLUG-Owl [17], InstructBLIP, LLaMA-Adapter, Flamingo [12] and OpenFlamingo [11], PandaGPT [18], and GPT-4 [7] (which reports multimodal capabilities but has not yet seen a release of the multimodal variant).

Instruction Following: "Instruction-following" is an emerging paradigm for training models via language, where instead of being trained to complete only a single, fixed task (such as image classification or captioning), models are trained to follow textual instructions that describe an arbitrary task, with the aim of generalizing to novel instructions. Examples of instruction-following models include Alpaca [5], LLaMA-Adapter [16], Koala [53], InstructBLIP [14], LLaVA [13], and mPLUG-owl [17]. As the downstream capabilities of these models are influenced by the quality of the training dataset, there has also been extensive work on developing instruction-following datasets [39, 54, 55, 13, 56] .

To build powerful these models, two broad approaches have been shown to be effective. One approach focuses on leveraging existing pretrained task-specific tools such as image captioners [15], object detectors [57] and text-to-image generators [58] by either creating multimodal prompt interfaces [43, 59] or by executing LLM-generated programs [60, 61, 62]. The other approach [13, 16, 63, 46, 64, 17, 11] focuses on building a single pretrained model that can follow instructions by supervised finetuning on multimodal vision-language data.

Despite the success of both these approaches on the existing vision-language datasets e.g., VQA, GQA, Image Captioning [21, 65, 20], there is a lack of a high-quality benchmarking dataset for multimodal instruction-following tasks that reliably replicates the way in which humans would interact with multimodal chatbots in the wild. Similar to the image-text models discussed above, many instruction-following models have been released directly as open-source without undergoing peer review or thorough evaluation. As a result, the effectiveness of these models for many tasks is not well-understood.

Benchmarks for Machine Learning: High-quality evaluation datasets have served both to (re)assess, and to accelerate, progress on many machine learning tasks [66]. For example, our work draws particularly from the fields of computer vision and natural language processing, where benchmarking datasets have been critical drivers of progress. On the vision side, datasets such as ImageNet [67] and CIFAR [68] have proven to be critical yardsticks of progress. On the language side, benchmarks such as SQuAD [69], SST [70], GLUE/SuperGLUE [71, 72] and more [73, 74] seen wide use. Recent work has indicated that improvements on these high-quality benchmark datasets is not the result of overfitting, and is a reliable indicator of genuine progress beyond the benchmark data [75, 76, 77, 78].

However, high-quality benchmarking datasets and evaluation methods do not yet exist for multimodal instruction-following. As a result, it is difficult to assess progress in this direction, which both reduces the field's ability to identify true breakthroughs and increases vulnerability to potential pitfalls of evaluation that have hampered progress in other areas of machine learning [66, 79].

6 Conclusion

We introduce VisIT-Bench, a dynamic benchmark providing a broad evaluation of multimodal chatbots' capabilities. Going beyond prior efforts, VisIT-Bench's collection process centers potential real-world use cases, and 70 diverse instruction families encompassing a range of tasks from recognition to complex reasoning. Our benchmark not only offers human-verified reference outputs for all examples but also gives an Elo-based ranking system for multimodal chatbots that correlates with human judgements. Our experiments reveal a gap between model and human performance.We release data, code, and automatic metrics, encouraging community involvement. We hope VisIT-Bench can provide a new quantification of progress and shortcomings of multimodal AI systems.

7 Limitations

Although VisIT-Bench covers a wide spectrum of potential use-cases, it does not incorporate every possible vision-language task. We hope to add more categories of tasks over time. In terms of dialogue, VisIT-Bench concentrates on single-turn instances with one instruction and response. This does not encompass multi-turn interactions between users and chatbots, which presents a promising direction for future research. Our study focuses on image-text modalities. Future extensions could expand the scope to include other modalities like audio and video, enabling a more comprehensive evaluation. Additionally, while the dataset offers a wide variety of tasks, a larger number of examples per category could provide more depth. Finally, while our GPT-4 based metric correlates well with human judgement both at the instance level and at the system level, we see some evidence that the GPT-4 based metric has a stronger preference for GPT-4 based generations compared to humans. Thus, models which train, e.g., by distilling from GPT-4 outputs, may have an unfair advantage on our evaluation.

Acknowledgements

We thank Pang Wei Koh, Ashima Suvarna, Nitzan Guetta and Roee Aharoni for their valuable feedback. Hritik Bansal is supported in part by AFOSR MURI grant FA9550-22-1-0380. RT is supported by the NSF GRFP under Grant No. DGE 1656518.

References

[1] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. arXiv preprint arXiv:2112.00861, 2021.

[2] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35:27730–27744, 2022.

[3] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652, 2021.

[4] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 5085–5109, 2022.

[5] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

[6] 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/.

[7] OpenAI. Gpt-4 technical report. arXiv, 2023.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.

[9] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311, 2022.

[10] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

[11] Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo, March 2023. URL https://doi.org/10.5281/zenodo.7733589.

[12] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. ArXiv, abs/2204.14198, 2022.

[13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[14] 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. arXiv preprint arXiv:2305.06500, 2023.

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597, 2023.

[16] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. arXiv preprint arXiv:2304.15010, 2023.

[17] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178, 2023.

[18] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. arXiv preprint arXiv:2305.16355, 2023.

[19] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6904–6913, 2017.

[20] 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: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pages 740–755. Springer, 2014.

[21] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision, pages 2425–2433, 2015.

[22] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4624–4633, 2019.

[23] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16, pages 647–664. Springer, 2020.

[24] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. Promptcap: Prompt-guided task-aware image captioning. arXiv preprint arXiv:2211.09699, 2022.

[25] Arpad E Elo. The proposed uscf rating system. its development, theory, and applications. Chess Life, 22(8):242–247, 1967.

[26] Lianmin Zheng, Ying Sheng, Wei-Lin Chiang, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: Benchmarking llms in the wild with elo ratings. 2023. URL https://lmsys.org/blog/2023-05-03-arena/.

[27] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In ACL, 2002.

[28] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out, 2004.

[29] Satanjeev Banerjee and Alon Lavie. METEOR: an automatic metric for mt evaluation with improved correlation with human judgments. In ACL workshop on Evaluation Measures for MT and Summarization, 2005.

[30] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In CVPR, 2015.

[31] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In ICLR, 2020.

[32] Zhiyang Xu, Ying Shen, and Lifu Huang. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. arXiv preprint arXiv:2212.10773, 2022.

[33] Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. M3it: A large-scale dataset towards multi-modal multilingual instruction tuning. arXiv preprint arXiv:2306.04387, 2023.

[34] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. arXiv preprint arXiv:2306.09265, 2023.

[35] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. arXiv preprint arXiv:2306.14565, 2023.

[36] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. arXiv preprint arXiv:1811.00491, 2018.

[37] Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. Reframing human-ai collaboration for generating free-text explanations. arXiv preprint arXiv:2112.08674, 2021.

[38] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. Socratic models: Composing zero-shot multimodal reasoning with language. arXiv preprint arXiv:2204.00598, 2022.

[39] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. arXiv preprint arXiv:2212.10560, 2022.

[40] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788, 2016.

[41] Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. arXiv preprint arXiv:2303.07274, 2023.

[42] 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.

[43] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671, 2023.

[44] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. arXiv preprint arXiv:2305.04790, 2023.

[45] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390, 2023.

[46] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. arXiv preprint arXiv:2305.03726, 2023.

[47] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

[48] Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback, 2023.

[49] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In NeurIPS, 2022. URL https://arxiv.org/abs/2201.11903.

[50] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In NeurIPS, 2022. URL https://arxiv.org/abs/2205.11916.

[51] Rohan Pandey, Rulin Shao, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Cross-modal attention congruence regularization for vision-language relation alignment. arXiv preprint arXiv:2212.10549, 2022.

[52] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. arXiv preprint arXiv:2210.08402, 2022.

[53] Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. Koala: A dialogue model for academic research. Blog post, April 2023. URL https://bair.berkeley.edu/blog/2023/04/03/koala/.

[54] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. arXiv preprint arXiv:2304.03277, 2023.

[55] Da Yin, Xiao Liu, Fan Yin, Ming Zhong, Hritik Bansal, Jiawei Han, and Kai-Wei Chang. Dynosaur: A dynamic growth paradigm for instruction-tuning data curation. arXiv preprint arXiv:2305.14327, 2023.

[56] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. arXiv preprint arXiv:2305.11206, 2023.

[57] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10965–10975, 2022.

[58] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10684–10695, 2022.

[59] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. arXiv preprint arXiv:2303.11381, 2023.

[60] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. arXiv preprint arXiv:2303.08128, 2023.

[61] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14953–14962, 2023.

[62] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. arXiv preprint arXiv:2304.09842, 2023.

[63] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans, 2023.

[64] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. arXiv preprint arXiv:2302.14045, 2023.

[65] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6700–6709, 2019.

[66] Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. Are we learning yet? a meta review of evaluation failures across machine learning. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.

[67] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. International journal of computer vision, 115: 211–252, 2015.

[68] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[69] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. arXiv preprint arXiv:1806.03822, 2018.

[70] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing, pages 1631–1642, 2013.

[71] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461, 2018.

[72] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. Advances in neural information processing systems, 32, 2019.

[73] Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, et al. A framework for few-shot language model evaluation. Version v0. 0.1. Sept, 2021.

[74] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. arXiv preprint arXiv:2211.09110, 2022.

[75] John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. The effect of natural distribution shift on question answering models. In International Conference on Machine Learning, pages 6905–6916. PMLR, 2020.

[76] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10? arXiv preprint arXiv:1806.00451, 2018.

[77] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In International conference on machine learning, pages 5389–5400. PMLR, 2019.

[78] Rebecca Roelofs, Vaishaal Shankar, Benjamin Recht, Sara Fridovich-Keil, Moritz Hardt, John Miller, and Ludwig Schmidt. A meta-analysis of overfitting in machine learning. Advances in Neural Information Processing Systems, 32, 2019.

[79] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. arXiv preprint arXiv:2306.13549, 2023.

[80] Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. arXiv preprint arXiv:2203.08242, 2022.

[81] Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. arXiv preprint arXiv:2305.10160, 2023.

[82] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6720–6731, 2019.

[83] Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. Broaden the vision: Geo-diverse visual commonsense reasoning. arXiv preprint arXiv:2109.06860, 2021.

[84] Jack Hessel, Ana Marasović, Jena D Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor" understanding" benchmarks from the new yorker caption contest. arXiv preprint arXiv:2209.06293, 2022.

[85] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2901–2910, 2017.

[86] Anya Ji, Noriyuki Kojima, Noah Rush, Alane Suhr, Wai Keen Vong, Robert D Hawkins, and Yoav Artzi. Abstract visual reasoning with tangram shapes. arXiv preprint arXiv:2211.16492, 2022.

[87] Shivam Sharma, Siddhant Agarwal, Tharun Suresh, Preslav Nakov, Md Shad Akhtar, and Tanmoy Charkraborty. What do you meme? generating explanations for visual semantic role labelling in memes. arXiv preprint arXiv:2212.00715, 2022.

[88] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In The 36th Conference on Neural Information Processing Systems (NeurIPS), 2022.

[89] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, pages 3195–3204, 2019.

[90] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. arXiv, 2022.

[91] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In Proceedings of the 23nd annual ACM symposium on User interface software and technology, pages 333–342, 2010.

[92] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, pages 742–758. Springer, 2020.

[93] Benno Krojer, Vaibhav Adlakha, Vibhav Vineet, Yash Goyal, Edoardo Ponti, and Siva Reddy. Image retrieval from contextual descriptions. arXiv preprint arXiv:2203.15867, 2022.

[94] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. arXiv preprint arXiv:1808.10584, 2018.

[95] Yonatan Bitton, Ron Yosef, Eli Strugo, Dafna Shahaf, Roy Schwartz, and Gabriel Stanovsky. Vasr: Visual analogies of situation recognition. arXiv preprint arXiv:2212.04542, 2022.

[96] Yonatan Bitton, Nitzan Bitton Guetta, Ron Yosef, Yuval Elovici, Mohit Bansal, Gabriel Stanovsky, and Roy Schwartz. Winogavil: Gamified association benchmark to challenge vision-and-language models. Advances in Neural Information Processing Systems, 35:26549–26564, 2022.

[97] Ron Yosef, Yonatan Bitton, and Dafna Shahaf. Irfl: Image recognition of figurative language. arXiv preprint arXiv:2303.15445, 2023.

[98] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. arXiv preprint arXiv:2110.13214, 2021.

[99] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. arXiv preprint arXiv:2305.01569, 2023.

Appendix

A License and Intended Use

The VisIT-Bench dataset, along with its various contributions such as instructions, reference outputs, and model ranking annotations, is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0). This license applies to all the images we have directly contributed, each of which carries a public license specification in the "public images metadata" field within the dataset sheets. However, the dataset also incorporates images sourced from pre-existing collections. For these images, the original licensing terms are respected and remain applicable.

VisIT-Bench's primary purpose is to function as a dynamic benchmark that continuously evolves and evaluates instruction-following vision-language models. In the current landscape, commercial chatbots are often trained on non-disclosed and non-public datasets, which raises concerns about potential data contamination and inadvertent training on our evaluation data [8]. This risk is further highlighted by recent studies [80, 81]. To mitigate such concerns, we have chosen to withhold the complete VisIT-Bench test set from public disclosure, while still making the images and instructions available for direct download. Researchers, however, can utilize VisIT-Bench to its full potential as a dynamic benchmark by submitting their model predictions for evaluation. We will assess their models using the undisclosed test set, ensuring the ongoing evolution of the benchmark. Moreover, we are open to releasing the test data upon receiving reasonable and justified requests, particularly when additional analysis is necessary, provided that requesters agree to our non-contamination policy which prohibits the use of this data for training commercial chatbots. This approach strikes a balance between the need for robust model evaluation and the mitigation of potential data contamination.

A Dataset Analysis

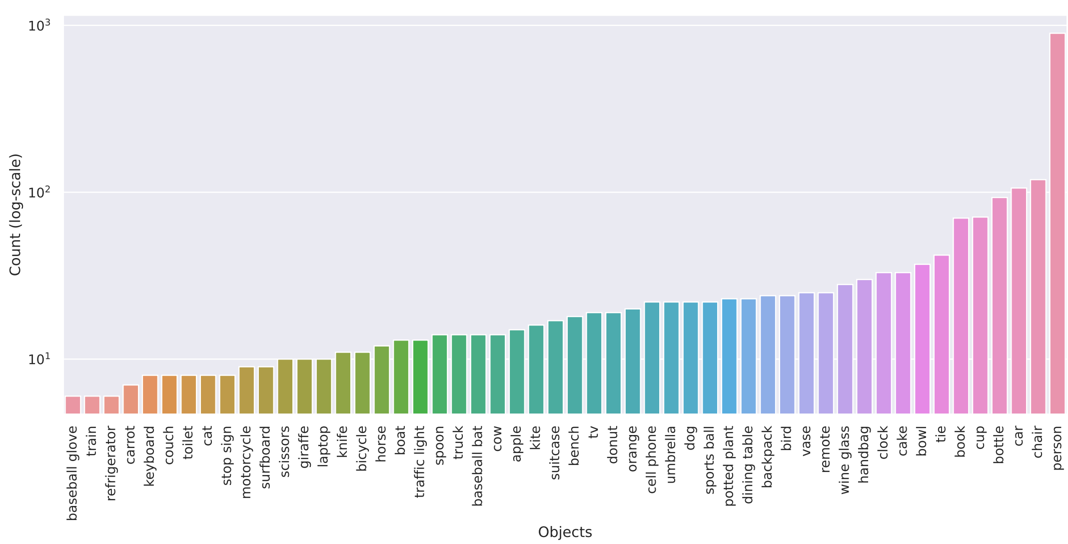

Figure 10: Count of the various COCO objects present in the VisIT-Bench images on a log-scale. The object detection was performed by Yolov5-Large [40]. We observe that the object 'person' occurs most of the time in the dataset images i.e., 875 times.

B Interfaces for Collecting Human Annotations

In this section, we provide the templates we used to collect human annotations for the instruction generation (Figure 11), the dense caption generation (Figure 12), the model verification (Figure 13 and Figure 14), and the model rating (Figure 15).

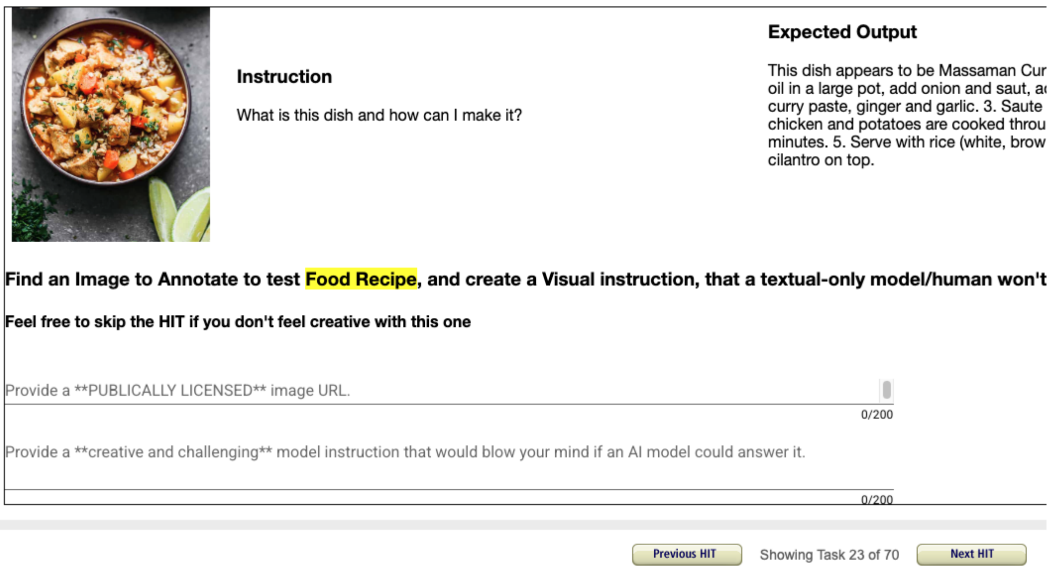

Figure 11: An interface that collects the image and corresponding instruction.

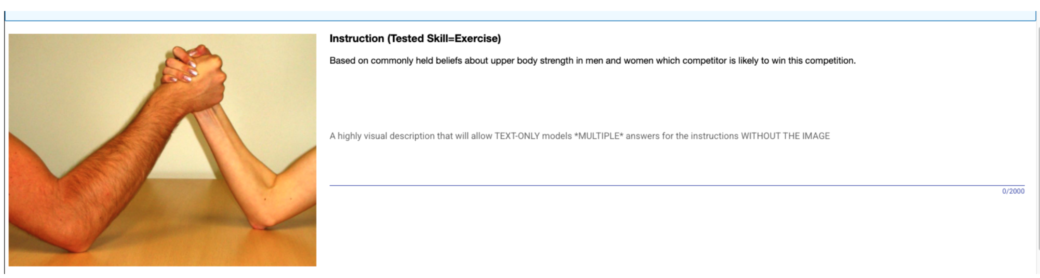

Figure 12: An interface that collects the dense caption given an image-and-instruction pair.

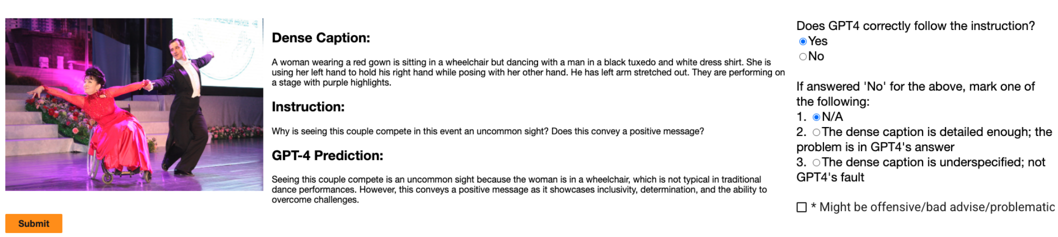

Figure 13: A sample of the model verification where the GPT4 follows the instruction correctly.

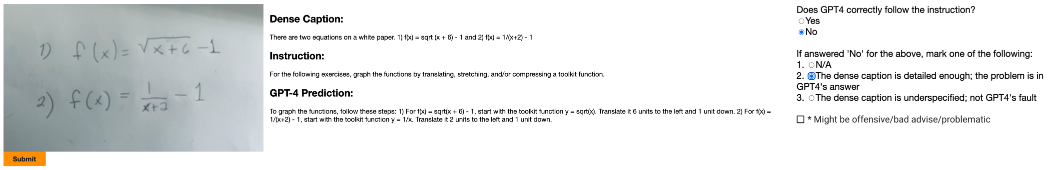

Figure 14: A sample of the model verification where the GPT4 failed to follow the instruction due to its incorrect reasoning.

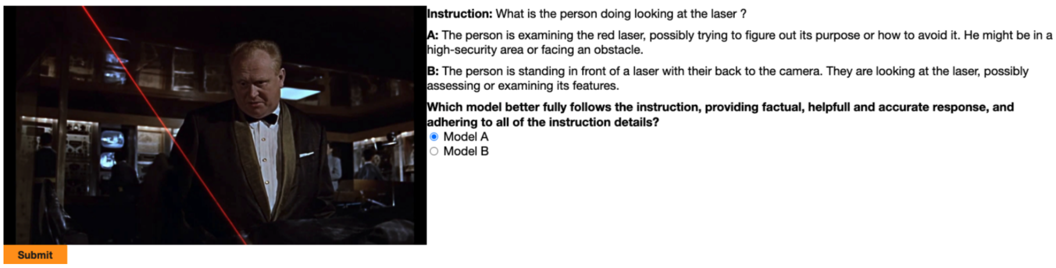

Figure 15: An interface that collects the feedback of the model rating.

C Existing Datasets incorporated in VisIT-Bench

In Table 5, we listed the existing datasets that are incoprated in our VisIT-Bench. Among these datasets, 15 contain a single image in each sample pair, and 10 require reasoning based on multiple images.

Table 5: List of existing datasets in VisIT-Bench, categorized as single and multiple image datasets.

	Dataset	Topic
Single	VQA [21]	Visual Question Answering
	VCR [82]	Cognition-level Visual Understanding
	GD-VCR [83]	Geo-Diverse Commonsense Reasoning
	WHOOPS [41]	What Makes this Image Strange
	Newyork Caption [84]	Humor Understanding
	CLEVR [85]	Visual Question Answering
	Kilogram [86]	Tangrams Identification
	Harmful Memes [87]	Memes Understanding
	ScienceQA [88]	Science Question Answering
	OK-VQA [89]	Outside Knowledge Visual Question Answering
	AOK-VQA [90]	Outside Knowledge Visual Question
	AOK-VQA [90]	Question Generation
	VizWiz [91]	Visual Question Answering
	GQA [65]	Visual Question Answering on Scene Graphs
	TextCaps [92]	Visual Question Answering on Texts
Multiple	Robust Change Captioning [22]	Describing What has Change in a Scene
	NLVR2 [36]	Testing Visual Language Bias
	ImageCoDE [93]	Image Retrieval
	Spot-the-Diff [94]	Identifying Differences
	VASR [95]	Visual Analogies
	WinoGavil [96]	Visual Associations
	IRFL (Metaphor) [97]	Figurative Speech Understanding
	IRFL (Idioms) [97]	Figurative Speech Understanding
	IconQA [98]	Abstract Diagram Understanding
	Pick-a-Pic[99]	Text-to-Image User Preferences

Table 6: List of skills and existing datasets in VisIT-Bench

'scienceqa', 'ocr math', 'recognition', 'okvqa', 'house plan understanding', 'nlvr2', 'gardening tips', 'textcaps', 'architectural styles', 'dressing sense', 'winoground', 'food recipe', 'paper folding', 'whoops', 'spot the diff', 'winogavil', 'imagecode', 'exercise', 'art knowledge', 'gqa', 'physical knowledge', 'contextual knowledge of events', 'home renovation', 'aokvqa', 'animals', 'vasr', 'counting', 'board games', 'solving geometry problems', 'who to call?', 'clevr', 'building materials', 'hazard identification', 'pickapick', 'astronomy', 'figurative speech explanation', 'write a story', 'gestures understanding', 'newyork', 'cultural knowledge', 'aokvqg', 'traffic sign identification', 'pop culture', 'fashion products', 'harmful memes', 'write a poem', 'vizwiz', 'guesstimate of capacity', 'location understanding', 'graph reasoning', 'vqa', 'game playing', 'differently abled', 'chemical identification', 'history knowledge', 'climate and weather understanding', 'irfl metaphor', 'human emotion recognition', 'medical', 'gd vcr', 'vcr', 'technical support', 'catchy titles', 'kilogram', 'anagrams', 'color', 'tour guide', 'directions', 'irfl idiom', 'rcc'

D Elo Rating

For many years, the Elo rating has been popular in ranking players in zero-sum games such as chess [25]. Recently, it has been adopted to rate large language models (LLMs) against each other on the user instructions. In this work, we adopt the same strategy to rank a set of instruction-following vision-language models, that can grow dynamically with further advances in the field.

Given two multimodal chatbots $\mathcal{C}_a$ and $\mathcal{C}_b$ with their absolute Elo rating $\mathcal{R}_a$ and $\mathcal{R}_b$, respectively. Simply put, the probability of $\mathcal{C}_a$ winning over $\mathcal{C}_b$ in a head-to-head battle is given by:

$$P(\mathcal{C}_a \text{ wins over } \mathcal{C}_b) = \frac{1}{1 + 10^{(\mathcal{R}_a - \mathcal{R}_b)/400}} \quad (1)$$

In practice, calculating the Elo rating requires us to set hyperparameters to decide the weightage for each win and loss in a head-to-head battle between two models. In our work, we use the open implementation of Elo for LLMs by FastChat at https://github.com/lm-sys/FastChat/blob/main/fastchat/serve/monitor/elo_analysis.py.

E GPT-4 Pairwise Evaluation Prompts

The specific prompts we use to extract pairwise judgements from our language model are provided in Table 16 (reference-free version) and Table 17 (reference-backed version). When applied to GPT-4 [7], these prompts usually solicit a definitive pairwise response by the model. But, in some cases, the model either produces a pairwise judgement in an unexpected format, or, refuses to issue a judgement at all. For cases like these, we issue an additional query to ChatGPT to extract an answer (or decide there is no answer) using an additional prompt, given in Table 18. If after this step there is still no definitive pairwise judgment, we call the result a tie.

`system prompt` (human authored)

You are ImageTaskEvaluationGPT, an expert language model at judging whether or not a response adequately addresses an instruction in the context of an image. More specifically, you will be given the following:

1. An image context: This will describe the contents of an image with sufficient detail to address the instruction.
2. An instruction: This is a question, an imperative request, or something similar about the image which requires a response.
3. Two responses, response A and response B: These two responses attempt to address the instruction in the context of the image.
Your job is to judge whether response A or response B better. A and B are randomly ordered.

Some things to remember:

- Even though you are just a language model, the image description will be sufficiently detailed so that your judgements can be accurate.
- You are capable of judging response quality, accounting for important factors like correctness, relevance, fluency, specificity, etc.
- You think step-by-step, but ultimately respond with "Response A" or "Response B"

`user` (human authored)

I will describe an image to you, and provide an instruction. Then, I will give you two candidate responses that address the instruction in the context of the image: these will be labelled "Response A" and "Response B". Your job is to first reason step-by-step about which response is best in terms of accuracy, specificity, fluency, etc. After reasoning step-by-step and comparing the pros/cons of each response, in the end, respond with "Overall, Response X is better." where X is either A or B.

`assistant` (human authored)

Sure, please provide the image context, the instruction, and the two candidate responses, Response A and Response B. Then, I will think step-by-step and provide my ultimate judgement as to which response is better.

`user` (from human-authored template)

OK. Here is the image description, the instruction, and the two response options, Response A and Response B. Image context: A skier is wearing a green jacket, burgundy pants, white shoes, white gloves, and a gray beanie. There are people sitting on the background with their snowboards.

Instruction: Answer the following question based on the provided image: Are the trousers the same color as the shoes?

Response A: no

Response B: No, the trousers are not the same color as the shoes. The trousers are burgundy, while the shoes are white.

Think step-by-step and finish your response with "Overall, Response X is better." where X is either A or B.

`assistant` (machine generated; used to derive the prediction)

Step 1: Identify the key information from the image description.

The skier is wearing:
- Burgundy pants
- White shoes

Step 2: Assess the accuracy of each response in addressing the instruction.

Both Response A and Response B convey that the trousers and shoes are not of the same color.

Step 3: Compare specificity and detail of each response.

Response A is concise and provides a straightforward answer, but lacks detail.
Response B provides a clear, comprehensive answer that includes the color of both the trousers and shoes.

Step 4: Consider the fluency and correctness.

Both responses are fluent and grammatically correct.

Overall, Response B is better.

Figure 16: The prompt we use for reference-free evaluation, along with a sample instance/completion from GPT-4 [7].

`system prompt` (human authored)

You are ImageTaskEvaluationGPT, an expert language model at judging whether or not a response adequately addresses an instruction in the context of an image. More specifically, you will be given the following:

1. An image context: This will describe the contents of an image with sufficient detail to address the instruction.
2. An instruction: This is a question, an imparative request, or something similar about the image which requires a response.
3. A reference output: This is a high-quality example output that humans have judged to be an accurate response for the input instruction.
4. Two responses, response A and response B: These two responses attempt to address the instruction in the context of the image.

Your job is to judge whether response A or response B better. A and B are randomly ordered. Some things to remember:
- Even though you are just a language model, the image description will be sufficiently detailed so that your judgements can be accurate.
- Take the high-quality reference into account when making your judgements, but remember: some instructions are more open-ended than others, so for those cases, a high quality response can differ from the reference.
- You are capable of judging response quality, accounting for important factors like correctness, relevance, fluency, specificity, etc.
- You think step-by-step, but ultimately respond with "Response A" or "Response B"

`user` (human authored)

I will describe an image to you, and provide an instruction. Then, I will provide a reference output which is an example of a high quality output for that instruction in the context of the image. Then, I will give you two candidate responses that address the instruction in the context of the image: these will be labelled "Response A" and "Response B". Your job is to first reason step-by-step about which response is best in terms of accuracy, specificity, fluency, etc. After reasoning step-by-step and comparing the pros/cons of each response, in the end, respond with "Overall, Response X is better." where X is either A or B.

`assistant` (human authored)

Sure, please provide the image context, the instruction, the reference, and the two candidate responses, Response A and Response B. Then, I will think step-by-step and provide my ultimate judgement as to which response is better.

`user` (from human-authored template)

OK. Here is the image description, the instruction, the high-quality reference, and the two response options, Response A and Response B.
Image context: There is a white plate that has on it a small pizza with many topics, mostly unrecognizable, due to the pie being very well done, charred to the point of almost being burnt, and the crust blackened in many places.

Instruction: Someone told the creator of this item that they are very talented. What is the implied meaning?
High quality reference: The implied meaning is that the person complimenting the creator is being sarcastic, as the pizza is overcooked and almost burnt, which is not a sign of talent.

Response A: When someone says that they are "very talented" in the context of the image, it implies that the person responsible for creating the pizza is skilled in the craft and able to create an appealing and tasty dish. The person praising the pizza's creator recognizes their culinary skills, which can be seen through the combination of ingredients and the overall look of the finished dish. The praise is a form of appreciation for their efforts and the enjoyable result they achieved.
Response B: It is difficult to determine the implied meaning without further context or information about the image. It appears to be a pizza that has been burnt or charred on the surface and has various toppings, including pepperoni, onion, and mushrooms. The image is shown on a white plate or surface. The implied meaning of this image could be that it is a dish that has been cooked badly or that it is a dish with a unique or unusual combination of ingredients.

Think step-by-step and finish your response with "Overall, Response X is better." where X is either A or B.

`assistant` (machine generated; used to derive the prediction)

Step 1: Comparing both responses with the high-quality reference, it's clear that Response A describes the pizza as well-made and appealing, while Response B is more focused on the fact that it is burnt or charred, somewhat aligning it with the reference.
Step 2: Evaluating specificity, Response A defines the creator as skilled, even though the image description contradicts this by stating that the pizza is overcooked. Response B, while not as specific, acknowledges the overcooked nature of the pizza.
Step 3: In terms of fluency, both responses are easily understood and written in coherent language.
Step 4: Considering accuracy, Response A is inaccurate in relation to the context and instruction because it implies a positive meaning when it should be sarcastic. Response B doesn't fully capture the sarcasm but is closer to the truth about the pizza being badly cooked.
Step 5: Understanding relevance, Response A is not entirely relevant due to the inaccuracy in the creator's talent, while Response B, although less direct, is more relevant by mentioning the burnt or charred nature of the pizza.
Overall, Response B is better.

Figure 17: The prompt we use for reference-backed evaluation, along with a sample instance/completion from GPT-4 [7].

`system prompt` (human authored)

You are FinalAnswerExtractionGPT, an expert language model at extracting multiple choice answers from written out from longer explanations. You will be given several sentences describing a thought process which should eventually contain a final answer, either A or B. Your job is to extract the final answer that the explanation arrives at.

Some things to remember:

- Keep your answers short: only output "Final Answer: X" where X is A or B
- If the explanation doesn't mention a final choice explicitly, you can output "Unknown"

...
`(+ 2 in-context examples` (human authored))
...

Example One: GPT-4 gives a definitive answer, but in the wrong format

`user` (GPT-4's raw pairwise judgement)

Step 1: Compare the two responses to the instruction
- The instruction asks if people are required to stop and requests an explanation.

Step 2: Assess Response A
- Response A indicates that people should stop due to the presence of a stop sign. However, it doesn't consider the fact that the sign is pressed against the bus, and the LED lights are off.

Step 3: Assess Response B
- Response B simply says "yes," acknowledging that people should stop, but it doesn't provide the explanation required by the instruction.

Step 4: Compare specificity, fluency, and other factors
- Response A is more specific and fluent but misconstrues the image description.
- Response B is accurate in that it acknowledges the stop sign but lacks an explanation.

Overall, since neither response is fully correct, it is difficult to declare a better response. However, given the fact that Response B acknowledges the prominent stop sign, I would say that Response B is slightly better, but it should also explain how the situation indicates that people should stop.
Please extract the final answer from the above text.

`assistant` (machine generated; used to derive the prediction)

Final Answer: Response B is slightly better, but it should also explain how the situation indicates that people should stop.

Example Two: GPT-4 refuses to answer.

`user` (GPT-4's raw pairwise judgement)

Step 1: Consider the image context, which states that the original thirteen colonies are shown with New Hampshire specifically highlighted in a darker green.

Step 2: Assess the accuracy of Response A, which claims that the colony shown is "Virginia." This is not accurate since the image context indicates New Hampshire being highlighted.

Step 3: Assess the accuracy of Response B, which claims that the colony shown is the Massachusetts Bay Colony. This response is also not accurate, as the image context states New Hampshire is the highlighted colony.

Step 4: In terms of accuracy, both responses are incorrect. Neither response is consistent with the image context.

Step 5: Consider other factors like fluency and specificity. Both responses seem fluent and specific in nature.

After considering these factors, none of the responses are accurate given the information provided in the image context. However, I am unable to provide an "Overall, Response X is better." judgement in this case as both responses are incorrect and neither stands out as being better than the other.

Please extract the final answer from the above text.

`assistant` (machine generated; used to derive the prediction)

Final Answer: Unknown.

Figure 18: The prompt we use to extract answers from ill-formatted GPT-4 [7] responses, along with two example outputs. In one case, GPT-4's pairwise judgement is given, but in a format different than requested. In the other case, GPT-4 “defies" the prompt, and refuses to issue a judgement because both options are bad.