

---

# How to Determine the Preferred Image Distribution of a Black-Box Vision-Language Model?

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1 Large foundation models have revolutionized the field, yet challenges remain in  
2 optimizing multi-modal models for specialized visual tasks. We propose a novel,  
3 generalizable methodology to identify preferred image distributions for black-  
4 box Vision-Language Models (VLMs) by measuring output consistency across  
5 varied input prompts. Applying this to different rendering types of 3D objects, we  
6 demonstrate its efficacy across various domains requiring precise interpretation  
7 of complex structures, with a focus on Computer-Aided Design (CAD) as an  
8 exemplar field. We further refine VLM outputs using in-context learning with  
9 human feedback, significantly enhancing explanation quality. To address the lack  
10 of benchmarks in specialized domains, we introduce CAD-VQA, a new dataset for  
11 evaluating VLMs on CAD-related visual question answering tasks. Our evaluation  
12 of state-of-the-art VLMs on CAD-VQA establishes baseline performance levels,  
13 providing a framework for advancing VLM capabilities in complex visual reasoning  
14 tasks across various fields requiring expert-level visual interpretation. We release  
15 the dataset and evaluation codes at <https://github.com/...>

## 16 1 Introduction

17 Large foundation models have revolutionized the AI landscape, providing unparalleled capabilities  
18 across various domains [3]. Vision-Language Models (VLMs), a subset of these models, integrate  
19 visual and textual information, enabling complex tasks such as image captioning, visual question  
20 answering, and multi-modal reasoning [4, 28]. Despite their impressive performance, a significant  
21 challenge remains: extracting the most useful knowledge from these black-box models.

22 Prompt engineering has seen extensive research and application in large language models, optimizing  
23 inputs to elicit more accurate and relevant responses [25, 18]. However, the multi-modal nature of  
24 VLMs introduces additional layers of complexity. These models must interpret and integrate informa-  
25 tion from both visual and textual inputs, and the optimal prompting strategy can vary significantly  
26 based on the image distribution [19].

27 Understanding image view distribution is crucial across various domains. In mechanical design,  
28 different views of parts and assemblies enhance comprehension of complex structures, aiding design  
29 and analysis. In architecture and construction, multiple perspectives of building designs help assess  
30 structural integrity and plan activities. In robotics and autonomous driving, diverse viewpoints  
31 improve navigation and object manipulation. In surveillance and security, integrating views from  
32 multiple cameras enhances monitoring accuracy. In medical imaging, different views of scans like  
33 MRI and CT provide comprehensive insights for diagnosing diseases, requiring models to integrate  
34 information from various angles.

35 In this work, we address the challenge of determining which image distributions lead to better outputs  
36 from a black-box VLM. Specifically, we focus on scenarios where multiple views of objects are  
37 available, such as renderings of images taken under different conditions [29]. Given that we often  
38 lack information about the data on which the VLM was trained, and do not have access to the  
39 model weights, properties, or gradients, traditional methods for assessing model confidence are not  
40 applicable [6].

41 To overcome this, we propose a novel method to measure the confidence of a VLM without requiring  
42 access to its internal parameters. Our approach involves analyzing the outputs produced by the  
43 model under different image distributions. By systematically evaluating the model’s confidence  
44 across various distributions, we can infer the image distributions that the VLM "prefers," leading to  
45 more reliable and accurate outputs. Our approach is based on the hypothesis that higher consistency  
46 in a VLM’s outputs, despite variations in input prompts, indicates higher model confidence. This  
47 hypothesis is grounded in the principle that a model with a robust internal representation of the input  
48 should produce consistent outputs even when the input is paraphrased. This aligns with recent work  
49 on self-consistency in language models [33] and relates to the concept of model calibration [13].

50 We also apply in-context learning with human feedback (ICL-HF) to refine and improve VLM outputs.  
51 By incorporating expert knowledge through iterative feedback, we demonstrate enhancements in the  
52 quality and accuracy of VLM-generated explanations for complex 3D mechanical parts. This process  
53 provides valuable insights into the learning dynamics of VLMs in specialized domains.

54 Building upon these methods, we present CAD-VQA, a new dataset specifically designed to evaluate  
55 VLMs’ understanding of 3D mechanical parts in Computer-Aided Design (CAD) contexts. This  
56 dataset, comprising carefully curated images, questions, and answers, addresses a gap in the field by  
57 providing a benchmark for assessing VLM performance in specialized technical domains.

58 The main contributions of this work are:

- 59 1. A novel method for measuring VLM confidence based on output consistency across different  
60 image distributions, without access to internal model parameters.
- 61 2. An application of in-context learning with human feedback (ICL-HF) to improve VLM performance  
62 in the specialized domain of 3D mechanical part analysis.
- 63 3. CAD-VQA: A new dataset for evaluating VLMs on CAD-related visual question answering tasks,  
64 addressing the lack of benchmarks in this domain.
- 65 4. Evaluation of state-of-the-art VLMs on the CAD-VQA dataset, establishing baseline performance  
66 levels for future research.

67 While we acknowledge that high consistency in model outputs could potentially result from model  
68 biases or limitations, rather than true confidence, we believe our approach provides a valuable  
69 proxy for assessing the reliability of (black-box) VLM outputs across different image distributions.  
70 Moreover, the combination of our consistency measurement technique, application of ICL-HF, and  
71 the CAD-VQA dataset offers a comprehensive framework for advancing the capabilities of VLMs in  
72 specialized visual reasoning tasks.

## 73 2 Related Work

74 **Prompt engineering** for large language models has been extensively explored, as demonstrated  
75 by Reynolds and McDonell [25], Liu et al. [18], and Radford et al. [24]. These studies focus on  
76 designing effective prompts to elicit desired responses from language models, thereby enhancing their  
77 utility in various applications. Recent works such as Gao et al. [11], Lester et al. [16], Wei et al. [30],  
78 and Sanh et al. [26] have further expanded on prompt engineering techniques, introducing methods  
79 like prompt tuning and instruction-based learning. However, prompt engineering for multi-modal  
80 models remains relatively underexplored, particularly in the context of image distributions and their  
81 impact on model performance.

82 **Prompt engineering for vision-language models.** While much work has been done on prompt engi-  
83 neering for language models [18], the extension to multimodal scenarios presents unique challenges.  
84 Cho et al. [7] proposed a unified framework for vision-language prompt learning, demonstrating the  
85 potential of tailored prompts in improving model performance.

86 **The complexity of evaluating black-box models** without access to their internal parameters is a  
 87 well-known challenge. Tsimpoukelli et al. [29] investigate multimodal few-shot learning with frozen  
 88 language models, addressing the difficulties in adapting pre-trained models to new tasks with limited  
 89 data. Similarly, Chen et al. [6] evaluate large language models trained on code, proposing methods  
 90 to assess model confidence and performance without direct access to model internals. Our work  
 91 builds on these foundations by addressing the specific challenge of determining preferred image  
 92 distributions for VLMs. By focusing on scenarios with multiple views of objects, such as renderings  
 93 under different conditions, we propose a novel approach to measure model confidence and optimize  
 94 input data for better outputs. This contribution aims to bridge the gap in the existing literature  
 95 on prompt engineering and evaluation for multi-modal models. Hendricks et al. [14] proposed a  
 96 probing framework to assess the grounding capabilities of VLMs, highlighting the importance of  
 97 understanding how these models integrate visual and linguistic information. Similarly, Cao et al. [5]  
 98 investigated the inner workings of VLMs, providing insights into their decision-making processes.

99 **The concept of using consistency as a measure of model performance** has gained traction in recent  
 100 years. Xu et al. [33] demonstrated that self-consistency can improve chain-of-thought reasoning in  
 101 language models, which aligns with our approach of using consistency to assess VLM outputs. In the  
 102 context of vision-language tasks, Frank et al. [10] explored the use of consistency in visual question  
 103 answering, showing how it can be leveraged to improve model accuracy.

104 **The concept of in-context learning with human feedback**, which we employ in our study, draws  
 105 inspiration from recent advancements in reinforcement learning from human feedback (RLHF)  
 106 [8, 27, 22]. While we don't use reinforcement learning directly, the principle of incorporating human  
 107 feedback to improve model outputs is similar. This approach aligns with broader trends in interactive  
 108 and iterative learning paradigms [2], as well as methods for fine-tuning language models with human  
 109 preferences [34, 31]. The integration of expert knowledge through feedback mechanisms has also  
 110 been explored in various domain-specific applications [32].

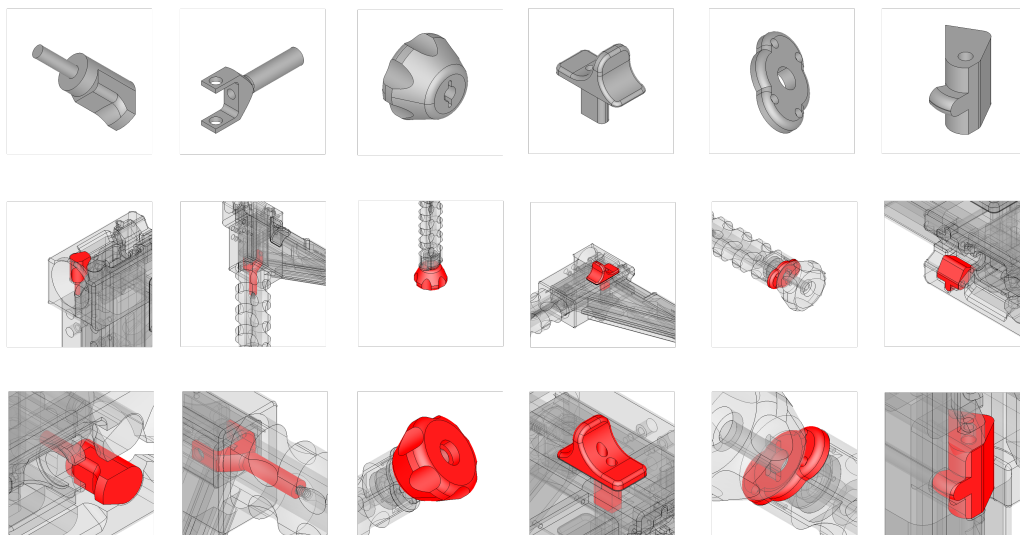


Figure 1: Sample data visualization showing different image distributions and generated explanations. First, second and third row correspond to distributions A, B, and C of the same object, respectively.

### 111 3 Method

112 In this work, we address the challenge of determining which image distributions lead to better  
 113 outputs from a black-box VLM. Specifically, we use GPT-4o [20] for our experiments, but it can  
 114 simply be replaced by any other VLM. GPT-4o is currently known for its state-of-the-art performance in  
 115 integrating visual and textual information. We focus on scenarios where multiple views of objects  
 116 are available, such as renderings of images taken under different conditions. Given that we often  
 117 lack information about the data on which the VLM was trained, and do not have access to the

118 model weights, properties, or gradients, traditional methods for assessing model confidence are not  
119 applicable.

120 Given  $N$  image distributions  $\{I_1, I_2, \dots, I_N\}$ , our goal is to determine which distribution leads to  
121 better performance when using a black-box Vision-Language Model (VLM), such as GPT-4o [20],  
122 where we do not have access to model weights, gradients, or probabilities. To achieve this, we propose  
123 a method to measure the consistency of the VLM’s outputs across different image distributions.

124 The underlying hypothesis of this methodology is that higher consistency in the VLM’s outputs,  
125 despite variations in the textual prompts, indicates higher model confidence. Model confidence refers  
126 to the certainty with which a model produces an output given an input. For a VLM, confidence can  
127 be understood as the model’s ability to generate consistent and reliable outputs despite variations in  
128 the input prompts. A robust model should produce similar outputs when presented with semantically  
129 equivalent but syntactically different prompts. This robustness indicates that the model has a stable  
130 and reliable understanding of the input image, suggesting higher confidence in its outputs.

131 Let  $P$  be the set of paraphrased prompts and  $I$  be an image distribution. For a robust and confident  
132 model, the outputs  $O_{P,I}$  should exhibit minimal variance. Formally, for paraphrased prompts  $P_1,$   
133  $P_2, \dots, P_C$ , the outputs  $O_{I,P_1}, O_{I,P_2}, \dots, O_{I,P_C}$  should be similar:

$$\text{Var}(O_{I,P_1}, O_{I,P_2}, \dots, O_{I,P_C}) \approx 0$$

134 Here, variance (Var) is a measure of inconsistency. Lower variance implies higher consistency, which  
135 can be interpreted as higher confidence.

136 **Prompt paraphrasing.** Given a textual prompt  $P$  that aims to extract information from an image,  
137 we generate  $C = 3$  different paraphrased commands  $\{P_1, P_2, P_3\}$  using the chat version of GPT-  
138 4 and manually verify them. These paraphrased commands are designed to maintain the same  
139 semantic meaning while varying the phrasing. The full prompt used for paraphrasing can be found in  
140 Appendix A.1.

141 After generation, we manually review the paraphrases to ensure they meet our criteria for semantic  
142 equivalence and diversity. Any paraphrases that deviate too far from the original meaning or don’t  
143 provide sufficient variation are replaced with manually crafted alternatives.

144 **Collecting VLM outputs.** For each image distribution  $I_n$  and each paraphrased command  $P_c$ , we  
145 collect the VLM’s output  $O_{n,c}$ :

$$O_{n,c} = \text{VLM}(I_n, P_c)$$

146 where  $n \in \{1, 2, \dots, N\}$  and  $c \in \{1, 2, \dots, C\}$ . This results in a set of outputs  
147  $\{O_{n,1}, O_{n,2}, \dots, O_{n,C}\}$  for each image distribution  $I_n$ .

### 148 3.1 Measuring Consistency

149 We measure the consistency of the outputs for each image distribution using three different methods:

150 *ROUGE and BLEU Scores.* We calculate the ROUGE [17] score i.e., ROUGE-1, ROUGE-2, ROUGE-  
151 L, and BLEU [23] scores for the outputs within each image distribution. Let  $S_{n,c}$  be the score  
152 between  $O_{n,c}$  and a reference output. The consistency score  $C_{\text{ROUGE/BLEU},n}$  for image distribution  $I_n$   
153 is defined as the average score across all paraphrased commands:

$$C_{\text{ROUGE/BLEU},n} = \frac{1}{C} \sum_{c=1}^C S_{n,c}$$

154 *BERT Embedding Cosine Similarity.* We embed each output  $O_{n,c}$  using a BERT model [9] and  
155 calculate the cosine similarity between the embeddings. Let  $\text{BERT}(O_{n,c})$  be the embedding of  $O_{n,c}$ .  
156 The consistency score  $C_{\text{BERT},n}$  for image distribution  $I_n$  is defined as the average cosine similarity  
157 between all pairs of embeddings:

$$C_{\text{BERT},n} = \frac{2}{C(C-1)} \sum_{i=1}^C \sum_{j=i+1}^C \cos(\text{BERT}(O_{n,i}), \text{BERT}(O_{n,j}))$$

158 **GPT-based Consistency Judgement.** We use GPT-4o to act as a judge and provide a consistency  
159 score for the outputs within each image distribution. The detailed prompt for consistency judgment is  
160 provided in Appendix A.2.

161 GPT-4o then provides a consistency score between 0 and 1, where 0 means completely inconsistent  
162 and 1 means perfectly consistent. Let  $G(O_{n,1}, O_{n,2}, \dots, O_{n,C})$  be the consistency score given by  
163 GPT-4o. The consistency score  $C_{\text{GPT},n}$  for image distribution  $I_n$  is:

$$C_{\text{GPT},n} = G(O_{n,1}, O_{n,2}, \dots, O_{n,C})$$

164 This approach allows us to leverage GPT-4o’s natural language understanding capabilities to assess  
165 the semantic consistency of the generated descriptions, providing a more nuanced evaluation than  
166 purely statistical methods.

167 Determining Preferred Image Distribution. Finally, we determine the preferred image distribution by  
168 comparing the consistency scores across all image distributions. The distribution with the highest  
169 average consistency score, considering all measurement methods (ROUGE/BLEU, BERT, and GPT-  
170 based), is considered the preferred distribution. This approach allows us to identify which image  
171 distribution leads to the most consistent and reliable outputs from the VLM.

## 172 3.2 Human Expert Rating and Dataset Creation

173 While consistency is a key indicator of model confidence, it is not sufficient on its own as the  
174 responses could be consistently incorrect. Therefore, we involve a mechanical expert to rate the  
175 explanations provided by the VLM for each part in different image distributions. The ratings focus on  
176 both accuracy and usefulness of the explanations. The overall expert rating results across all samples  
177 and explanations (i.e., 25 samples times 3 explanations for each) are summarized in Table 3. The  
178 criteria for expert ratings include *Relevance*, *Accuracy*, *Detail*, *Fluency*, and *Overall Quality*. We  
179 convert the options for these criteria to numerical values between 1 and 5 to calculate the values  
180 in Table 3. We also ask the human experts to add comments when necessary to provide additional  
181 insights.

182 The *relevance* and *accuracy* were evaluated by first analyzing the congruency between the name and  
183 the depicted image. A lower rating was assigned if the preliminary assessment revealed a lack of  
184 alignment. Subsequently, the rating was adjusted if the name and the content of the text did not align.  
185 A higher level of congruity indicated higher accuracy. From there, the contents were assessed for their  
186 ability to accurately describe the component design features, characteristics, industry, intended use,  
187 etc. The *detail* evaluation was assessed based on whether the provided data sufficed to conceptualize  
188 the design. *Fluency* was gauged by the grammatical correctness and the coherence of the descriptions.  
189 The *overall quality* was determined by the total of the scores from the indicated categories. While  
190 evaluating the different categories, an emerging trend was noticed. If the visual language model  
191 correctly identified the object’s name, the subsequent details tended to align correctly. However,  
192 when the model misidentified the geometry, the details tended to correspond to the wrong item  
193 identification. For parts that were highly specialized for assembly, a more general example of industry  
194 standards was often indicated, rather than a specific standard as a starting point for further analysis by  
195 the end user.

196 From top-rated explanations, we developed a specialized dataset comprising CAD images paired with  
197 questions and answers extracted from the explanations. This dataset is designed to evaluate VLMs  
198 on visual question answering (VQA) tasks specific to CAD objects. By grounding our dataset in  
199 expert-validated explanations, we provide a reliable benchmark for assessing VLM performance in  
200 the CAD domain, bridging the gap between consistency and domain-specific accuracy.

## 201 4 CAD-VQA Dataset

202 We present CAD-VQA (Computer-Aided Design Visual Question Answering), a novel dataset  
203 designed to evaluate Vision-Language Models’ understanding of 3D mechanical parts in CAD  
204 contexts.

### 205 4.1 Dataset Creation Process

206 Building upon the high-quality explanations generated through our iterative process of VLM output  
207 and human expert evaluation, we developed a novel dataset for evaluating Vision-Language Models  
208 on CAD tasks. The dataset creation process involved the following steps:

209 *Selection of top-rated explanations:* We chose explanations for 17 parts that received excellent ratings  
 210 from human experts.

211 *Question generation:* Using Claude 3.5 Sonnet, we generated an initial set of questions based on  
 212 these top-rated explanations. The questions cover various aspects including part names, geometrical  
 213 features, assembly features, and functionality.

214 *Visual focus:* We designed questions to require analysis of the provided images, ensuring that answers  
 215 couldn't be derived solely from common knowledge of 3D design.

216 *Comprehensive coverage:* A total of 85 multiple-choice questions were created, providing a diverse  
 217 range of queries about the 17 selected parts.

218 *Quality assurance:* We conducted rigorous post-processing to ensure consistency in question style,  
 219 eliminate errors, and maintain a uniform difficulty level across the dataset.

220 This dataset addresses a gap in the field of VLM evaluation for CAD applications. Currently, there is  
 221 a scarcity of publicly available datasets specifically designed to assess VLMs' understanding of 3D  
 222 mechanical parts and their features. Our dataset, while compact, represents one of the first efforts to  
 223 create a benchmark for evaluating VLMs in the context of CAD and mechanical engineering.

224 The uniqueness of this dataset lies in its focus on:

- 225 • Specialized vocabulary and concepts from mechanical engineering and CAD
- 226 • Visual interpretation of 3D parts from multiple perspectives
- 227 • Understanding of both individual part features and their roles in larger assemblies
- 228 • Application of domain-specific knowledge to answer questions based on visual input

229 By providing this dataset, we aim to stimulate further research in improving VLMs' capabilities in  
 230 specialized technical domains, particularly in the field of mechanical design and engineering.

231 To illustrate the nature of our CAD-VQA dataset, we provide a few representative examples in Table 1.  
 232 These examples demonstrate the diversity of questions and the necessity of properly analyzing the  
 233 provided images to correctly answer them.

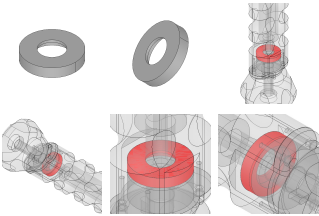
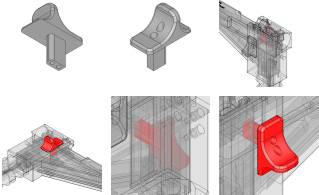
Image	Question and Options
	<p><b>Q:</b> <i>Based on the visual representation, what role does the washer likely play in the assembly?</i></p> <hr/> <p>A) It acts as a pivot point, B) It provides electrical insulation, C) It distributes the load of a fastener, D) It serves as a decorative element, E) It functions as a seal, F) It acts as a heat sink, G) It provides cushioning, H) It serves as a wear surface, K) It functions as a locking device, J) Both C and H are correct</p>
	<p><b>Q:</b> <i>Looking at the 2D images, which of the following names best describes this part?</i></p> <hr/> <p>A) Gear Assembly, B) Piston Rod, C) Bracket Mount, D) Camshaft, E) Flywheel, F) Crankshaft, G) Bracket with Mounting Holes, H) Valve Cover, K) Both C and G are correct, J) Timing Belt</p>

Table 1: Sample data points from the CAD-VQA dataset

234 **5 Results**

235 For our preliminary experiments, we use a relatively small dataset due to the difficulty in scaling  
 236 the rating process of detailed explanations by mechanical experts. Our dataset consists of 25 3D  
 237 mechanical parts from the ABC collection [15], each part appearing within a larger assembly context.  
 238 We evaluate four different image distributions for rendering these parts. *Distribution A*: Each part is  
 239 rendered as an individual solid. *Distribution B*: Each part is rendered in the assembly along with other  
 240 parts, where the other parts are transparent. *Distribution C*: Similar to Distribution B but slightly  
 241 zoomed. *Distribution D*: A mix of Distributions A, B, and C (two samples from each).

242 These distributions were chosen to cover a range of contexts, although many other rendering methods  
 243 are possible. For each part, we generate three different paraphrased prompts aimed at explaining the  
 244 part’s function and significance within the assembly. A sample of how the data looks is shown in  
 245 Figure 1.

246 **Consistency Measurement.** We measure the consistency of the outputs using the methods described  
 247 previously: ROUGE and BLEU scores, BERT embedding cosine similarity, and GPT-based consi-  
 248 stency judgment. The results for each image distribution are summarized in Table 2. The results  
 249 indicate that Distribution D, which includes a mix of the different rendering methods, consistently  
 250 achieves the highest scores in both consistency metrics and expert ratings. This suggests that provid-  
 251 ing multiple perspectives of the parts helps the VLM generate more accurate and reliable explanations.  
 252 Additionally, the use of in-context learning with expert feedback shows a noticeable improvement in  
 253 the quality of the explanations, demonstrating the effectiveness of iterative refinement in enhancing  
 254 model performance.

Metric	Distribution A	Distribution B	Distribution C	Distribution D
ROUGE-1	0.4831 $\pm$ 0.0483	0.4479 $\pm$ 0.0606	0.4569 $\pm$ 0.0747	<b>0.5159</b> $\pm$ 0.0609
ROUGE-2	0.1398 $\pm$ 0.0277	0.1298 $\pm$ 0.0369	0.1326 $\pm$ 0.0313	<b>0.2055</b> $\pm$ 0.034
ROUGE-L	0.2324 $\pm$ 0.0238	0.2267 $\pm$ 0.0307	0.2287 $\pm$ 0.0283	<b>0.2916</b> $\pm$ 0.027
BLEU	0.0874 $\pm$ 0.0176	0.0837 $\pm$ 0.0213	0.0865 $\pm$ 0.0173	<b>0.1613</b> $\pm$ 0.0216
Cosine Similarity	<b>0.8988</b> $\pm$ 0.0289	0.8902 $\pm$ 0.0401	0.8756 $\pm$ 0.055	0.8887 $\pm$ 0.041
GPT Score	0.6212 $\pm$ 0.2403	0.4365 $\pm$ 0.2716	0.4269 $\pm$ 0.2207	<b>0.6308</b> $\pm$ 0.2504
<b>Average</b>	0.4104 $\pm$ 0.0644	0.3691 $\pm$ 0.0769	0.3679 $\pm$ 0.0712	<b>0.4490</b> $\pm$ 0.0723

Table 2: Consistency scores across different image distributions. For all distributions, we randomly select 5 images rendered from various angles. For Distribution D ("All"), these 5 images are a mix drawn from the other three distributions.

255 **5.1 In-Context Learning with Human Feedback**

256 To further refine the model’s performance, we use the expert ratings as feedback for in-context  
 257 learning. The VLM is shown the expert ratings to learn and correct the explanations that received  
 258 lower scores. After incorporating this feedback, we re-evaluate the model with human experts to  
 259 assess improvement. The updated ratings are shown in Table 3.

260 Based on our consistency scores, Distribution D (a mix of single object renders, assembly renders with  
 261 transparent parts, and zoomed assembly renders) performed best. We apply an in-context learning  
 262 process to our dataset, using a prompt that provides the model with images, descriptions, and expert  
 263 ratings for each part. The full in-context learning prompt can be found in Appendix A.3.

264 For our current dataset of parts, we provide GPT-4o with a comprehensive prompt containing all  
 265 parts’ information simultaneously: images from Distribution D for each part, descriptions per part,  
 266 and their corresponding human expert ratings. The model then generates new descriptions for parts  
 267 based on this extensive in-context learning.

268 However, for larger datasets where providing all information at once may exceed the model’s context  
 269 length, we suggest two alternative approaches: a Sliding Window Approach and a Sequential Process-  
 270 ing Approach. Details of these approaches and a visual comparison can be found in Appendix A.4.

<b>Metric</b>	<b>Before ICL-HF <math>\uparrow</math></b>	<b>After ICL-HF <math>\uparrow</math></b>
Relevance	3.88 $\pm$ 1.34	<b>3.96 <math>\pm</math> 1.27</b>
Accuracy	3.98 $\pm$ 0.80	<b>4.10 <math>\pm</math> 0.75</b>
Detail	4.14 $\pm$ 0.69	<b>4.16 <math>\pm</math> 0.68</b>
Fluency	4.06 $\pm$ 0.75	<b>4.10 <math>\pm</math> 0.73</b>
Overall Quality	4.07 $\pm$ 0.79	<b>4.14 <math>\pm</math> 0.73</b>

Table 3: Expert ratings across all samples and explanations before and after in-context learning. Score range is 1 to 5. ICL-HF refers to in-context learning with human feedback.

## 271 5.2 Performance of State-of-the-Art VLMs on our CAD-VQA dataset

272 We evaluated several state-of-the-art Vision-Language Models on our CAD-VQA dataset to establish  
 273 baseline performance levels. The models tested include Claude-3.5-Sonnet [1], OpenAI’s GPT-4o[21]  
 274 and O1-preview, and Gemini-1.5-Pro [12].

275 Table 4 presents the accuracy of each model on our dataset:

<b>Model</b>	<b>Accuracy (%)</b>
Claude-3.5-Sonnet	<b>61.17</b>
Gemini-1.5-Pro	54.12
GPT-4o	54.11
O1-preview	42.35

Table 4: Performance of state-of-the-art VLMs on the CAD-VQA dataset

276 These results demonstrate that even the most advanced VLMs face significant challenges in accurately  
 277 interpreting and reasoning about CAD objects. Claude-3.5-Sonnet shows the highest accuracy  
 278 at 61.17%, while Gemini-1.5-Pro achieves 54.12% accuracy. These scores, while above random  
 279 guessing (10% for 10-option multiple choice questions), indicate substantial room for improvement  
 280 in VLMs’ understanding of specialized technical domains like mechanical engineering and CAD.

281 The performance gap between these models and human experts underscores the need for continued  
 282 research and development in enhancing VLMs’ capabilities in domain-specific visual reasoning  
 283 tasks.

## 284 6 Conclusion

285 Our study addressed the challenge of optimizing image distributions for black-box Vision-Language  
 286 Models (VLMs). Experimenting with 3D mechanical parts and GPT-4o, we evaluated four image  
 287 distributions using a novel methodology based on output consistency across paraphrased prompts.  
 288 The mixed distribution, combining various rendering perspectives, consistently outperformed others,  
 289 indicating that multiple viewpoints enhance VLM performance in generating accurate explanations.  
 290 Expert ratings validated these findings and demonstrated the effectiveness of in-context learning  
 291 with human feedback in improving explanation quality. Building on these insights, we developed  
 292 CAD-VQA, a new dataset for evaluating VLMs on CAD-related visual question answering tasks.  
 293 This dataset addresses a gap in the field and provides a benchmark for assessing VLM performance  
 294 in specialized technical domains.

295 Our approach of automated consistency checks, followed by expert evaluation, offers a scalable  
 296 method for assessing VLM outputs. The evaluation of state-of-the-art VLMs on CAD-VQA estab-  
 297 lishes baseline performance levels, highlighting both the potential and current limitations of VLMs  
 298 in interpreting specialized visual data. While our experiments focused on CAD applications, this  
 299 methodology and the principles behind CAD-VQA are broadly applicable to other domains requiring  
 300 specialized visual interpretation. Future work should explore scaling this approach to diverse fields,  
 301 applying the dataset creation process to other specialized domains, and investigating the relationship  
 302 between output consistency and model confidence through comparison with explicit confidence  
 303 estimation techniques and human evaluations.



## 304 References

- 305 [1] Anthropic. Claude-3-sonnet. <https://www.anthropic.com>, 2024. Accessed: 2024-08-27.
- 306 [2] Surya Bhupatiraju, Hao Hu, Karm Singh, Nate Brown, and Jared Kaplan. Interactive language  
307 learning by question answering. *arXiv preprint arXiv:2201.08540*, 2022.
- 308 [3] Rishi Bommasani et al. On the opportunities and risks of foundation models. *arXiv preprint*  
309 *arXiv:2108.07258*, 2021.
- 310 [4] Tom Brown et al. Language models are few-shot learners. In *Advances in Neural Information*  
311 *Processing Systems*, volume 33, 2020.
- 312 [5] Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, and Yen-Chun Chen. Behind the scene: Revealing  
313 the secrets of pre-trained vision-and-language models. In *European Conference on Computer*  
314 *Vision*, pages 565–580. Springer, 2020.
- 315 [6] Mark Chen et al. Evaluating large language models trained on code. *arXiv preprint*  
316 *arXiv:2107.03374*, 2021.
- 317 [7] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text  
318 generation. *arXiv preprint arXiv:2102.02779*, 2021.
- 319 [8] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep  
320 reinforcement learning from human preferences. *Advances in neural information processing*  
321 *systems*, 30, 2017.
- 322 [9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of  
323 deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*,  
324 2018.
- 325 [10] Stella Frank, Pranava Madhyastha, Lucia Barbu, and Matthias Buch-Kromann. Vision-and-  
326 language or vision-for-language? on cross-modal influence in multimodal transformers. *arXiv*  
327 *preprint arXiv:2109.04448*, 2021.
- 328 [11] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot  
329 learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational*  
330 *Linguistics (ACL)*, 2021.
- 331 [12] Google. Gemini-1.5-pro. <https://deepmind.google/technologies/gemini/>, 2024. Ac-  
332 cessed: 2024-08-27.
- 333 [13] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural  
334 networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.
- 335 [14] Lisa Anne Hendricks and Aishwarya Arnab. Probing image–language transformers for verb  
336 understanding. *arXiv preprint arXiv:2106.09141*, 2021.
- 337 [15] Sebastian Koch, Albert Matveev, Zhongshi Jiang, Francis Williams, Alexey Artemov, Evgeny  
338 Burnaev, Marc Alexa, Denis Zorin, and Daniele Panozzo. Abc: A big cad model dataset for  
339 geometric deep learning. In *Proceedings of the IEEE/CVF conference on computer vision and*  
340 *pattern recognition*, pages 9601–9611, 2019.
- 341 [16] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient  
342 prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.
- 343 [17] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization*  
344 *Branches Out*, pages 74–81, 2004.
- 345 [18] Pengfei Liu et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in  
346 natural language processing. *arXiv preprint arXiv:2107.13586*, 2021.
- 347 [19] Jiasen Lu et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-  
348 language tasks. In *NeurIPS*, 2019.
- 349 [20] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- 350 [21] OpenAI. Gpt-4o. <https://openai.com/research/gpt-4>, 2023. Accessed: 2024-08-27.
- 351 [22] Long Ouyang et al. Training language models to follow instructions with human feedback.  
352 *arXiv preprint arXiv:2203.02155*, 2022.

- 353 [23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic  
354 evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for*  
355 *computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.
- 356 [24] Alec Radford et al. Learning transferable visual models from natural language supervision.  
357 *arXiv preprint arXiv:2103.00020*, 2021.
- 358 [25] Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond  
359 the few-shot paradigm. *arXiv preprint arXiv:2102.07350*, 2021.
- 360 [26] Victor Sanh et al. Multitask prompted training enables zero-shot task generalization. *arXiv*  
361 *preprint arXiv:2110.08207*, 2022.
- 362 [27] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec  
363 Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback.  
364 In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021, 2020.
- 365 [28] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from  
366 transformers. In *EMNLP*, 2019.
- 367 [29] Maria Tsimpoukelli et al. Multimodal few-shot learning with frozen language models. In  
368 *NeurIPS*, 2021.
- 369 [30] Jason Wei et al. Chain-of-thought prompting elicits reasoning in large language models. In  
370 *Advances in Neural Information Processing Systems*, 2022.
- 371 [31] Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and  
372 Paul F Christiano. Recursively summarizing books with human feedback. *arXiv preprint*  
373 *arXiv:2109.10862*, 2021.
- 374 [32] Yuqing Wu, Siqi Zhao, Hao Chen, Yifan Zhang, Zhijie Gao, Guan Zheng, Shihan Yan, Yuxuan  
375 Lin, Zhaowei Peng, Jingwei Jiang, et al. Human preference optimization: Economic decision  
376 making for ai alignment. *arXiv preprint arXiv:2310.03026*, 2023.
- 377 [33] Xuezhi Xu, Yifei Wang, Junyi Zhang, Yixin Chen, Heng-Tze Cheng, Jason Wei, Zhitao Zhao,  
378 Kathy Lee, Mohammad Shoeybi, Mostafa Dehghani, et al. Self-consistency improves chain of  
379 thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.
- 380 [34] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei,  
381 Paul F Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences.  
382 *arXiv preprint arXiv:1909.08593*, 2019.

383 **A Supplementary Material**

384 **A.1 Paraphrasing Prompt**

385 The following prompt was used to generate paraphrases for our experiments:

**Paraphrasing Prompt**

Please generate 3 paraphrases of the following prompt. Each paraphrase should maintain the same core meaning but vary in phrasing and complexity. Ensure a mix of minor variations (e.g., word order changes, synonym substitution) and more significant restructuring. The paraphrases should be diverse enough to test a language model’s robustness to input variations, but not so different that they alter the fundamental query.

**Original prompt:**

“Please analyze the object shown in the image. Note that in some images, the 3D part might appear red when shown in an assembly format, while in others, it might look grey when presented as an individual part. Provide a detailed explanation of the object’s name or type, its geometric features and shape, and its likely function or purpose within a larger system or assembly. Be as specific and comprehensive as possible in your description.”

Generate your 3 paraphrases below:

1. [Paraphrase 1] 2. [Paraphrase 2] 3. [Paraphrase 3]

386

387 **A.2 Consistency Judgment Prompt**

388 The following prompt was used for GPT-based consistency judgment:

**Consistency Judgment Prompt**

You are tasked with evaluating the consistency of multiple descriptions of the same 3D mechanical part. These descriptions were generated by an AI model in response to slightly different prompts about the same image. Your job is to assess how consistent these descriptions are with each other in terms of content, details, and overall interpretation of the part.

Please consider the following aspects:

1. Name/Type Consistency: Do all descriptions refer to the part using the same or very similar names/types?
2. Geometric Features Consistency: Are the descriptions of the part’s shape, size, and key geometric features consistent across all versions?
3. Functionality Consistency: Do all descriptions attribute the same or very similar functions or purposes to the part?
4. Detail Level Consistency: Is the level of detail provided about the part similar across all descriptions?
5. Context Consistency: If the part’s position or role within a larger assembly is mentioned, is this consistent across descriptions?

After analyzing the descriptions, please provide:

1. A consistency score from 0 to 1, where 0 means completely inconsistent and 1 means perfectly consistent.
2. A brief explanation (2-3 sentences) justifying your score.

Descriptions to evaluate: 1. [Description 1] 2. [Description 2] 3. [Description 3]

Your consistency score and explanation: [Score]: [Explanation]:

389

390 **A.3 In-Context Learning with Human Feedback Prompt**

391 The following prompt was used for in-context learning with human feedback:

### ICL-HF Prompt

You are an AI assistant specializing in describing 3D mechanical parts. You will be provided with information for different parts. For each part, you will receive:

1. Five images (various perspectives of the part) 2. Three descriptions of the part 3. Human expert ratings for each description

Analyze this information and generate improved descriptions. Here's the format for each part:

#### Part 1

[Image 1], ... , [Image 5]

**Description 1** [Description text] Relevance: [ ] Accuracy: [ ] Detail: [ ] Fluency: [ ] Overall: [ ]

**Description 2** [Description text] Relevance: [ ] Accuracy: [ ] Detail: [ ] Fluency: [ ] Overall: [ ]

**Description 3** [Description text] Relevance: [ ] Accuracy: [ ] Detail: [ ] Fluency: [ ] Overall: [ ]

..., **Part 25**, ...

According to the ratings, generate an improved description that:

- Accurately identifies and names the part
- Describes its geometric features and shape in detail, referencing specific views from the five images
- Explains its likely function or purpose within a larger system or assembly
- Maintains consistency with the high-rated aspects of previous descriptions
- Improves upon areas that received lower ratings
- Integrates information from all provided perspectives

Your new description should aim to maximize all five rating categories: Relevance, Accuracy, Detail, Fluency, and Overall Quality.

Please provide your improved description.

392

#### 393 A.4 Alternative Approaches for Large Datasets

394 For larger datasets where providing all information at once may exceed the model's context length,  
395 we suggest two alternative approaches:

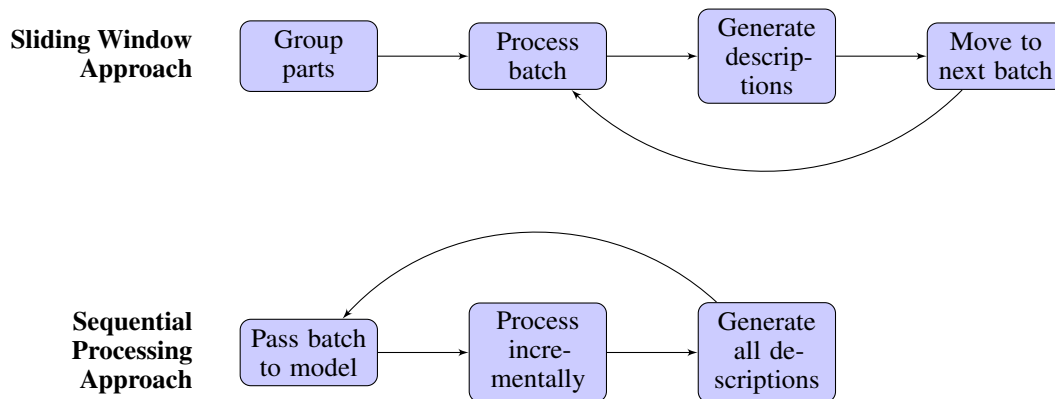


Figure 2: Visual comparison of algorithms for processing large datasets

396 These methods allow the model to learn from a substantial amount of context while remaining within  
397 practical limits. The Sliding Window Approach processes the data in overlapping batches, while  
398 the Sequential Processing Approach passes batches to the model incrementally before generating all  
399 descriptions at once.