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

Yonatan Bitton^{*1} Hritik Bansal^{*2} Jack Hessel^{*3} Rulin Shao⁴ Wanrong Zhu⁵
Anas Awadalla⁴ Josh Gardner⁴ Rohan Taori⁶ Ludwig Schimdt^{3,4,7}

Abstract

Recent advances in instruction-following vision-language models have led to a surge in large-scale multimodal chatbots. However, existing works lack a comprehensive evaluation of these models' capabilities to understand human instructions and provide useful, fluent, and safe responses. We introduce VisIT-Bench, a robust benchmark for diverse real-life vision-language instructions across 70 tasks, from recognition to reasoning. VisIT-Bench offers an in-depth understanding of a model's conversational abilities. Our dataset uniquely includes *instruction-conditioned caption* for all the images, facilitating the generation of *reference outputs* from a large language model (GPT-4) that are further verified by humans. For the model evaluation, we collect the reviewer's preference between the pairwise predictions from the set of candidate models (including reference output). These preferences are further used to assign an Elo rating and a Win-rate vs reference output (human-verified GPT-4 response) to the candidate models. While utilizing humans as reviewers, we find that the best-performing model wins against the reference output in just 27% (Win-rate vs GPT-4) of the comparisons, highlighting the complexity of the tasks and the scope for further development in this area. Finally, we design an *automatic preference framework* that demonstrates strong agreement with the human preferences as a reviewer, demonstrating its reliability for large-scale evaluations. VisIT-Bench is dynamic and can integrate and evaluate new models. We make the VisIT-Bench benchmark publicly available, including data, code, automatic metrics, and the human annotation data at <https://github.com/mlfoundations/VisIT-Bench>.

^{*}Equal contribution. Contact yonatanbitton1@gmail.com, hbansal@ucla.edu, jackh@allenai.org.
¹Hebrew University ²UCLA ³Allen Institute for AI ⁴University of Washington ⁵UCSB ⁶Stanford ⁷LAION

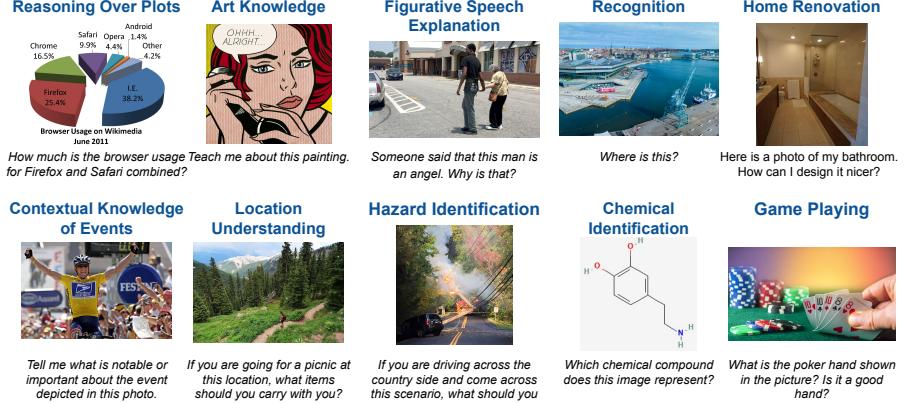


Figure 1: A sample from the 70 instruction families in VisIT-Bench.

23 1 Introduction

24 A long-standing challenge for artificial intelligence is to build general-purpose assistants that can
 25 understand human intent to solve real-world tasks [1]. Several recent works [2, 3, 4, 5, 6, 7] have
 26 shown that fine-tuning with human instructions enables large language models (LLMs) such as GPT-3
 27 and LLaMA [8, 9, 10] to solve new tasks ‘in the wild’ without any further task-specific training. Since
 28 humans interact with their surroundings through various modalities, there has been growing interest
 29 in developing *multimodal* assistants [11, 12, 13, 14, 15, 7, 16] that can understand information from
 30 different modalities, particularly vision and language.

31 Several multimodal instruction-following models have been proposed (described in Section 5), with
 32 promising preliminary results. However, their zero-shot instruction-following capabilities have
 33 mostly been evaluated informally, e.g., through qualitative comparison of outputs. As a result, it is
 34 difficult to quantify the progress and pitfalls of multimodal instruction-tuned models. To bridge this
 35 gap, we present VisIT-Bench, a dynamic benchmark consisting of 679 challenging vision-language
 36 instructions. Each instruction consists of an image paired with an imperative request or question,
 37 e.g., for an image of pancakes, a user asks “*how can I cook this in a healthy way?*”. Different from
 38 existing zero-shot evaluations, many of the instructions focus on open-ended generation requests
 39 (e.g., “*write a poem...*” or “*what should I bring if I were to visit here?*”). We collected VisIT-Bench
 40 in a structured manner, with the goal of spanning a diverse set of “instruction families”. Our starting
 41 point was a set of 70 “wish-list” tasks such as ‘home renovation’ and ‘gardening tips’ collected by
 42 the authors: these different families require different high-level skills ranging from recognition and
 43 open-world knowledge to complex reasoning tasks (Figure 1).

44 Every instruction family includes 10 images, resulting in 679 instances. Each instance contains
 45 an instruction, input image(s), and a *instruction-conditioned caption*: a distinctive human-crafted
 46 caption for the image(s). These captions, unlike prior work [17], are meticulously designed to adhere
 47 to the instruction, providing sufficient detail for text-only models to follow, thus paving the way for
 48 a text-only evaluation approach for multi-modal chatbots. Specifically, the *instruction-conditioned*
 49 *caption* allows us to generate a reference output from a text-only large language model like GPT-4 that
 50 is further verified by humans for its correctness. For more specifics on VisIT-Bench’s construction,
 51 refer to §2 and Figure 2. Further, VisIT-Bench innovatively transforms tasks such as Visual Question
 52 Answering (VQA) [18] and robust change captioning [19] into chatbot-style representation. For
 53 instance, the Visual Commonsense Reasoning (VCR) task is reframed as a zero-shot text generation
 54 task, utilizing the task description and input image. Importantly, our ground-truth responses mimic
 55 chatbot interactions, differing from the short answers typical in previous studies [12, 16, 11]. This
 56 advancement enables a more effective and rigorous evaluation of chatbots.

57 We conduct a large-scale empirical comparison of multimodal instruction-tuned models using VisIT-
 58 Bench (§4). We first gather the predictions for each instance from the candidate models. Then, we

59 perform preference selection between the pairwise predictions of the candidate models (including
60 reference output from GPT-4) in head-to-head battles. Subsequently, these preference selections are
61 summarized to assign an Elo-rating [20] and a *Win-rate vs human-verified GPT-4* response to the
62 candidate models. In our work, we first employ human labelers to select their preferences from the
63 pairwise predictions (§4.2). We find that the best-performing model wins against the reference output
64 in just 27% (Win-rate vs GPT-4) of the comparisons, highlighting the complexity of the tasks and the
65 scope for further development in this area.

66 Since human preferences are time-consuming and unscalable, we propose an *automatic preference*
67 *framework* that utilizes GPT-4 to automatically select a preference between the pairwise predictions.
68 Specifically, we create two variants of this framework: *reference-free* and *reference-backed*. In the
69 *reference-free* variant, GPT-4 is just provided with the *instruction-conditioned caption* for the test
70 instance whereas we also provide human-verified GPT-4 reference in the *reference-backed* variant
71 of the framework. In both cases, the framework has access to a detailed human-curated image
72 description (though the models being judged only have access to the image content), anticipating
73 its judgments on criteria such as correctness, relevance, fluency, specificity, etc., to align well with
74 human judgment. In §4, we present an analysis of results from both approaches, revealing a slight
75 edge for the *reference-backed* method, albeit not significantly so. Following [20], we summarize the
76 pairwise judgments using Elo ratings. See Figure 5 for a schematic of the process.

77 We validate our evaluation method by collecting a set of 2.4K pairwise judgements from humans and
78 observe that the model ranking induced by human judgements closely matches the model ranking
79 induced by the proposed automatic approach (see §4.2, and an agreement analysis in §4.4). Our
80 experiments, which cover several recent instruction-following visual language models, quantify
81 performance gaps that prior evaluation methods miss. While it is difficult to *a priori* envision all of
82 the possible scenarios under which more performant multimodal chatbots might be used, we hope
83 the corpus and evaluation supported by VisIT-Bench can provide a path to improving instruction-
84 following vision-language models ‘in the wild’. Table 1 presents a summary of our contributions in
85 comparison to the recent works [21, 12, 16, 22, 23] in the evaluation of multimodal chatbots.

86 VisIT-Bench operates dynamically, allowing updates to the text-only evaluation engine (currently
87 using models such as GPT3.5 and GPT4) with the potential for future integration of improved models
88 as they emerge. This setup facilitates continuous expansion of instruction families for evaluating a
89 diverse range of vision-and-language models, especially those emerging in real-life scenarios. We
90 publicly release VisIT-Bench data, code, and automatic metrics to facilitate future model evaluations.

91 In summary, our main contributions are as follows:

92 **VisIT-Bench Benchmark:** Introducing VisIT-Bench, an open, unique, and timely dataset for evaluating
93 instruction-following vision-language models on a set of challenging and diverse test examples
94 that aim to simulate human multimodal chatbot interactions, accessible in the following link: [VisIT-
95 Bench Sheet](#).

96 **Empirical Study:** Using VisIT-Bench, we generate the model outputs on the examples in our dataset
97 for a battery of instruction-following multimodal models, and make them publicly available. We
98 perform an in-depth analysis of the dataset diversity and provide an analysis on the model strengths
99 and weaknesses across different skills.

100 **Automatic Evaluation Framework for Multimodal Chatbots:** Establishing a verified automatic
101 evaluation method, utilizing the Elo system, for dynamic assessment of multimodal chatbots’ perfor-
102 mance in instruction-following tasks, fostering an ongoing and reliable leaderboard.

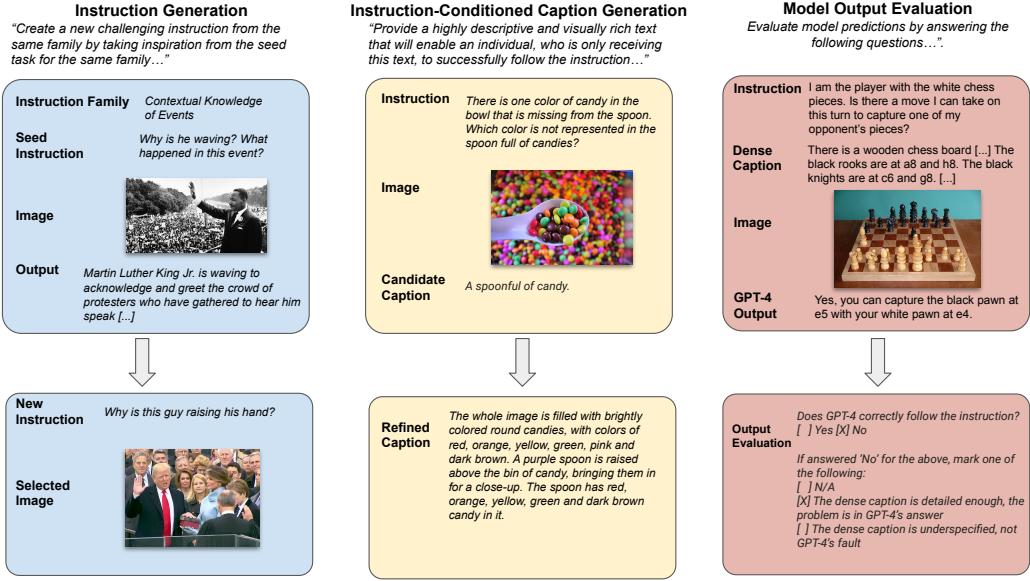


Figure 2: Human-in-the-loop data collection steps. For each step, the instructions to raters and inputs provided to raters are shown in the top block; the outputs from raters are shown in the bottom block. The data collection process progresses from left to right.

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

	MultiInstruct [21]	Owl [16]	InstructBLIP [12]	M ³ IT [22]	LVLM-eHUB [23]	VisIT-Bench (Ours)
Number of Models	1	5	3	4	8	10
Number of Skills Tested	9	6	13	13	47	70
Multiple-Images	X	✓	X	X	X	✓
Video	X	X	✓	✓	X	X
Multi-Turn Conversations	✓	✓	✓	✓	✓	X
Multilingual Conversations	X	✓	X	✓	X	X
Instruction-conditioned Caption	X	X	X	X	X	✓
Chatbot-style Reference Output	X	X	X	X	X	✓
Dataset-specific Evaluation	✓	✓	✓	✓	✓	X
Human Evaluation	X	✓	X	X	✓	✓
Auto/GPT-4 Evaluation	X	✓	X	✓	X	✓
Win-rates*	X	✓	X	✓	X	✓
Elo Rating	X	X	X	X	✓	✓

103 2 VisIT-Bench: A Comprehensive V&L Instruction-Tuning Benchmark

104 Our main motivation in constructing VisIT-Bench was to construct a benchmark composed of images
 105 and text instructions spanning a set of high-quality, challenging, and diverse tasks. Unlike traditional
 106 question-oriented approaches that often involve repetitive questioning patterns, we sought to develop
 107 instructions that mirror real-world usage of multimodal models. To this end, we began by formulating
 108 a set of tasks we envision as practical uses of instruction-tuned vision-and-language models. Our
 109 emphasis on diversity over depth was driven by the need for comprehensive testing across different
 110 problem categories, a critical factor for evaluating the robustness and generalization abilities of these
 111 models. The data includes overall 679 instances and 1,578 images and is published in its entirety and
 112 can be accessed at [VisIT-Bench Sheet](#) and [VisIT-Bench Sheet Multi-Images](#).

113 Creating the image-text components of VisIT-Bench involved a two-fold approach: we gathered
 114 new data and repurposed existing datasets. In the data we collected, instances are grouped into
 115 specific “skills” classified into what we call instruction families. We further expanded the breadth
 116 of our newly-collected dataset by incorporating 25 established datasets, which provided additional
 117 instruction families (Table 6). Notably, 10 of these are multi-image datasets.

118 A distinct feature of VisIT-Bench is the incorporation of chatbot-style responses as the benchmarking
119 standard. This technique enhances the relevance of VisIT-Bench to practical instruction-following
120 models. Existing datasets, detailed in the appendix, are reinterpreted, retaining their essence while
121 reframing tasks into instruction-following format, complete with corresponding images. fig. 4 illus-
122 trates this process. Incorporating 10 samples from each existing dataset, VisIT-Bench encompasses a
123 total of 250 instances. These repurposed tasks are subsequently integrated into our data collection
124 process, ensuring uniformity between the chatbot-style answers in VisIT-Bench and the reinterpreted
125 tasks. Through this methodology, we harness the valuable insights from previous studies, tailoring
126 them to a contemporary context and providing a comprehensive instrument for the evaluation of
127 instruction-following models.

128 2.1 Data Collection Framework

129 Our data-gathering process unfolds in four successive steps. (1) In the first step, we (the authors of
130 this paper) manually annotate a set of 70 seed “instruction families”. For each instruction family,
131 we generate a name (e.g. “Contextual Knowledge of Events”), an image-instruction pair that are
132 designed to elicit a specific form of reasoning, and an example output (“Martin Luther King Jr. is
133 waving to acknowledge and greet the crowd of protesters [...]”). We show a sample of 10 seed tasks
134 from our dataset in Figure 1. The following three steps involve a set of human raters. These steps
135 are summarized in Figure 2, and are designed to (2) propagate the seed examples to generate more
136 image-instruction pairs (“instruction generation”), (3) create *instruction-conditioned caption*: rich
137 captions that describe the image and allow an entity, relying solely on this text, to interpret and
138 execute the given instruction successfully (“instruction-conditioned caption generation”), and (4)
139 assess the correctness of GPT-4’s response to the instruction (“model output evaluation”). We further
140 elaborate on the three steps using human annotators below.

141 **Instruction Generation.** This step allows us to propagate a single example from an instruction
142 family into several high-quality examples. In this step, human annotators begin with a reference
143 from an instruction family, comprised of an instruction family, instruction, image, and exemplar
144 model output. Their task is to create a new instance from the same instruction family, along with
145 an instruction and corresponding image. For instance, in Figure 2 (left), the instruction family
146 under consideration is “Contextual Knowledge of Events”, and the example instruction is “*Why is*
147 *he waving? What happened in this event?*” alongside an image of Martin Luther King, Jr. The
148 annotators then devise a new instruction inspired by the seed task from the same family. They also
149 provide a publicly licensed image URL.¹

150 **Instruction-Conditioned Caption Generation.** This step creates high-quality textual captions for
151 each image in the dataset. Annotators are provided with an instruction and an image and are tasked
152 to construct a caption that is rich enough to allow an entity, solely receiving the text, to follow the
153 instruction. This caption will later facilitate GPT-4 output generation and will be used for text-only
154 auto-evaluation. See Figure 2 (middle) for an example, where the annotator creates a dense caption,
155 indicating the exact colors of skittle candies, the required information for solving the instruction.

156 **Model Output Evaluation.** The primary aim of this stage is to accumulate validated chatbot
157 responses from human annotators. To achieve this, we initially obtain response candidates from GPT-
158 4, then facilitate human verification, where annotators determine the correctness of each response. If
159 a response is marked incorrect, the annotator identifies whether the issue lies with the detail level of
160 the dense caption or with GPT-4’s response itself. This approach helps us assess GPT-4’s competency
161 in following instructions based on text-only (instruction, dense caption) pairs, establishing a baseline
162 for future comparisons. Alongside this, the annotators are tasked to screen for any offensive, unsound,
163 or harmful advice present in the responses. An example of this evaluation phase is depicted in
164 Figure 2 (right), which demonstrates the validation of a GPT-4 response. The prompt used to solicit
165 GPT-4’s outputs in this step reads: “*Consider an image depicted by: <caption>*’. Now, briefly follow

¹Annotators were advised to use Openverse (<https://openverse.org/>) for Creative Commons image search.

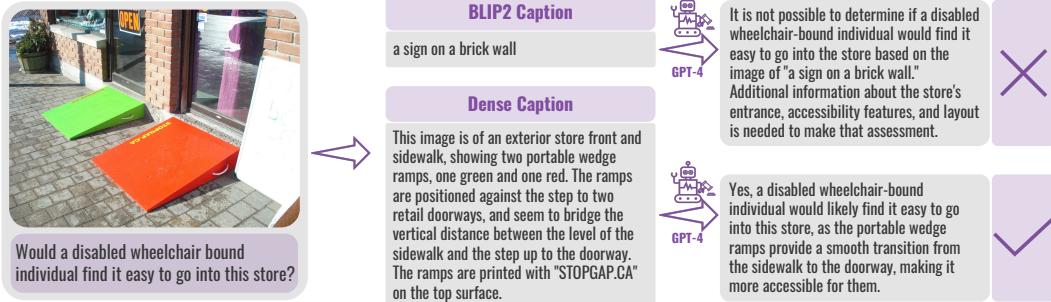


Figure 3: Demonstration of the importance of dense captions for accurate instruction-following instances. Using an image and instruction, GPT-4 produces two responses, one with a dense caption, another with a BLIP2 caption. The latter’s lack of detail results in an incorrect response, underscoring the importance of detailed captions that are specifically tailored to match the given instruction.

166 *this instruction, and you can add a short explanation: <instruction>’. Response: ’li’ This prompt
 167 structure is employed for both single and multiple image instances, with appropriate modifications
 168 for the latter.*

169 **2.2 Data Collection Annotation and Results**

170 We conduct the data collection steps in Figure 2 using Amazon’s Mechanical Turk (MTurk) platform.
 171 Prior to annotating, each MTurk worker passed a qualification test, which involved 5-10 sample tasks
 172 designed to assess their ability to generate high-quality annotations. Annotators were paid at a rate of
 173 18\$ per hour. More detailed information about the execution process and full user interface examples
 174 can be found in Appendix D.

175 Our annotation results are summarized in
 176 Table 2. We measure how well our data
 177 collection pipeline succeeds in generation
 178 correctly-verified chatbot responses. In §4,
 179 we will use the verified responses as ref-
 180 erences for our automatic evaluation. Our
 181 data processing method showed solid per-
 182 formance , particularly with single-image
 183 tasks, scoring 91.5% correctness in GPT-
 184 4’s responses . These results underline the
 185 effectiveness of using the *instruction-conditioned caption* coupled with a text model. However, the
 186 success rate dropped to 63.0% in the more complex multi-image tasks, accompanied by an uptick
 187 in issues either in the captions (6.0%) or GPT-4’s responses (30.0%). This drop hints at the more
 188 difficult challenge posed by multi-image tasks, indicating room for growth.

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

189 **2.3 The Necessity of Dense Captions**

190 To assess the significance of our dense captions, we conducted an experiment involving 150 single-
 191 image instances, depicted in Figure 3. In this experiment, instead of using our instruction-conditioned
 192 captions, we use BLIP2 [24] image captions, which is a state-of-the-art image captioning model. We
 193 extract image captions, and feed to GPT-4 for text-only generation as detailed earlier, to provide a
 194 text-based chatbot response.

195 We manually evaluated whether the resulting output accurately followed the instructions. We find that
 196 while dense captions led to correct outputs in 91% of the cases (table 2), the success rate fell to 31%
 197 when using BLIP2 captions. These results highlight the pivotal role of our *instruction-conditioned*
 198 *caption* to generate correct instruction-following responses.

199 **3 VisIT-Bench Analysis**

200 Here, we analyze VisIT-Bench diversity in terms of the instructions, tasks, and the objects in the
201 dataset images.

202 VisIT-Bench is a composite of 70 unique instruction families, each representing a different capability
203 a chatbot model needs to exhibit, mirroring an array of practical real-world chatbot interactions.
204 Examples of these instruction families are showcased in Figure 1. The construction of VisIT-Bench
205 incorporates a wide variety of data by including 25 prior datasets. Importantly, among these, 10
206 datasets involve multiple images, introducing an elevated layer of task complexity, as illustrated in
207 Figure 4. Hence, the formation of VisIT-Bench assembles a rich spectrum of realistic challenges for
208 assessing the effectiveness of multimodal chatbot models.

209 Following [25], we plot the most frequent root verbs and their direct nouns instructions, from VisIT-
210 Bench, in Appendix Figure 7. We find that the instructions in our dataset are quite diverse, and
211 demand extensive comprehension and commonsense understanding e.g., the instructions range from
212 ‘*identifying objects*’ to ‘*writing stories and poems*’. Moreover, we provide the list of skills that are
213 tested in VisIT-Bench in Appendix Table 7. We highlight the various levels of difficulty present in
214 our dataset ranging from recognition such as ‘emotion identification’ to complex reasoning tasks
215 such as ‘paper folding’. Finally, we detect all the COCO [26] objects present in the images from our
216 dataset using Yolov5-Large [27] in Appendix Figure 8. We observe that the images in the datasets
217 contain a variety of objects with the highest count of ‘person’ ~ 900 times. This indicates that the
218 instruction-following vision-language models need to recognize and reason over a diverse class of
219 objects for good performance on our dataset.

220 We present a summary of the key contributions and comparisons against the related works in Table 1.
221 The ‘wish-list’ approach, a defining feature of VisIT-Bench, is designed to emulate realistic human
222 interactions with instruction-following vision-language models. Moreover, we make an effort to
223 examine a broader range of skills (70) than previous studies. A unique element of VisIT-Bench is
224 its inclusion of *instruction-conditioned caption* and reference outputs for all test instances, which
225 enhances the granularity of performance gap analysis by calculating win rates against reference
226 outputs. Another distinguishing factor of VisIT-Bench is its inclusion of multi-image instruction
227 instances, broadening the diversity of tasks and challenges for the models. Crucially, our automatic
228 evaluation methodology exhibits high congruence with human evaluations, solidifying its credibility
229 as a trustworthy and scalable tool for dynamic model assessments.

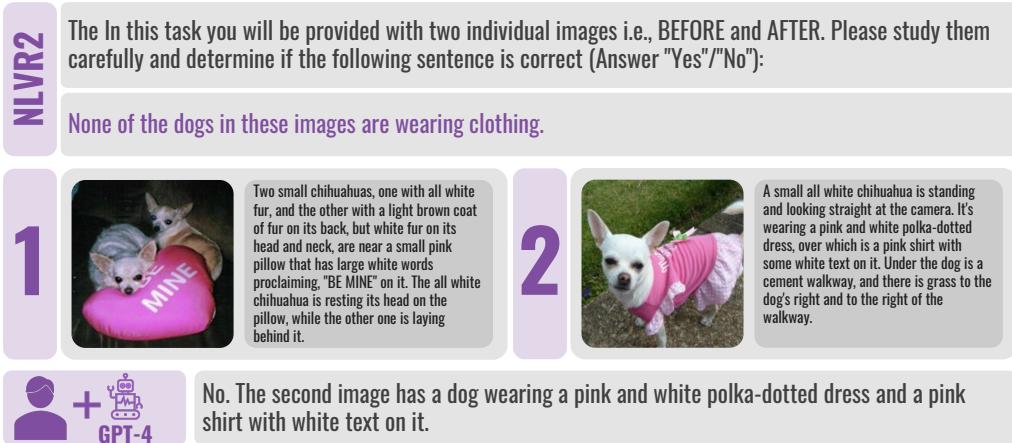


Figure 4: This figure showcases an example of a multi-image instruction task from VisIT-Bench. Originally sourced from NLVR2 [28], this task aims to evaluate visual reasoning skills. The NLVR2 format includes a sentence to analyze, two images, and a binary response. Our adaptation adds a zero-shot instruction 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.

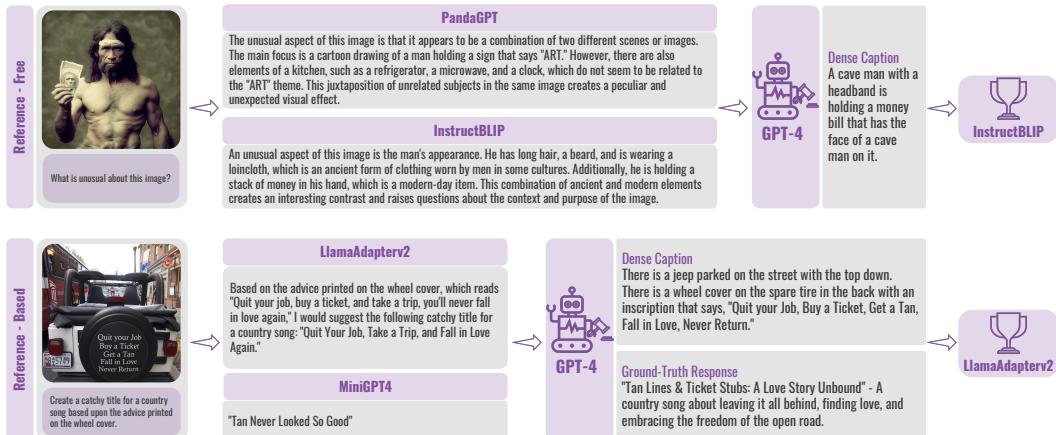


Figure 5: ELO-based automated evaluation for VisIT-Bench: Top - our *reference-free* approach: Our GPT4 based evaluator compare two instruction-following models using the instruction and a *instruction-conditioned caption*. The instance is obtained from an existing dataset, WHOOPS! [29]. Bottom - our *reference-backed* approach: the evaluator now includes a human-verified chatbot response to provide a more comprehensive judgment. The instance was created with our data collection.

230 4 Experiments

231 In this section, we evaluate state-of-the-art vision-and-language chatbots on VisIT-Bench. In §4.1, we
 232 provide the details of the instruction-following models in our benchmark. Following this, we collect
 233 the human preferences for pairwise model generations to achieve a human-guided Elo ranking and
 234 the win-rates against the reference of the models in §4.2. We then develop automatic evaluation on
 235 VisIT-Bench in §4.3, that can be scaled and improved given new and improved models. Finally, we
 236 establish the trustworthiness of our automatic evaluation method by performing agreement analysis
 237 with the human judgments in §4.4. Although VisIT-Bench encompasses 679 instances, our analysis
 238 utilizes the subset of 592 instances with correct verified GPT4 responses (see §2.2), providing a basis
 239 for automated evaluation.

240 **4.1 Models**

241 Our evaluation utilises a selection of publicly accessible vision-language models that have been
 242 fine-tuned with multimodal instructions or conduct execution based on LLM outputs. These mod-
 243 els include LLaVA-13B [30], InstructBLIP-13B [12], MiniGPT4-7B [31], mPLUG-Owl-7B [16],
 244 LlamaAdapter-v2-7B [32], PandaGPT-13B [15], VisualChatGPT [33], Multimodal GPT [34], Open-
 245 Flamingo [35] and Otter [36]. For the execution-based VisualChatGPT [33], we implement a chat
 246 window for each sample, hold inputs and intermediate chains of thoughts and actions in memory, and
 247 feed the images and the instruction sequentially. For the other models [12, 31, 15, 32, 16], we feed
 248 the image to the vision feature extractor and feed the instruction as a prompt for the text encoder.

249 **4.2 Human-Guided Elo Rankings**

250 Here, we first collect human preferences for the pairwise model generations and then assign an Elo
 251 rating to the models based on their wins against their counterparts. We follow the same setup as [20]
 252 for computing the Elo ratings (more details in Appendix F).² In addition, we also provide a win-rate
 253 vs reference score for every model that indicates the majority vote prefers the output from that model
 254 over the GPT-4 reference output. Here we provide the results for the models on the single-image
 255 examples of our dataset. We provide the results for multiple-image examples in the Appendix §4.5.

Table 3: Human scoring results for the models, shown as both an ELO rating and win-rate against the reference. In total, this summarizes 2.4K pairwise judgements.

Model	Elo	Win-rate vs. GPT-4
Human Verified GPT-4 Reference	1208	-
LLaVA (13B)	1053	24.39%
LlamaAdapter-v2 (7B)	1030	27.27%
mPLUG-Owl (7B)	1020	15.62%
InstructBLIP (13B)	974	11.11%
MiniGPT-4 (7B)	917	14.89%
PandaGPT (13B)	798	12.50%

256 To generate the model outputs, we prompt some of the models in §4.4 with the single-image test
 257 examples from VisIT-Bench. Subsequently, we create a total of 5,000 pairwise comparison data
 258 from the model outputs such that each model occurs in ~ 700 comparisons. These pairs also include
 259 the correct reference outputs generated by GPT-4. Then, we select 1,000 samples, and three human
 260 annotators mark their preference, based on accuracy, helpfulness, and detail, for each of the pairwise
 261 comparisons unbeknownst to the model responsible for output generation. We provide the template
 262 for the human annotation process in Appendix Figure 13.

263 In Table 3, we analyze the performance of various vision-language models based on Elo ratings and
 264 win rates against reference outputs. Human-verified GPT-4 responses excel on our single-image
 265 dataset, likely due to their capacity for intricate reasoning with the aid of the *instruction-conditioned*
 266 *captions*. Among multimodal models, LLaVA (13B) achieves the top performance. Its 150K
 267 instruction-tuning data, rich in diversity and quality, might contribute to this result. Interestingly,
 268 LlamaAdapter-v2 (7B) wins the most head-to-head battles against the reference outputs. This
 269 suggests that a focus on fine-tuning language instructions could enhance visual instruction-following
 270 capabilities. We note significant performance variation among models of the same size (e.g., 7B and
 271 13B), hinting at the possible influence of fine-tuning strategies and instruction-tuning data quality.
 272 However, the complexity and variety of models and tasks in VisIT-Bench makes it challenging to
 273 definitively pinpoint the factors influencing performance. While we make a preliminary attempt to
 274 unravel these intricacies in section 4.3, a comprehensive understanding will necessitate more nuanced
 275 and extensive future research.

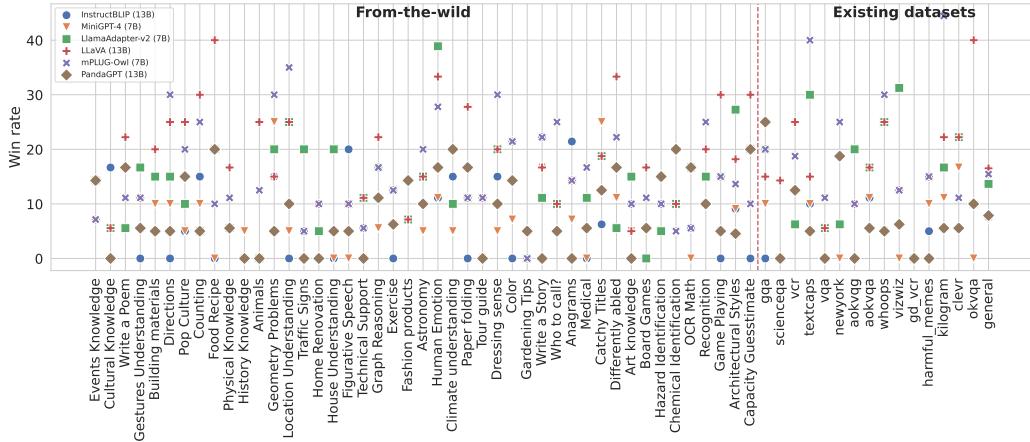


Figure 6: Win rate vs. human-verified GPT4 response for each instruction category in our reference-free assessment. Axes: win rate (Y), instruction categories (X). Categories are from-the-wild or existing datasets. Our benchmark facilitates analysis of diverse instruction tuning tasks.

Table 4: Current reference-free elo rankings as of June 14th, 2023. In total, these rankings summarize 12K "battles" between models; each battle 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	RFE	Battles	Win vs. Verified GPT-4
Human Verified GPT-4 Reference	1363	3274	-
LLaVA (13B)	1099	3274	5.03%
mPLUG-Owl (7B)	1053	3284	4.55%
LlamaAdapter-v2 (7B)	1037	3281	3.8%
Otter (9B)	998	154	2.50%
InstructBLIP (13B)	992	3274	2.37%
VisualGPT (Da Vinci 003)	967	251	1.92%
MiniGPT-4 (7B)	925	3291	2.09%
OpenFlamingo (9B)	892	441	0.0%
Multimodal GPT	854	267	0.0%
PandaGPT (13B)	820	3275	0.85%

276 4.3 Automatic Evaluation on VisIT-Bench and Dynamic Leaderboard

277 We propose an automatic evaluation framework to rank the model outputs. We summarize those
 278 rankings with Elo ratings, and win-rate against the ground-truth. Following the recent line of work in
 279 utilizing strong API LLMs for automated evaluation [6, 37], we adopt GPT-4 to provide pairwise
 280 preferences given an instruction and two candidate outputs. We acknowledge the potential bias
 281 introduced by using GPT-4 to evaluate its own output. To address this concern, we measure and
 282 demonstrate high agreement with human ratings in section 4.4. Specifically, we provide the LLM
 283 with: 1) a system prompt describing the desired evaluation behavior; 2) the dense description of
 284 the image; 3) the instruction to be followed (and, optionally, a reference); and 4) two candidate
 285 generations dubbed “Response A” and “Response B”. We provide our prompts in Appendix G. To
 286 mitigate bias in “A” and “B” positioning, for all pairs of candidates, we run two queries covering
 287 both possible orderings. Our prompt encourages the model to think step-by-step so that its chain-
 288 of-thought process is made explicit [38, 39]. We summarize the LLM’s judgements using three
 289 automated metrics: **(1) Reference-free Elo score:** this setup closely mirrors the human scoring setup

²We re-use the following code for Elo ratings: https://github.com/lm-sys/FastChat/blob/main/fastchat/serve/monitor/elo_analysis.py

Table 5: Overall win rates of multi-image instruction-following models via human assessment.

Model	Overall Win Rate (%)	# Battles
GPT-4 Reference	94.2	70
mPLUG-Owl	48	66
Otter	25.5	51
OpenFlamingo	20	59

in §4.2; (2) **Reference-backed Elo score**: this metric is computed similarly, except that the LLM is additionally provided the optional reference in the prompt (and the prompt is slightly adjusted to account for that); (3) **Win-rate versus reference**: because we also run pairwise LLM judgments treating the human-curated reference as an option, we can compute a non-Elo metric as simply: the percent of time the LLM perfers the model’s generation to the reference.

The results are presented in Table 4 in the descending order of the Elo ratings. We observe that the Elo rankings from the reference-free and reference-backed evaluations are consistent with each other. In addition, we find that the win rates of the existing instruction-following models range from 0% – 5.94% further highlighting the large gap in their performance against the reference outputs. We note that there is a gap between the absolute win-rate vs GPT-4 scores Table 4 vs. the scores in Table 3, which might be attributed to the fact that GPT-4 evaluation prefers GPT-4 generated responses unlike an unbiased human evaluator.

In Figure 6, we further highlight 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 usecases.

4.4 Analysis and Validity of Automatic Evaluation

Given the ground truth human ranking and automatic evaluation ranking results, we analyze the automatic evaluation results from Table 4, and verify their validity by computing agreement metrics with human ground truth. Overall, we observe that human rankings scores are highly correlated with the automatic evaluation scores. GPT-4 reference scores the highest on both the reference-free ELO metric and human evaluation. Additionally, both automatic ELO scores suggest a large gap between the rest of the models and MiniGPT-4 and PandaGPT-13B, a finding which is also replicated in human evaluation.

We additionally verify our automatic evaluation metrics with an instance-level agreement analysis. Specifically, we compare the human majority vote preference and the automatic evaluation preference for 500 examples in the pairwise data. We find that the reference-free automatic evaluation agrees with the human majority votes 71.02% of the time while the reference-backed automatic evaluation agrees with the human majority votes 69.3% of the time. To further verify the trustworthiness of the automatic evaluation, we compute the instance-level agreement between the human majority votes and always choosing the longer response output from the pair of the model outputs. Here, we find that the agreement score is 48.73%, suggesting the trustworthiness of our automatic evaluation .

4.5 Results on Instructions with Multiple Images

Since VisIT-Bench supports instances with multiple images that are derived from the existing datasets, we evaluate the models including OpenFlamingo, Otter, and mPLUG-Owl on them.

In a preliminary human evaluation study consisting of 246 pairwise human judgements, the human-verified GPT-4 reference is preferred in 95% of cases; mPLUG-OWL in 48% of cases; Otter in 25% of cases; and OpenFlamingo in 20% of cases. Because not all models support multi-image inputs, we are building a separate leaderboard submission process. We show these preliminary numbers along with the number of battles in Table 5.

330 **5 Related Work**

331 Our work is related to previous work on multimodal models for image-text understanding, instruction-
332 following, and benchmarking in machine learning. We provide a detailed overview of related work
333 in §C. Multi-modal models for image-text understanding have recently emerged as powerful and
334 useful methods for many image-language reasoning tasks [14, 24, 11, 40, 15, 12, 16, 35, 7]. Both
335 language and multimodal models are often trained to follow language instruction, a paradigm known
336 as “instruction following” [5, 32, 41, 12, 11, 16]. Despite the success of these approaches on existing
337 vision-language datasets (GQA, Image Captioning [18, 42, 26]), there is a lack of a high-quality
338 benchmarking dataset for multimodal instruction-following tasks that reliably replicates the way
339 in which humans would interact with multimodal chatbots in the wild. This lack of benchmarking
340 data makes reliable assessments of progress difficult [43], and has contributed to a lack of empirical
341 evaluations of instruction-following models.

342 **6 Conclusion**

343 In conclusion, we introduce VisIT-Bench, a dynamic benchmark providing a broad, rigorous assess-
344 ment of multimodal chatbots’ capabilities. It spans 70 diverse instruction families, encompassing a
345 range of tasks from recognition to complex reasoning. Our benchmark not only offers human-verified
346 reference outputs for all test examples but also integrates an Elo-based ranking system for multimodal
347 chatbots. The experiments conducted reveal that current models, despite significant advances, still
348 trail behind human performance. To further the development of more robust and reliable multimodal
349 models, we openly provide the VisIT-Bench data, code, and automatic metrics, encouraging commu-
350 nity involvement. We hope future advancements in multimodal AI will leverage this benchmark to
351 refine their models and address more nuanced real-world human instructions effectively.

352 **7 Limitations**

353 Although VisIT-Bench covers a wide spectrum of tasks, it does not incorporate every potential
354 vision-language dataset. As part of our ongoing refinement process, we aim to integrate more datasets
355 dynamically for comprehensive evaluations. Our current focus on single-turn dialogues, though
356 insightful, does not mimic real-world multi-turn interactions between users and chatbots. A future
357 direction could involve incorporating such dialogues into the evaluation process. The scope of our
358 study is limited to image-text modality. Future extensions might incorporate additional modalities,
359 such as audio and video, for a more comprehensive evaluation. Lastly, while we’ve strived to mitigate
360 potential biases in images and GPT-4 outputs, interpretations of what is considered offensive can vary.
361 Thus, the results should be interpreted with an awareness of these potential biases.

362 **Acknowledgements**

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

366 **References**

- 367 [1] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy
368 Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a
369 laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.
- 370 [2] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin,
371 Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to
372 follow instructions with human feedback. *Advances in Neural Information Processing Systems*,
373 35:27730–27744, 2022.

- 374 [3] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan
 375 Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. [arXiv](#)
 376 preprint arXiv:2109.01652, 2021.
- 377 [4] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei,
 378 Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al.
 379 Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In
 380 [Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing](#),
 381 pages 5085–5109, 2022.
- 382 [5] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy
 383 Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model.
 384 https://github.com/tatsu-lab/stanford_alpaca, 2023.
- 385 [6] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng,
 386 Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna:
 387 An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL
 388 <https://lmsys.org/blog/2023-03-30-vicuna/>.
- 389 [7] OpenAI. Gpt-4 technical report. [arXiv](#), 2023.
- 390 [8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal,
 391 Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are
 392 few-shot learners. [Advances in neural information processing systems](#), 33:1877–1901, 2020.
- 393 [9] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam
 394 Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm:
 395 Scaling language modeling with pathways. [arXiv preprint arXiv:2204.02311](#), 2022.
- 396 [10] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timo-
 397 thée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open
 398 and efficient foundation language models. [arXiv preprint arXiv:2302.13971](#), 2023.
- 399 [11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.
- 400 [12] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng
 401 Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose
 402 vision-language models with instruction tuning. [arXiv preprint arXiv:2305.06500](#), 2023.
- 403 [13] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning
 404 with gpt-4. [arXiv preprint arXiv:2304.03277](#), 2023.
- 405 [14] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson,
 406 Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford,
 407 Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick,
 408 Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski,
 409 Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual
 410 language model for few-shot learning. [ArXiv](#), abs/2204.14198, 2022.
- 411 [15] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to
 412 instruction-follow them all. [arXiv preprint arXiv:2305.16355](#), 2023.
- 413 [16] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang,
 414 Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large
 415 language models with multimodality. [arXiv preprint arXiv:2304.14178](#), 2023.
- 416 [17] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari.
 417 Connecting vision and language with localized narratives. In [Computer Vision–ECCV 2020:](#)
 418 [16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16](#), pages
 419 647–664. Springer, 2020.

- 420 [18] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence
421 Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE*
422 *international conference on computer vision*, pages 2425–2433, 2015.
- 423 [19] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *Proceedings*
424 *of the IEEE/CVF International Conference on Computer Vision*, pages 4624–4633, 2019.
- 425 [20] Lianmin Zheng, Ying Sheng, Wei-Lin Chiang, Hao Zhang, Joseph E. Gonzalez, and Ion
426 Stoica. Chatbot arena: Benchmarking llms in the wild with elo ratings. 2023. URL <https://lmsys.org/blog/2023-05-03-area/>.
- 428 [21] Zhiyang Xu, Ying Shen, and Lifu Huang. Multiinstruct: Improving multi-modal zero-shot
429 learning via instruction tuning. *arXiv preprint arXiv:2212.10773*, 2022.
- 430 [22] Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng
431 Yang, Jingjing Xu, Xu Sun, et al. M3 it: A large-scale dataset towards multi-modal multilingual
432 instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023.
- 433 [23] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan
434 Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large
435 vision-language models. *arXiv preprint arXiv:2306.09265*, 2023.
- 436 [24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-
437 image pre-training with frozen image encoders and large language models. *arXiv preprint*
438 *arXiv:2301.12597*, 2023.
- 439 [25] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi,
440 and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instruc-
441 tions. *arXiv preprint arXiv:2212.10560*, 2022.
- 442 [26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr
443 Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer*
444 *Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014,*
445 *Proceedings, Part V 13*, pages 740–755. Springer, 2014.
- 446 [27] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified,
447 real-time object detection. In *Proceedings of the IEEE conference on computer vision and*
448 *pattern recognition*, pages 779–788, 2016.
- 449 [28] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for
450 reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*,
451 2018.
- 452 [29] Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel
453 Stanovsky, and Roy Schwartz. Breaking common sense: Whoops! a vision-and-language
454 benchmark of synthetic and compositional images. *arXiv preprint arXiv:2303.07274*, 2023.
- 455 [30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv*
456 *preprint arXiv:2304.08485*, 2023.
- 457 [31] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: En-
458 hancing vision-language understanding with advanced large language models. *arXiv preprint*
459 *arXiv:2304.10592*, 2023.
- 460 [32] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan
461 Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction
462 model. *arXiv preprint arXiv:2304.15010*, 2023.

- 463 [33] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan.
 464 Visual chatgpt: Talking, drawing and editing with visual foundation models. [arXiv preprint](#)
 465 [arXiv:2303.04671](#), 2023.
- 466 [34] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu,
 467 Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for
 468 dialogue with humans. [arXiv preprint arXiv:2305.04790](#), 2023.
- 469 [35] Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani
 470 Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel
 471 Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo, March 2023. URL <https://doi.org/10.5281/zenodo.7733589>.
- 473 [36] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A
 474 multi-modal model with in-context instruction tuning. [arXiv preprint arXiv:2305.03726](#), 2023.
- 475 [37] Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos
 476 Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacafarm: A simulation framework for
 477 methods that learn from human feedback, 2023.
- 478 [38] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny
 479 Zhou. Chain-of-thought prompting elicits reasoning in large language models. In [NeurIPS](#),
 480 2022. URL <https://arxiv.org/abs/2201.11903>.
- 481 [39] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large
 482 language models are zero-shot reasoners. In [NeurIPS](#), 2022. URL <https://arxiv.org/abs/2205.11916>.
- 484 [40] Rohan Pandey, Rulin Shao, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency.
 485 Cross-modal attention congruence regularization for vision-language relation alignment. [arXiv](#)
 486 [preprint arXiv:2212.10549](#), 2022.
- 487 [41] Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and
 488 Dawn Song. Koala: A dialogue model for academic research. Blog post, April 2023. URL
 489 <https://bair.berkeley.edu/blog/2023/04/03/koala/>.
- 490 [42] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual
 491 reasoning and compositional question answering. In [Proceedings of the IEEE/CVF conference](#)
 492 [on computer vision and pattern recognition](#), pages 6700–6709, 2019.
- 493 [43] Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. Are we learning
 494 yet? a meta review of evaluation failures across machine learning. In [Thirty-fifth Conference](#)
 495 [on Neural Information Processing Systems Datasets and Benchmarks Track \(Round 2\)](#), 2021.
- 496 [44] Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. [arXiv](#)
 497 [preprint arXiv:2203.08242](#), 2022.
- 498 [45] Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. Stop uploading test data
 499 in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks.
 500 [arXiv preprint arXiv:2305.10160](#), 2023.
- 501 [46] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman,
 502 Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-
 503 5b: An open large-scale dataset for training next generation image-text models. [arXiv preprint](#)
 504 [arXiv:2210.08402](#), 2022.
- 505 [47] Da Yin, Xiao Liu, Fan Yin, Ming Zhong, Hritik Bansal, Jiawei Han, and Kai-Wei Chang.
 506 Dynosaur: A dynamic growth paradigm for instruction-tuning data curation. [arXiv preprint](#)
 507 [arXiv:2305.14327](#), 2023.

- 508 [48] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivas Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma,
 509 Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. [arXiv preprint](#)
 510 [arXiv:2305.11206](#), 2023.
- 511 [49] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu
 512 Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image
 513 pre-training. In [Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern](#)
 514 [Recognition](#), pages 10965–10975, 2022.
- 515 [50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-
 516 resolution image synthesis with latent diffusion models. In [Proceedings of the IEEE/CVF](#)
 517 [Conference on Computer Vision and Pattern Recognition](#), pages 10684–10695, 2022.
- 518 [51] Zhengyu Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed,
 519 Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for
 520 multimodal reasoning and action. [arXiv preprint arXiv:2303.11381](#), 2023.
- 521 [52] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution
 522 for reasoning. [arXiv preprint arXiv:2303.08128](#), 2023.
- 523 [53] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning
 524 without training. In [Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern](#)
 525 [Recognition](#), pages 14953–14962, 2023.
- 526 [54] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun
 527 Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language
 528 models. [arXiv preprint arXiv:2304.09842](#), 2023.
- 529 [55] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu,
 530 Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for
 531 dialogue with humans, 2023.
- 532 [56] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao
 533 Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning
 534 perception with language models. [arXiv preprint arXiv:2302.14045](#), 2023.
- 535 [57] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng
 536 Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual
 537 recognition challenge. [International journal of computer vision](#), 115:211–252, 2015.
- 538 [58] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.
 539 2009.
- 540 [59] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don’t know: Unanswerable
 541 questions for squad. [arXiv preprint arXiv:1806.03822](#), 2018.
- 542 [60] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y
 543 Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a
 544 sentiment treebank. In [Proceedings of the 2013 conference on empirical methods in natural](#)
 545 [language processing](#), pages 1631–1642, 2013.
- 546 [61] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman.
 547 Glue: A multi-task benchmark and analysis platform for natural language understanding. [arXiv](#)
 548 [preprint arXiv:1804.07461](#), 2018.
- 549 [62] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix
 550 Hill, Omer Levy, and Samuel Bowman. Super glue: A stickier benchmark for general-purpose
 551 language understanding systems. [Advances in neural information processing systems](#), 32, 2019.

- 552 [63] Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence
 553 Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, et al. A framework for few-shot
 554 language model evaluation. *Version v0. 0.1.* Sept, 2021.
- 555 [64] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga,
 556 Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of
 557 language models. *arXiv preprint arXiv:2211.09110*, 2022.
- 558 [65] John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. The effect of natural distri-
 559 bution shift on question answering models. In *International Conference on Machine Learning*,
 560 pages 6905–6916. PMLR, 2020.
- 561 [66] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10
 562 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018.
- 563 [67] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet
 564 classifiers generalize to imagenet? In *International conference on machine learning*, pages
 565 5389–5400. PMLR, 2019.
- 566 [68] Rebecca Roelofs, Vaishaal Shankar, Benjamin Recht, Sara Fridovich-Keil, Moritz Hardt, John
 567 Miller, and Ludwig Schmidt. A meta-analysis of overfitting in machine learning. *Advances in
 568 Neural Information Processing Systems*, 32, 2019.
- 569 [69] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition:
 570 Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer
 571 vision and pattern recognition*, pages 6720–6731, 2019.
- 572 [70] Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. Broaden the vision:
 573 Geo-diverse visual commonsense reasoning. *arXiv preprint arXiv:2109.06860*, 2021.
- 574 [71] Jack Hessel, Ana Marasović, Jena D Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert
 575 Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor" understanding"
 576 benchmarks from the new yorker caption contest. *arXiv preprint arXiv:2209.06293*, 2022.
- 577 [72] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick,
 578 and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary
 579 visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern
 580 recognition*, pages 2901–2910, 2017.
- 581 [73] Anya Ji, Noriyuki Kojima, Noah Rush, Alane Suhr, Wai Keen Vong, Robert D Hawkins, and
 582 Yoav Artzi. Abstract visual reasoning with tangram shapes. *arXiv preprint arXiv:2211.16492*,
 583 2022.
- 584 [74] Shivam Sharma, Siddhant Agarwal, Tharun Suresh, Preslav Nakov, Md Shad Akhtar, and
 585 Tanmoy Charkraborty. What do you meme? generating explanations for visual semantic role
 586 labelling in memes. *arXiv preprint arXiv:2212.00715*, 2022.
- 587 [75] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind
 588 Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via
 589 thought chains for science question answering. In *The 36th Conference on Neural Information
 590 Processing Systems (NeurIPS)*, 2022.
- 591 [76] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual
 592 question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf
 593 conference on computer vision and pattern recognition*, pages 3195–3204, 2019.
- 594 [77] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh
 595 Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. *arXiv*,
 596 2022.

- 597 [78] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller,
598 Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-
599 time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User*
600 *interface software and technology*, pages 333–342, 2010.
- 601 [79] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset
602 for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th*
603 *European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–
604 758. Springer, 2020.
- 605 [80] Benno Krojer, Vaibhav Adlakha, Vibhav Vineet, Yash Goyal, Edoardo Ponti, and Siva Reddy.
606 Image retrieval from contextual descriptions. *arXiv preprint arXiv:2203.15867*, 2022.
- 607 [81] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of
608 similar images. *arXiv preprint arXiv:1808.10584*, 2018.
- 609 [82] Yonatan Bitton, Ron Yosef, Eli Strugo, Dafna Shahaf, Roy Schwartz, and Gabriel Stanovsky.
610 Vavr: Visual analogies of situation recognition. *arXiv preprint arXiv:2212.04542*, 2022.
- 611 [83] Yonatan Bitton, Nitzan Bitton Guetta, Ron Yosef, Yuval Elovici, Mohit Bansal, Gabriel
612 Stanovsky, and Roy Schwartz. Winogavil: Gamified association benchmark to challenge
613 vision-and-language models. *Advances in Neural Information Processing Systems*, 35:26549–
614 26564, 2022.
- 615 [84] Ron Yosef, Yonatan Bitton, and Dafna Shahaf. Irlf: Image recognition of figurative language.
616 *arXiv preprint arXiv:2303.15445*, 2023.
- 617 [85] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang,
618 and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual
619 language reasoning. *arXiv preprint arXiv:2110.13214*, 2021.
- 620 [86] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy.
621 Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint*
622 *arXiv:2305.01569*, 2023.
- 623 [87] Arpad E Elo. The proposed uscf rating system. its development, theory, and applications. *Chess*
624 *Life*, 22(8):242–247, 1967.

625 8 Checklist

- 626 1. For all authors...
 - 627 (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s
628 contributions and scope? **[Yes]**
 - 629 (b) Did you describe the limitations of your work? **[Yes]**
 - 630 (c) Did you discuss any potential negative societal impacts of your work? **[N/A]**
 - 631 (d) Have you read the ethics review guidelines and ensured that your paper conforms to
632 them? **[Yes]**
- 633 2. If you are including theoretical results...
 - 634 (a) Did you state the full set of assumptions of all theoretical results? **[N/A]**
 - 635 (b) Did you include complete proofs of all theoretical results? **[N/A]**
- 636 3. If you ran experiments (e.g. for benchmarks)...
 - 637 (a) Did you include the code, data, and instructions needed to reproduce the main experi-
638 mental results (either in the supplemental material or as a URL)? **[Yes]**
 - 639 (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they
640 were chosen)? **[N/A]**

- (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

 - (a) If your work uses existing assets, did you cite the creators? [Yes]
 - (b) Did you mention the license of the assets? [Yes]
 - (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
 - (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]

5. If you used crowdsourcing or conducted research with human subjects...

 - (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes]
 - (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [No]
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes]

660 Appendix

661 A License and Intended Use

662 The Visit-Bench dataset, along with its various contributions such as instructions, reference outputs,
663 and model ranking annotations, is licensed under the Creative Commons Attribution 4.0 International
664 License (CC BY 4.0). This license applies to all the images we have directly contributed, each of
665 which carries a public license specification in the “public images metadata” field within the dataset
666 sheets. However, the dataset also incorporates images sourced from pre-existing collections. For
667 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 [44, 45]. 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.

681 B Dataset Analysis

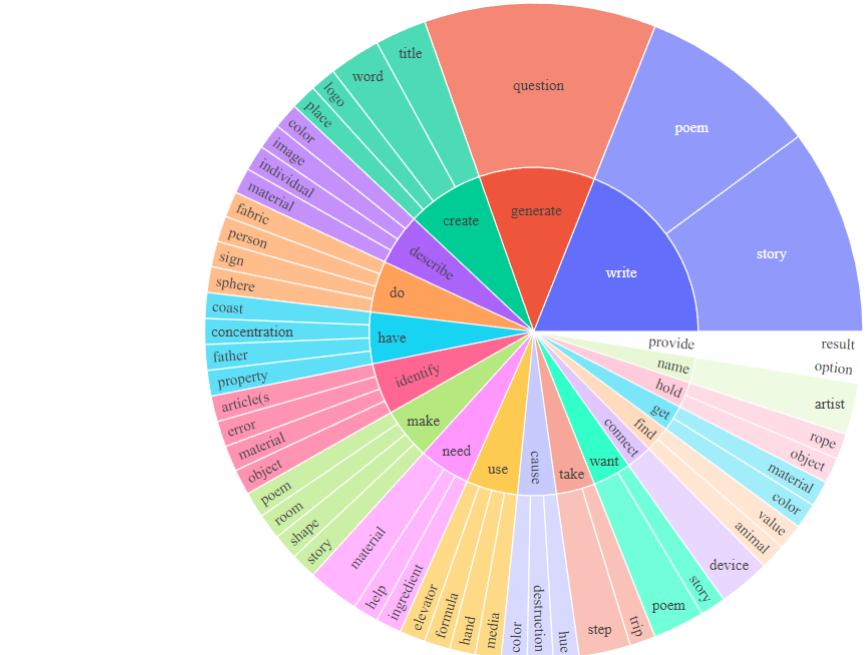


Figure 7: Most frequently occurring verbs (inner circle) and their top 4 direct nouns (outer circle) in the VisIT-Bench instructions.

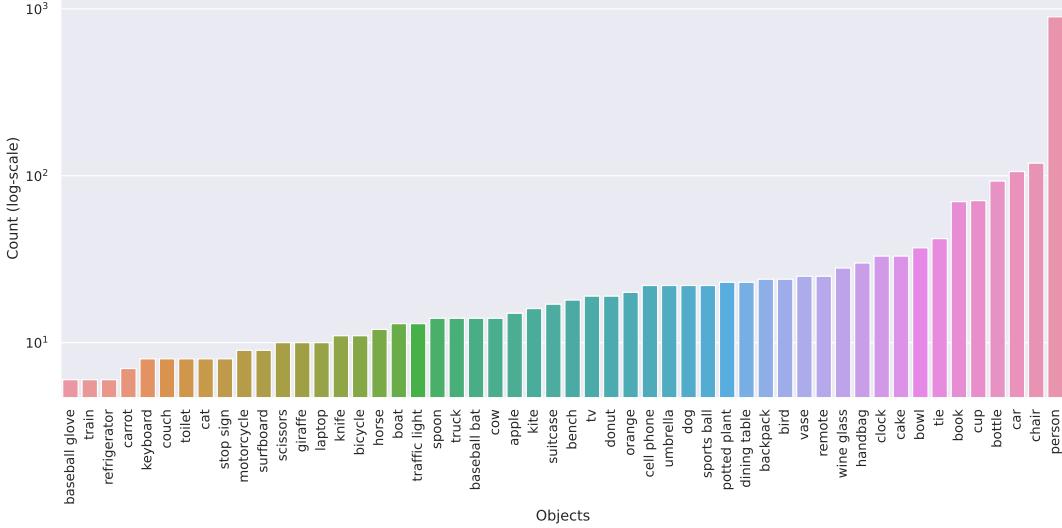


Figure 8: Count of the various COCO objects present in the VisIT-Bench images on a log-scale. The object detection was performed by Yolov5-Large [27]. We observe that the object ‘person’ occurs most of the time in the dataset images i.e., 875 times.

682 C Related Work

683 **Multimodal Models for Image-Text Understanding:** Recently, the field of machine learning
 684 has experienced a rapid proliferation of new models which can perform various image-text tasks
 685 [14, 24, 11, 40, 15, 12]. This growth has been driven by several factors, including the emergence of
 686 large-scale multimodal datasets (e.g. LAION-5B [46], Multimodal C4 [35]), improved software and
 687 hardware frameworks, and advances in modality-specific models such as language models (e.g., [10]).
 688 Our work specifically evaluates models which can generate textual outputs, given one or more images,
 689 and text. Recent examples of such models include LLaVA [11], mPLUG-Owl [16], InstructBLIP,
 690 LLaMA-Adapter, Flamingo [14] and OpenFlamingo [35], PandaGPT [15], and GPT-4 [7] (which
 691 reports multimodal capabilities but has not yet seen a release of the multimodal variant).

692 **Instruction Following:** “Instruction-following” is an emerging paradigm for training models via
 693 language, where instead of being trained to complete only a single, fixed task (such as image
 694 classification or captioning), models are trained to follow textual instructions that describe an
 695 arbitrary task, with the aim of generalizing to novel instructions. Examples of instruction-following
 696 models include Alpaca [5], LLaMA-Adapter [32], Koala [41], InstructBLIP [12], LLaVA [11], and
 697 mPLUG-owl [16]. As the downstream capabilities of these models are influenced by the quality of
 698 the training dataset, there has also been extensive work on developing instruction-following datasets
 699 [25, 13, 47, 11, 48].

700 To build powerful these models, two broad approaches have been shown to be effective. One
 701 approach focuses on leveraging existing pretrained task-specific tools such as image captioners
 702 [24], object detectors [49] and text-to-image generators [50] by either creating multimodal prompt
 703 interfaces [33, 51] or by executing LLM-generated programs [52, 53, 54]. The other approach
 704 [30, 32, 55, 36, 56, 16, 35] focuses on building a single pretrained model that can follow instructions
 705 by supervised finetuning on multimodal vision-language data.

706 Despite the success of both these approaches on the existing vision-language datasets e.g., VQA,
 707 GQA, Image Captioning [18, 42, 26], there is a lack of a high-quality benchmarking dataset for
 708 multimodal instruction-following tasks that reliably replicates the way in which humans would
 709 interact with multimodal chatbots in the wild. Similar to the image-text models discussed above,
 710 many instruction-following models have been released directly as open-source without undergoing

711 peer review or thorough evaluation. As a result, the effectiveness of these models for many tasks is
712 not well-understood.

713 **Benchmarks for Machine Learning:** High-quality evaluation datasets have served both to (re)assess,
714 and to accelerate, progress on many machine learning tasks [43]. For example, our work draws
715 particularly from the fields of computer vision and natural language processing, where benchmarking
716 datasets have been critical drivers of progress. On the vision side, datasets such as ImageNet [57] and
717 CIFAR [58] have proven to be critical yardsticks of progress. On the language side, benchmarks such
718 as SQuAD [59], SST [60], GLUE/SuperGLUE [61, 62] and more [63, 64] seen wide use. Recent
719 work has indicated that improvements on these high-quality benchmark datasets is not the result of
720 overfitting, and is a reliable indicator of genuine progress beyond the benchmark data [65, 66, 67, 68].

721 However, high-quality benchmarking datasets and evaluation methods do not yet exist for multimodal
722 instruction-following. As a result, it is difficult to assess progress in this direction, which both reduces
723 the field’s ability to identify true breakthroughs and increases vulnerability to potential pitfalls of
724 evaluation that have hampered progress in other areas of machine learning [43].

725 D Interfaces for Collecting Human Annotations

726 In this section, we provide the templates we used to collect human annotations for the instruction
727 generation (Figure 9), the dense caption generation (Figure 10), the model verification (Figure 11 and
Figure 12), and the model rating (Figure 13).

	Instruction What is this dish and how can I make it?	Expected Output This dish appears to be Massaman Cur oil in a large pot, add onion and saut, a curry paste, ginger and garlic. 3. Saute chicken and potatoes are cooked throu minutes. 5. Serve with rice (white, brow cilantro on top.
<p>Find an Image to Annotate to test Food Recipe, and create a Visual instruction, that a textual-only model/human won't</p> <p>Feel free to skip the HIT if you don't feel creative with this one</p> <p>Provide a **PUBLICALLY LICENSED** image URL. 0/200</p> <p>Provide a **creative and challenging** model instruction that would blow your mind if an AI model could answer it. 0/200</p> <p style="text-align: center;">Previous HIT Showing Task 23 of 70 Next HIT</p>		

Figure 9: An interface that collects the image and corresponding instruction.

	Instruction (Tested Skill=Exercise) Based on commonly held beliefs about upper body strength in men and women which competitor is likely to win this competition. A highly visual description that will allow TEXT-ONLY models *MULTIPLE* answers for the instructions WITHOUT THE IMAGE 0/2000
---	---

Figure 10: An interface that collects the dense caption given an image-and-instruction pair.



Dense Caption:
A woman wearing a red gown is sitting in a wheelchair but dancing with a man in a black tuxedo and white dress shirt. She is using her left hand to hold his right hand while posing with her other hand. He has left arm stretched out. They are performing on a stage with purple highlights.

Instruction:
Why is seeing this couple compete in this event an uncommon sight? Does this convey a positive message?

GPT-4 Prediction:
Seeing this couple compete is an uncommon sight because the woman is in a wheelchair, which is not typical in traditional dance performances. However, this conveys a positive message as it showcases inclusivity, determination, and the ability to overcome challenges.

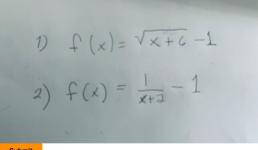
Does GPT4 correctly follow the instruction?
 Yes
 No

If answered 'No' for the above, mark one of the following:
1. N/A
2. The dense caption is detailed enough; the problem is in GPT4's answer
3. The dense caption is underspecified; not GPT4's fault

* Might be offensive/bad advise/problematic

Submit

Figure 11: A sample of the model verification where the GPT4 follows the instruction correctly.



Dense Caption:
There are two equations on a white paper: 1) $f(x) = \sqrt{x+6} - 1$ and 2) $f(x) = \frac{1}{x+2} - 1$.

Instruction:
For the following exercises, graph the functions by translating, stretching, and/or compressing a toolkit function.

GPT-4 Prediction:
To graph the functions, follow these steps: 1) For $y(x) = \sqrt{x+6} - 1$, start with the toolkit function $y = \sqrt{x}$. Translate it 6 units to the left and 1 unit down. 2) For $y(x) = \frac{1}{x+2} - 1$, start with the toolkit function $y = \frac{1}{x}$. Translate it 2 units to the left and 1 unit down.

Does GPT4 correctly follow the instruction?
 Yes
 No

If answered 'No' for the above, mark one of the following:
1. N/A
2. The dense caption is detailed enough; the problem is in GPT4's answer
3. The dense caption is underspecified; not GPT4's fault

* Might be offensive/bad advise/problematic

Submit

Figure 12: A sample of the model verification where the GPT4 failed to follow the instruction due to its incorrect reasoning.



Instruction: What is the person doing looking at the laser ?

A: The person is examining the red laser, possibly trying to figure out its purpose or how to avoid it. He might be in a high-security area or facing an obstacle.

B: The person is standing in front of a laser with their back to the camera. They are looking at the laser, possibly assessing or examining its features.

Which model better fully follows the instruction, providing factual, helpful and accurate response, and adhering to all of the instruction details?

Model A
 Model B

Submit

Figure 13: An interface that collects the feedback of the model rating.

729 **E Existing Datasets incorporated in VisIT-Bench**

730 In Table 6, we listed the existing datasets that are incorporated in our VisIT-Bench. Among these
 731 datasets, 15 contain a single image in each sample pair, and 10 require reasoning based on multiple
 732 images.

Table 6: List of existing datasets in VisIT-Bench, categorized as single and multiple image datasets.

	Dataset	Topic
Single	VQA [18]	Visual Question Answering
	VCR [69]	Cognition-level Visual Understanding
	GD-VCR [70]	Geo-Diverse Commonsense Reasoning
	WHOOPS [29]	What Makes this Image Strange
	Newyork Caption [71]	Humor Understanding
	CLEVR [72]	Visual Question Answering
	Kilogram [73]	Tangrams Identification
	Harmful Memes [74]	Memes Understanding
	ScienceQA [75]	Science Question Answering
	OK-VQA [76]	Outside Knowledge Visual Question Answering
Multiple	AOK-VQA [77]	Outside Knowledge Visual Question
	AOK-VQA [77]	Question Generation
	VizWiz [78]	Visual Question Answering
	GQA [42]	Visual Question Answering on Scene Graphs
	TextCaps [79]	Visual Question Answering on Texts
	Robust Change Captioning [19]	Describing What has Change in a Scene
	NLVR2 [28]	Testing Visual Language Bias
	ImageCoDE [80]	Image Retrieval
	Spot-the-Diff [81]	Identifying Differences
	VASR [82]	Visual Analogies
	WinoGavil [83]	Visual Associations
	IRFL (Metaphor) [84]	Figurative Speech Understanding
	IRFL (Idioms) [84]	Figurative Speech Understanding
	IconQA [85]	Abstract Diagram Understanding
	Pick-a-Pic[86]	Text-to-Image User Preferences

Table 7: 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', 'wino-gavil', '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'

733 **F Elo Rating**

734 For many years, the Elo rating has been popular in ranking players in zero-sum games such as chess
735 [87]. Recently, it has been adopted to rate large language models (LLMs) against each other on the
736 user instructions. In this work, we adopt the same strategy to rank a set of instruction-following
737 vision-language models, that can grow dynamically with further advances in the field.

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

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

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

744 **G GPT-4 Pairwise Evaluation Prompts**

745 The specific prompts we use to extract pairwise judgements from our language model are provided in
746 Table 14 (reference-free version) and Table 15 (reference-backed version). When applied to GPT-4
747 [7], these prompts usually solicit a definitive pairwise response by the model. But, in some cases, the
748 model either produces a pairwise judgement in an unexpected format, or, refuses to issue a judgement
749 at all. For cases like these, we issue an additional query to ChatGPT to extract an answer (or decide
750 there is no answer) using an additional prompt, given in Table 16. If after this step there is still no
751 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 14: 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 imperative 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 toppings, 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 15: 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 16: 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.