

# Benchmarking VLMs’ Reasoning About Persuasive Atypical Images (Supplementary Material)

Sina Malakouti<sup>1,\*</sup> Aysan Aghazadeh<sup>1,\*</sup> Ashmit Khandelwal<sup>2</sup> Adriana Kovashka<sup>1</sup>

<sup>1</sup> University of Pittsburgh <sup>2</sup>BITS Pilani

{sem238, aya34}@pitt.edu f20200980@goa.bits-pilani.ac.in kovashka@cs.pitt.edu

## 1. Overview

We aimed to investigate the effectiveness of VLMs for understanding persuasive advertisement. Concretely, we hypothesized that understanding atypicality can aid understanding advertisements. Hence, we first compared state-of-the-art VLMs on three novel atypicality understanding tasks: (1) Multi-label Atypicality Classification (MAC), (2) Atypicality Statement Retrieval (ASR), and (3) Atypicality Object Recognition (AOR). Table 1 compares the performance of VLMs with our proposed strategies on the MAC task, offering a more comprehensive evaluation of metrics than Table 1 in the main paper. Table 2 summarizes the results on the small-set for the AOR task. Full-set results on ASR and AOR tasks can be found in Table 1 and Table 2 of the main paper.

Secondly, to evaluate the impact of atypicality in ad understanding and analyze VLMs’ reasoning ability about atypicality, we proposed a novel atypicality-aware verbalization method. We compared our method with VLMs and verbalization baselines (i.e.  $V + T$ ). Table 5 compares various methods of constructing atypicality-aware verbalization, including concatenation and LLM-based combinations, when used with CLIP. We also benchmark these against the CLIP ( $I$ ) baseline and a related zero-shot for KAFA (CLIP ( $I + T$ )). Full-set results on ARR are in Table 3 in the main paper. Table 6 ablate different types of verbalization and shows effectiveness of each component in our proposed verbalization method, which is discussed in Sec.2.2 and Sec.5.3 in the main paper. Table 7 shows the evaluation of the our method’s generalization to the typical images. In Table 8, we evaluated LLaVA and our method on WHOOPS! [2]. We further provide analysis for the validation of the generated semantically hard negatives by GPT-4 in Sec. 2.2 (analysis are in text). An example of our full pipeline for multi action-reason retrieval tasks is demonstrated in Fig. 4.

Figs. 1 and 3 visualize examples of semantically hard negatives and a comparison between the predictions of our proposed method and LLaVA, respectively. Finally, the

Model	Verb.	AUC-ROC		AUC-PR		Subset-Acc	
		✓	×	✓	×	✓	×
LLaVA	-	50.16	50.12	35.81	30.46	0.94	2.83
InstructBLIP	-	50.13	50.03	35.81	30.44	0.51	1.54
Vicuna	$T + V$	50.63	50.31	36.03	30.53	3.51	6.68
	$IN$	<b>52.26</b>	52.25	<b>36.84</b>	<b>31.50</b>	4.28	7.88
	$UH$	52.03	<b>52.26</b>	36.64	31.40	<b>5.22</b>	<b>10.70</b>
GPT-3.5	$T + V$	52.40	51.83	37.01	31.34	<b>10.10</b>	<b>24.32</b>
	$IN$	53.28	52.71	37.69	32.13	4.20	9.08
	$UH$	<b>54.36</b>	<b>54.64</b>	<b>38.34</b>	<b>33.17</b>	7.62	20.89
GPT 4	$T + V$	51.10	50.91	36.34	30.94	1.71	3.68
	$IN$	54.13	53.88	38.46	33.16	4.79	9.85
	$UH$	<b>55.51</b>	<b>56.00</b>	<b>39.32</b>	<b>34.41</b>	<b>11.22</b>	<b>28.00</b>

Table 1. **Multi-label atypicality classification on Full-set.** ✓/× denotes performance with/without No Atypicality (NA) class. **Bolded** numbers indicate best-performing strategy per LLM.

Model	Avg. similarity score	% of scores		
		> 0.7	> 0.6	> 0.5
BLIP2 [9]	0.45	8.13	19.11	36.59
InstructBLIP [6]	0.47	10.57	23.58	43.90
MiniGPT4 [13]	0.52	15.45	31.71	56.50
LLaVA [11]	0.60	29.79	56.17	69.36
GPT-4V [1]	<b>0.67</b>	<b>46.94</b>	<b>61.63</b>	<b>77.14</b>

Table 2. **Atypical Object Recognition (AOR) on Small-set.** MP-Net sentence similarity scores and score thresholds are reported.

prompts utilized in this study are detailed in Sec. 3.

## 2. Results

### 2.1. Atypicality Understanding Results

**Multi-label Atypicality Classification and Atypicality Statement Retrieval.** Table 1 presents additional evaluation metrics (i.e., AUC-ROC, AUC-PR, and Subset-Acc) compared to the main paper’s table. We observe that  $UH$  consistently outperforms all other strategies in AUC-ROC when excluding NA (denoted as ×). Similarly, in most LLMs,  $UH$  surpasses both  $IN$  and  $V + T$ , showcasing the effectiveness of  $UH$  in highlighting the unusualness of the image to facilitate understanding of atypicality. A significant difference is observed between the performance of

Classifier	Verb.	Precision@k			
		k=1	k=2	k=3	avg
LLaVA	$I$	59.67	38.27	26.06	41.33
	$I (CoT)$	66.53	42.24	28.29	45.68
Vicuna	$\mathcal{T}_V + \hat{s}_{IN}$ (Ours)	<b>71.77</b>	<b>46.77</b>	<b>31.59</b>	<b>50.04</b>
GPT-4	$\mathcal{T}_V + \hat{s}_{IN}$ (Ours)	<b>96.77</b>	<b>87.77</b>	<b>73.65</b>	<b>86.06</b>
	$\mathcal{T}_V + \hat{s}_{IN}(CoT)$	95.97	86.29	72.17	84.81

Table 3. Chain-of-thought prompting for ARR on Small-set

Classifier	Verb.	Precision@k			
		k=1	k=2	k=3	avg
LLaVA 1.6	$I$	74.79	52.00	35.73	54.17
Vicuna	$\mathcal{T}_V$ (Ours)	<b>86.40</b>	<b>62.40</b>	<b>43.19</b>	<b>63.99</b>
InternVL2-8B	$I$	91.12	75.40	55.64	74.05
InternLM	$\mathcal{T}_V$ (Ours)	<b>93.60</b>	<b>78.20</b>	<b>57.20</b>	<b>76.33</b>

Table 4. Additional VLMs for ARR on Small-set. InternLM is ‘InternLM2-5-7b-chat’.

LLMs on  $UH$  and VLMs on subset-acc. Subset-acc is a challenging metric where a prediction is considered correct only if it can successfully identify all atypicalities of the image. For instance,  $UH$  on GPT-4 achieves 28% accuracy, improving LLaVA and InstructBLIP by 25.17 and 26.46 percent, respectively. This underscores the limitations of VLMs in directly recognizing atypicality.

**Atypical Object Retrieval.** Table 2 compares current state-of-the-art VLMs on the Atypical Object Recognition (AOR) task. GPT-4V achieves the avg. similarity score of 0.67 between generated statement  $\hat{s} = (a^+, \hat{o}^p, \hat{o}^s)$  and ground-truth statement  $s = (a^+, o^{+p}, o^{+s})$  with 46.94% of the scores above 0.7. This is significantly higher than public LLMs, led by LLaVA, where only 29.79 scores are higher than 0.7. While these results show that GPT-4V is more powerful than public VLMs, it is still limited in accurately recognizing the first/second objects and the atypical relationship among them.

## 2.2. Action-Reason Retrieval Results

**Comparison against Chain-of-Thought.** Table 3 compares our proposed atypicality-aware verbalization against CoT (‘think step-by-step’) in [8]. While CoT reasoning yields marginal improvements in LLaVA due to the multi-step nature of the problem, it still falls short when compared to our approach, with significant differences of 4.36 in Vicuna and 40.38 in GPT-4.

We also observed that applying CoT on top of our method in GPT-4 results in lower performance. This happens because our approach already includes a form of implicit reasoning similar to CoT. Adding explicit CoT reasoning creates redundancy, which complicates the reasoning process and may introduce unnecessary steps. This overlap leads to the performance drop, as the extra reasoning adds complexity without improving results.

**Comparison against more VLMs** In Tab. 4, we compare our method with two state-of-the-art VLMs: LLaVA 1.6 [10] and InternVL2-8B [4]. We utilized the lan-

Classifier	Precision@k			Top-k Acc		
	k=1	k=2	k=3	k=1	k=2	k=3
CLIP( $I$ )	61.04	33.86	22.66	23.72	44.61	61.04
CLIP( $I + T$ )	46.15	24.36	16.24	15.15	31.25	46.15
CLIP( $I + T + V$ ) (Ours)	70.46	39.17	26.11	29.79	53.08	70.46
CLIP( $I + T + V + IN + UH$ ) (Ours)	<b>72.35</b>	<b>41.05</b>	<b>27.40</b>	<b>32.11</b>	<b>54.37</b>	<b>72.35</b>
CLIP( $I + \mathcal{T}_V$ ) (Ours)	63.53	34.25	22.83	24.14	45.38	63.53


Table 5. Evaluation of CLIP-based models on Full-set. Bolded numbers indicate the best performing model.


Classifier	Verb.	Multi					Single	
		Precision@k			Top-k Acc		Avg	Acc
		k=1	k=2	k=3	k=1	k=2		
Vicuna	$V + T$	64.11	41.53	27.69	24.19	45.56	34.62	44.35
	$IN$	64.92	43.55	29.17	<b>24.60</b>	43.55	41.16	45.56
	$UH$	60.89	38.71	25.94	20.97	40.32	37.37	37.90
Ours (Vicuna)	$V + T + IN + UH$	69.35	45.33	30.88	23.80	45.21	42.91	48.39
	$\mathcal{T}_V \setminus UH$	69.35	44.35	29.57	22.98	45.56	42.36	48.39
	$\mathcal{T}_V$	71.37	<b>46.77</b>	31.45	23.39	45.16	43.63	46.37
	$\mathcal{T}_V + \hat{s}_{IN}$	<b>71.77</b>	<b>46.77</b>	<b>31.59</b>	23.79	<b>46.77</b>	<b>44.14</b>	<b>48.79</b>
GPT-3.5	$V + T$	85.43	59.51	40.62	46.56	68.02	60.02	72.46
	$IN$	89.07	64.78	45.48	<b>65.99</b>	79.76	69.01	77.41
	$UH$	84.62	58.91	40.89	52.63	70.45	61.10	76.61
Ours (GPT-3.5)	$V + T + IN + UH$	90.32	64.92	45.43	48.39	74.60	64.73	74.39
	$\mathcal{T}_V \setminus UH$	91.09	66.81	46.55	62.75	78.94	69.23	78.54
	$\mathcal{T}_V$	<b>91.90</b>	<b>67.61</b>	<b>46.96</b>	63.15	79.75	69.87	<b>78.94</b>
	$\mathcal{T}_V + \hat{s}_{IN}$	<b>91.90</b>	67.20	46.69	<b>65.99</b>	<b>81.78</b>	<b>70.71</b>	72.87
GPT-4	$V + T$	92.71	84.62	72.47	84.55	89.52	84.77	95.55
	$IN$	89.92	78.23	64.65	77.42	85.08	79.06	93.88
	$UH$	80.41	63.67	50.48	62.04	72.65	64.85	95.08
Ours (GPT-4)	$V + T + IN + UH$	96.37	86.49	72.45	72.18	89.11	83.32	96.37
	$\mathcal{T}_V \setminus UH$	94.34	85.63	73.42	84.62	90.28	85.66	88.21
	$\mathcal{T}_V$	96.77	87.30	<b>74.60</b>	84.96	91.46	87.01	<b>96.77</b>
	$\mathcal{T}_V + \hat{s}_{IN}$	96.77	<b>87.77</b>	73.65	<b>87.09</b>	<b>91.54</b>	<b>87.36</b>	96.36
	$\mathcal{T}_V + \hat{s}_{\mathcal{T}_V}$	<b>97.17</b>	86.99	73.55	85.02	91.50	86.85	96.76

Table 6. ARR on Small-set. Best result per LLM/column is bolded. ‘Multi’ means we ask the LLM for multiple outputs, ‘Single’ for one.

guage models from these models (InternLM [3] against InternVL2-8B and Vicuna-13B [5] against LLaVA 1.6) to retrieve correct action-reason statements based on descriptions generated by LLaVA 1.5 and LLaVA 1.6 respectively. The results show that our approach, using InternLM, outperforms InternVL2-8B, even when using LLaVA 1.5 verbalization, and Vicuna-13B when using LLaVA 1.6 verbalization, outperforms LLaVA 1.6.

**CLIP ablation.** Table 5 demonstrates different verbalization strategies impact on CLIP zero-shot model. We observe that in contrast to Table 6 and Table 3 in the main paper, where the best results are mostly based on  $\mathcal{T}_V$ , simple concatenation (i.e.,  $U + T + IN + UH$ ) achieves the best performance on CLIP. This can be due to the more fine-grained (even noisy) information in  $T + V + IN + UH$ . Therefore, CLIP that has shown to have bag-of-words behavior [12] performs better when more information, such as object names, relations, etc., are explicitly noted. However, our proposed LLM-based approaches have more reasoning capabilities. Thus, a less noisy and more unified description in  $\mathcal{T}_V$  is a more suitable verbalization strategy.

	Correct Option	I should drink beer more often Because it would make me feel good
	Action Alter	I should <b>abstain from beer</b> because it would make me feel good.
	Reason Alter	I should drink beer more often <b>because it would make me feel bad.</b>
	Object Swap	I shouldn't drink <b>water</b> more often Because it would make me feel good
	Statement Alter	I should drink beer more often because <b>it enhances my physical fitness.</b>
Adjective Alter	I should <b>avoid</b> beer more often because it would make me feel <b>terrible.</b>	

	Correct Option	I should drink absolut vodka Because this vodka is like an island paradise
	Action Alter	I should <b>abstain from Absolut vodka</b> because this vodka is like an island paradise.
	Reason Alter	I should drink absolut vodka <b>because this vodka is nothing like an island paradise.</b>
	Object Swap	I should drink <b>coconut water</b> Because this vodka is like an island paradise
	Statement Alter	I should drink absolut vodka because this vodka is like a <b>winter blizzard.</b>
Adjective Alter	I should <b>avoid</b> absolut vodka because this vodka is like a <b>deserted</b> island.	


	Correct Option	I should drink red bull Because it will help me work hard
	Action Alter	I should <b>avoid drinking red bull</b> because it will help me work hard.
	Reason Alter	I should drink red bull <b>because it will hinder my ability to work hard.</b>
	Object Swap	I should drink <b>water</b> Because it will help me work hard
	Statement Alter	I should drink <b>water</b> because it will help me sleep well.
Adjective Alter	I should drink red bull because it will help me work <b>lazily.</b>	

Figure 1. For each correct action-reason statement, we construct 5 different types of hard negatives: (1) Action Alter, (2) Reason Alter, (3) Object Swap, (4) Statement Alter, and (5) Adjective Alter. **Green** denotes correct action-reason statements. **Red** indicates generated wrong phrases/statements.

**Hard Negative Validation.** To ensure the quality of the generated hard negatives using GPT-4, we sampled 100 images and had 5 human annotators classify each option (options constitute both ground-truth action-reason statements and the generated hard negative options using our proposed method) as negative or positive. Here, ‘positive’ indicates a correct action-reason statement for the corresponding image. Our observations revealed that 99.28% were marked as true negatives by the annotators. Specifically, out of a total of 1669 hard negative action-reason statements generated by the LLM, only 12 statements were identified as correct (i.e., positive), while 1657 were marked as incorrect (i.e., negative) action-reason statements for the images. This underscores the effectiveness of our method in generating valid, high-quality, semantically hard negatives for the action-reason retrieval task.

Fig. 1 shows different types of hard negatives generated by GPT-4 for three images.

**BLIP-2 Failure.** While we reported the performance of BLIP-2 [9] for the AOR task, it was not effective for other tasks. For instance, BLIP-2 failed to follow instructions and produce reasonable output for multi-option/multi-label tasks like multi-ARR and MAC. For example, in the multi-ARR task, BLIP-2 erroneously identified all provided options as correct action-reason statements when only three correct statements were required. This limitation could be

Classifier	Verb.	Precision@k		
		k=1	k=2	k=3
LLaVA [10]	$I$	66.4	42.2	28.3
Vicuna [5]	$\mathcal{T}_V$ (Ours)	71.2	48.6	33.2

Table 7. ARR on Typical images

due to the lack of instruction tuning in the pre-training phase of the BLIP-2 model compared to more recent models such as LLaVA [10]. Consequently, we explored InstructBLIP, an instruction-tuned version of the BLIP-2 model.

**Effectiveness of each component in atypicality-aware verbalization.** To further evaluate the effectiveness of different steps in atypicality-aware verbalization on the performance of different LLMs on ARR tasks, we repeated the experiments on the small set. We used Vicuna, GPT-3.5, and GPT-4 as the LLMs. As observed in Table 6  $\mathcal{T}_V + \hat{S}_{IN}$  verbalization performs better, with all the LLMs.  $V+T+U+H$  includes the atypicality; however, LLaVA generated descriptions might be noisy. Combining them and denoising the combination by an LLM improve the performance. Inspired by ASR task, we detect the atypicality statement for the image using  $IN$  description. The results in Table 6 shows directly adding the detected atypicality statement to the verbalization, rather than keeping it implicit, further improves performance.

**Generalization to typical images** PittAd dataset [7] includes both typical and atypical ad images. The focus of the experiments in the main paper is on the atypical images in the dataset. However, to evaluate the generalization of the proposed atypicality-aware verbalization method to images without atypicality, we used the typical images in the dataset. Results in Table 7 show that even in images without atypicality, our atypicality-aware verbalization outperforms LLaVA, demonstrating its generalizability.

### 2.3. Generalization beyond Ads (WHOOOPS!)

WHOOOPS! [2] generates common sense-defying images by placing normal objects in an unusual context. Unlike persuasive ads, WHOOPS! doesn’t include atypical objects and its unusualness isn’t designed to convey specific messages. Hence it does not need the further reasoning ability required in ads to connect the unusualness to the final message of the image. Despite these differences, we use WHOOPS! as the closest existing benchmark to test our atypicality-aware verbalization method beyond ads. Specifically, we focus on Explanation task which involves identifying explanation for why an image is unusual.

Initially, we used 15 random explanations as negative options, but this is inadequate to effectively evaluate reasoning ability of the models. These negatives may be unrelated to the image’s scene/content, contain objects absent in the


Correct Option: Elon Musk is known as the CEO of Twitter, so he would not wear a shirt with the logo of Meta, which holds a competitive social media company named Facebook.		
	Easy Negatives	Hard Negatives
	<ol style="list-style-type: none"> <li>1. Chess is a game designed for two players, it is impossible for all the pieces on the board to be of uniform color, because then the opposing players would not know where their pieces are.</li> <li>2. Corrective eyeglasses are not prescribed to babies to aid in reading as they lack the cognitive development necessary to become literate and enjoy books at this age.</li> <li>3. A knife is used to cut pieces of food into more manageable sizes for chewing and swallowing, and another utensil such as a fork is required for bringing food from the plate to the mouth.</li> <li>4. Swimming pools are supposed to be full of water, but jumping into an empty pool without the water to break one's fall leads to serious bodily injury.</li> </ol>	<ol style="list-style-type: none"> <li>1. Elon Musk is recognized as a passionate fan of Meta, so he would not wear a shirt with the logo of Twitter, which holds a competitive social media company named Facebook.</li> <li>2. Elon Musk is known as the CEO of Twitter, so he would likely wear a shirt with the logo of Meta, which holds a competitive social media company named Facebook.</li> <li>3. Elon Musk is known as the CEO of Microsoft, so he would likely wear a shirt with the logo of Meta, which holds a competitive social media company named Facebook.</li> <li>4. Elon Musk is famous for being the CEO of SpaceX, so he would likely sport a shirt with the logo of NASA, a collaborative space exploration organization.</li> </ol>

Figure 2. An example of Explanation task in WHOOPS! dataset [2] with easy and hard negative options. Green shows correct option and Red shows incorrect options.

image, or described irrelevant actions. As a result, models could easily eliminate these options using basic image understanding, such as object recognition. For example, in Fig. 2 (left) a model could simply rule out all options due to mentioning ‘chess’, ‘babies’, ‘knife’, or ‘swimming pool’ - objects clearly not in the image. Such easy negatives fail to effectively evaluate models’ reasoning and deeper image understanding capabilities.

To address this limitation, we employed GPT-4 to generate more challenging negative options by (1) *Random Options*: randomly chosen from the explanation of other images; (2) *Alter Verb*: replacing a verb in the correct explanation with another verb and changes the meaning of the sentence; (3) *Alter Object*: replacing an effective object in the correct explanation with an object visually similar to the original object; (4) *Alter Adjective*: replacing an adjective in the correct explanation or add an adjective that changes the sentence semantically; and (5) *Alter Causal*: changing the second half of the correct explanation while keeping the first half unchanged. Unlike easy negatives, these options (right column in Fig. 2) are closely related to the image content, making simple object recognition insufficient. Instead, these hard negatives demand deeper reasoning and more nuanced analysis.

Table 8 shows that LLaVA (i.e. LLaVA outperforms Vicuna with  $\mathcal{T}_V$  verbalization with easy negatives, while Vicuna( $\mathcal{T}_V$ ) have better performance on the Explanation task with hard negative options. This demonstrates that our proposed atypicality-aware verbalization method generalizes on **unusual images beyond ads**, especially when metaphorical reasoning is required to fully interpret the image.

Classifier	Verb.	Explanation Hard	Explanation Easy
LLaVA	-	18.8	<b>88.0</b>
Vicuna	$\mathcal{T}_V$	<b>20.4</b>	65.4

Table 8. **Explanation results on Whoops dataset.** Evaluation metric for Explanation is accuracy. Explanation Easy indicates Explanation task with hard negative options generated by GPT-4 and Explanation Easy indicates Explanation task with negative options randomly chosen from the explanation of other images.

### 3. Prompts

Throughout our experimentation, we explored various prompt strategies for each LLM (i.e., Vicuna and GPT models). We utilized a fixed prompt for each task that achieved the best performance for the respective LLM, ensuring adherence to the instructions and output format. It’s important to note that all methods were implemented using the same prompt for a given LLM to ensure correctness and fair evaluation.

**Verbalization prompts.** Prompts utilized to verbalize the image and obtain ‘list of top-5 objects’ ( $V$ ), ‘text-scene’ ( $T$ ), ‘image description’ ( $IN$ ), and ‘unusualness’ ( $UH$ ) are depicted in Listing 1, Listing 2, Listing 3, Listing 4, respectively. GPT4-V prompts use the same question without LLaVA’s specific prompt format (i.e., ‘USER:<image>’ and ‘ASSISTANT:’). Finally, Listing 5 illustrates the prompt for combining the LLaVA/GPT-4V verbalization to obtain  $\mathcal{T}_V$  for both GPT-4 and Vicuna.

**Atypicality Understanding prompts.** Listing 6 and Listing 7 showcase the Multi-label classification (MAC) prompt templates for GPT and Vicuna models, respectively. Listing 8 and Listing 9 are Atypical Statement Retrieval (ASR) prompt templates for GPT and Vicuna, re-

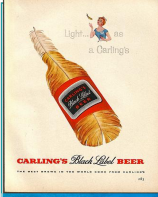
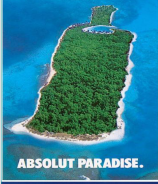


	Model: GPT4 Verb.: $\mathcal{T}_V + \hat{s}_{IN}$	LLaVA
	<ol style="list-style-type: none"> <li>1. I should drink carling's black label beer because it is as light as a Carling</li> <li>2. I should drink Carlings Because it's light</li> <li>3. I should drink beer more often Because it would make me feel good</li> </ol>	<ol style="list-style-type: none"> <li>1. I should drink Carling's black-label beer because it is as heavy as a Carling</li> <li>2. I should avoid beer more often because it would make me feel terrible</li> <li>3. I should drink Carling's black-label beer because it is as light as a Carling</li> </ol>
	<ol style="list-style-type: none"> <li>1. I should drink absolut vodka Because this vodka is like an island paradise</li> <li>2. I should buy Absolut Because it's relaxing</li> <li>3. I should buy Absolut Because it makes me more adventurous</li> </ol>	<ol style="list-style-type: none"> <li>1. I should drink absolut vodka Because this vodka is like an island paradise</li> <li>2. I should buy Absolut because it makes me more cautious</li> <li>3. I should drink absolut vodka because this vodka is nothing like an island paradise.</li> </ol>
	<ol style="list-style-type: none"> <li>1. I should eat natural salsa Because it's good for me</li> <li>2. I should buy Tostitos Because they're natural</li> <li>3. I should eat Tostitos chips Because they are all-natural</li> </ol>	<ol style="list-style-type: none"> <li>1. I should buy tostitos Because theyre natural</li> <li>2. I should eat Tostitos chips because they are all synthetic.</li> </ol>
	<ol style="list-style-type: none"> <li>1. I should buy skinny cow Because it's sweet</li> <li>2. I should avoid skinny cow because it's sweet.</li> <li>3. I should avoid skinny cow because it's unsweetened.</li> </ol>	<ol style="list-style-type: none"> <li>1. I should try some chocolates Because chocolate melts around the marshmallow and it looks good</li> <li>2. I should avoid all sweets Because chocolate melts around the marshmallow and it looks good</li> <li>3. I should try some chocolates because chocolate doesn't melt around the marshmallow and it looks unappealing.</li> </ol>

Figure 3. Examples of output from Ours (i.e., GPT-4 ( $\mathcal{T}_V + \hat{s}_{IN}$ )) and LLaVA in the multi-ARR task. Green/Red denote correct/in-correct predictions, respectively.

spectively. See Listing 10 for GPT and LLaVA, Listing 11 for MiniGPT4, and Listing 12 for BLIP2 and InstructBLIP, for examples of the prompts used in the Atypicality Object Recognition (AOR) task.

**Action-Reason Retrieval prompts.** Listing 13 and Listing 14 exhibits prompt templates for GPT-based language models for single-ARR and multi-ARR tasks. The corresponding prompts for the Vicuna language model can be found in Listing 15 and Listing 16 for the single and multi-tasks, respectively.

Listing 1. LLaVA’s prompt for list of top-5 objects  $V$

```

1 USER:
2 <image>
3 What are the non-textual objects visible
4 in this image? Carefully output AT MOST top
5 5 objects. If there are more than 5
6 objects, output major/important objects
7 according to the image. Words/Texts are
8 not considered as objects. Separate
9 with a comma.
10 ASSISTANT:

```

Listing 2. LLaVA’s prompt for text-scene  $T$

```

1 USER:
2 <image>
3 You are an OCR expert. What are the text
4 messages in the image? If there are no text
5 messages on the image, return only 'NO TXT'
6 ASSISTANT:

```

Listing 3. LLaVA’s Prompt for image description  $IN$

```

1 USER:
2 <image>
3 Describe the image in detail.
4 ASSISTANT:

```

Listing 4. LLaVA’s prompt for unusualness  $UH$

```

1 USER:
2 <image>
3 What is unusual about this image?
4 ASSISTANT:

```

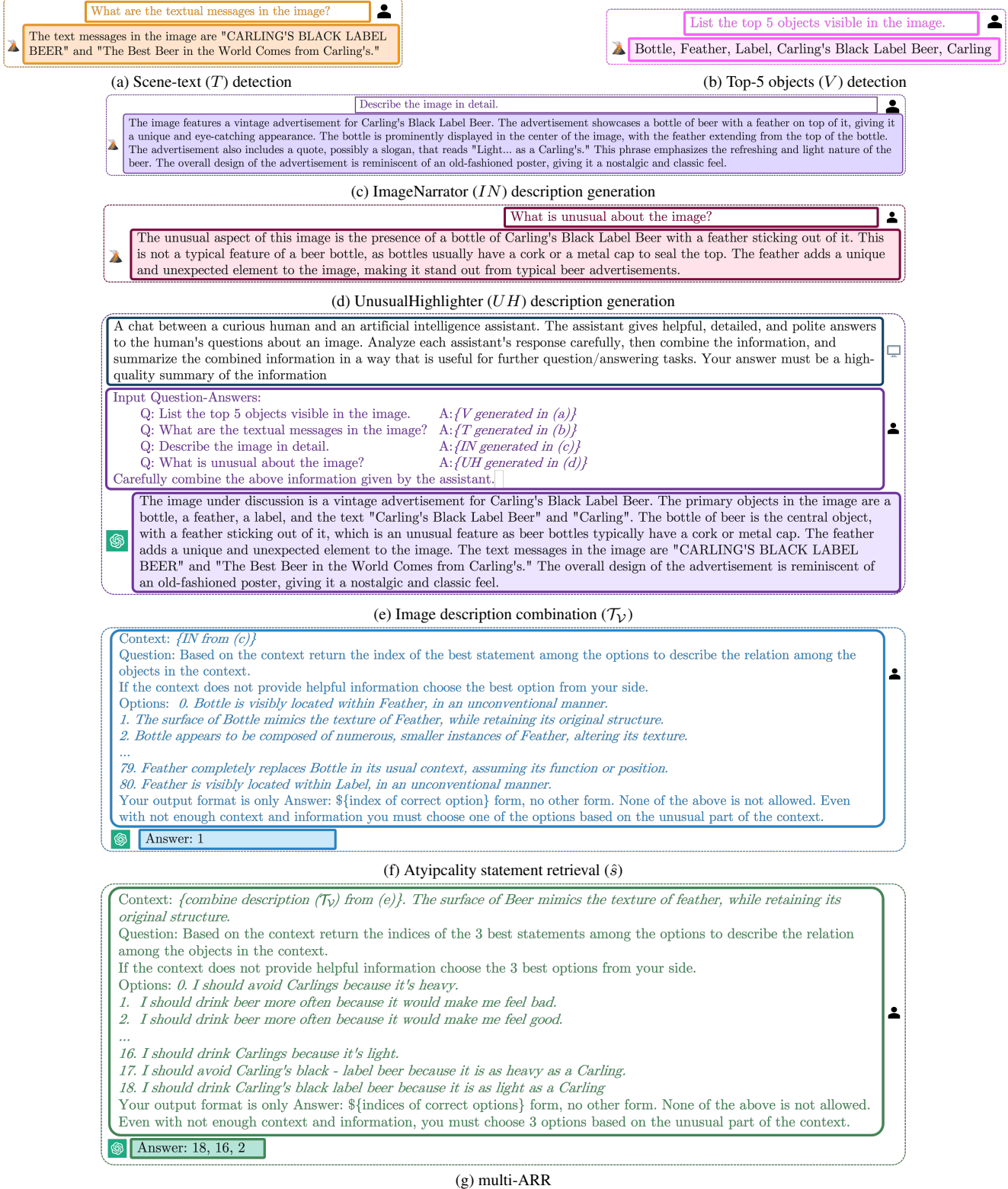


Figure 4. **Full pipeline for the multi-ARR task.** (a-d) Image verbalization with LLaVA, (e) Outputs of (a-d) are input into GPT-4 to generate the combined description  $\mathcal{T}_V$ , (f)  $V$  and atypicality statement templates  $\mathcal{S}_A$  generate atypicality statement options. Next, we use  $IN$  to retrieve the atypicality statement  $\hat{s}$ . (g) Finally, we concatenate  $\hat{s}$  with  $\mathcal{T}_V$  for multit-ARR.  $\{\}/i$  denote variable/dynamic information.

Listing 5. **GPT-4 and Vicuna Prompt template for combining LLaVA's/GPT-4V verbalizations to generate  $\mathcal{T}_V$ .** {Blue} denotes elements added dynamically.

```

1 A chat between a curious human and an
  artificial intelligence assistant. The
  assistant gives helpful, detailed, and
  polite answers to the human's questions
  about an image. Analyze each assistant's
  response carefully, then combine the
  information, and summarize the combined
  information in a way that is useful for
  further question/answering tasks. Your
  answer must be a high-quality summary of
  the information.
2
3 Input Question-Answers:
4
5 Question:
6     What are the non-textual objects
      visible in this image? Carefully
      output AT MOST top 5 objects. If there
      are more than 5 objects, output
      major/important objects according to
      the image. Words/Texts are not
      considered as objects. Separate with
      comma.
7 Answer:
8     {V (List of top-5 objects)}
9 Question:
10    You are an OCR expert. What are the
      text messages in the image?
11 Answer:
12    {T (List of scene-tests)}
13 Question:
14    Describe the image in detail.
15 Answer:
16    {IN (ImageNarrator)}
17 Question:
18    What is unusual about this image?
19 Answer:
20    {UH (UnusualHighlighter)}
21 Carefully combine the above information
    given by the assistant.

```

Listing 6. **GPT prompt template for Mac.** {Blue} denotes elements added dynamically.

```

1 Consider the following atypicality
  definition:
2     {DA atypicality definition}
3 Use the above definitions to help the user
  in classifying atypicalities in the images.
4 Question:
5 You are a highly intelligent and accurate
  image atypicality multi-label
  classification system. You take an Image
  Description as input and classify that
  into at most 4 appropriate atypicality

```

```

Categories from the given category list:
6     (1) TR1
7     (2) TR2
8     (3) OIO
9     (4) OR
10 You should select multiple atypicality
    categories ONLY if multiple atypicalities
    are present in the image.
11 If none of the atypicality categories
    exist, one of the predicted labels has to
    be "NA."
12 Your output format is only {{
    output_format|default("[{'1': 1st level
    Atypicality Category, '2': 2nd level
    Atypicality Category,...}]" )}} form, no
    other form.
13 Image Description:
14     {image-description (e.g., UH)}

```

Listing 7. **Vicuna prompt template for MAC.** {Blue} denotes elements added dynamically.

```

1 USER: You are a highly intelligent
  multi-label classification system. You
  will be given an Image Description and a
  Question. Answer the question based on the
  Image Description:
2 Image Description:
3     {description (e.g., UH)}
4 Question:
5 According to the Image Description and the
  atypicality definitions below, detect the
  atypicality categories:
6     {DA atypicality definition}
7 You should select multiple atypicality
  categories ONLY if multiple atypicalities
  are present in the image.
8 If none of the atypicality categories
  exist, one of the predicted labels has to
  be "NA".
9 You must choose the detected atypicalities
  from (OIO, TR1, TR2, and OR) and the
  acceptable output format is only {{
  output_format|default("[{'1': 1st level
  Atypicality Category, '2': 2nd level
  Atypicality Category,...}]" )}} form, no
  other form. Do NOT output any extra
  information or explanation.
10 ASSISTANT:

```

Listing 8. **GPT prompt template for Atypical Statement Retrieval (ASR).** {Blue} denotes elements added dynamically.

```

1 Context: {description (e.g., IN)}
2 Question: Based on the context return the
  index of best statement among the options
  to describe the relation among the objects
  in the context.

```

```

3 If the context does not provide helpful
  information, choose the best option from
  your side.
4 Options: {list of generated correct and
  incorrect atypicality statements}
5 Your output format is only Answer: ${index
  of correct statement} form, no other form.
  None of the above is not allowed. Even
  with not enough context and information,
  you must choose one of the options based
  on an unusual part of the context.

```

Listing 9. Vicuna prompt template for Atypical Statement Retrieval (ASR). {Blue} denotes elements added dynamically.

```

1 USER:
2 Context: {IN description}
3 Question: Based on the context return the
  index of best statement among the options
  to describe the relation among the objects
  in the context.
4 If the context does not provide helpful
  information, choose the best option.
5 Options: {list of generated correct and
  incorrect atypicality statements}
6 Your output format is only Answer: ${index
  of correct statement} form, no other form.
  None of the above is not allowed. Even
  with not enough context and information,
  you must choose one of the options based
  on an unusual part of the context.
7 ASSISTANT:

```

Listing 10. GPT and LLaVA prompt template for Atypical Object Recognition (AOR). {Blue} denotes elements added dynamically based on the atypicality relation. Here, we show the TR1 atypicality relation as an example.

```

1 USER:
2 <image>
3 A human has described this image as
  atypical. They have found it atypical
  because of: Texture Replacement 1, with
  objects' texture borrowed from another
  object.
4 More specifically, The surface of <object1>
  mimics the texture of <object2>, while
  retaining its original structure.
5 Fill in your answers for <object1> and
  <object2>. Make sure to include the
  angular brackets < and >.
6 An example output: The surface of <eleven>
  mimics the texture of <meat>, while
  retaining its original structure.
7 ASSISTANT:

```

Listing 11. MiniGPT4 prompt template for Atypical Object Recognition (AOR). {Blue} denotes elements added dynamically based on the atypicality relation. Here, we show the TR1 atypicality relation.

```

1 USER:
2 <image>
3 A human has described this image as
  atypical. They have found it atypical
  because of: Texture Replacement 1, with
  objects' texture borrowed from another
  object.
4 More specifically, The surface of <object1>
  mimics the texture of <object2>, while
  retaining its original structure.
5 Give short answers for what <object1> and
  <object2> are, in the format:
6 <object1>: <answer1>
7 <object2>: <answer2>
8 ASSISTANT:

```

Listing 12. BLIP2 and InstructBLIP prompt template for Atypical Object Recognition (AOR). We use a multi-step prompt to generate the primary and secondary objects separately. {Blue} denotes elements added dynamically based on the atypicality relation. Here, we show the TR1 atypicality relation.

```

1 <image>
2 A human has described this image as
  atypical. They have found it atypical
  because of: Texture Replacement 1, with
  objects' texture borrowed from another
  object.
3 More specifically, The surface of <object1>
  mimics the texture of <object2>, while
  retaining its original structure.
4 Give short answers for what <object1> and
  <object2> are.
5 <object1>: VLM prompted here
6 <object2>: VLM prompted here

```

Listing 13. GPT prompt template for Action-Reason Retrieval (ARR) choosing single correct option. {Blue} denotes elements added dynamically.

```

1 Context: {Tv description} {Atypicality
  statement}
2 Question: Based on the context return the
  index of the best statement among the
  options to interpret the described image.
3 Even without enough information return the
  index of the best option among the
  options.
4 Options: {list of correct and incorrect
  action-reason statements}
5 Your output format is only Answer: ${index
  of correct statement} form, no other form.
6 None of the above is not allowed. Even
  without enough information choose the best
  interpretation.

```

Listing 14. GPT prompt template for Action-Reason Retrieval (ARR) choosing all correct options. {Blue} denotes elements added dynamically.



```

1 | Context: {Tv description} {Atypicality
  | statement}
2 | Question: Based on the context return the
  | indices of the 3 best statements among the
  | options to interpret the described image.
3 | Separate the answers by comma and even
  | without enough information return the
  | indices of the 3 best options among the
  | options.
4 | Question: {list of correct and incorrect
  | action-reason statements}
5 | Your output format is only Answer:
  | ${indices of the 3 best
  | statements} form, no other form.
6 | None of the above is not allowed. Even
  | without enough information choose the 3
  | best interpretations.

```

Listing 15. Vicuna prompt template for Action-Reason Retrieval (ARR) choosing single correct option. {Blue} denotes elements added dynamically.

```

1 | USER:
2 | Context: {Tv description} {Atypicality
  | statement}
3 | Question: Based on the context return the
  | index of the best statement among the
  | options to interpret the described image.
4 | Options: {list of correct and incorrect
  | action-reason statements}
5 | None of the above is not allowed. Even
  | without enough information, choose the
  | best interpretations.
6 | Your output format is only Answer: ${index
  | of correct statement} form, no other form.
7 | ASSISTANT:

```

Listing 16. Vicuna prompt template for Action-Reason Retrieval (ARR) choosing all correct options. {Blue} denotes elements added dynamically.

```

1 | USER:
2 | Context: {Tv description} {Atypicality
  | statement}
3 | Question: Based on the context, return the
  | indices of the 3 best statements among the
  | options to interpret the described image.
4 | Separate the answers by comma, and even
  | without enough information, return the
  | indices of the 3 best options.
5 | Options: {list of correct and incorrect
  | action-reason statements}
6 | None of the above is not allowed. Even
  | without enough information, choose the 3
  | best interpretations.
7 | Your output format is only Answer:
  | ${indices of the 3 best
  | statements} form, no other form.
8 | ASSISTANT:

```

Listing 17. Prompt for generating Action Alter hard negatives. {Blue} denotes elements added dynamically, based on the correct option.

```

1 | Generate one hard negative statement that
  | semantically contradicts the action in the
  | following correct statement.
2 | The hard negative should be plausible but
  | must convey an opposite or entirely
  | different
3 | action, while the underlying reason
  | remains unchanged. This requires reversing
  | the
4 | action's intent or suggesting a completely
  | different concept that contrasts with
  | the original message, yet sounds coherent
  | when paired with the same rationale.
5 |
6 |
7 |
8 | Example:
9 |   - Correct Statement: "I should get
  |   involved with artistic expression
  |   because dressing in style is a type of
  |   art."
10 |   - Generated Hard Negative: "I should
  |   avoid artistic expression because
  |   dressing in style is a type of art."
11 | In this example, "I should get involved
  | with artistic expression" is the action,
  | which is inverted to "I should avoid
  | artistic expression" in the hard negative.
12 | The reason, "because dressing in style is
  | a type of art," remains constant.
13 |
14 | Correct Interpretation: {correct option}
15 |
16 | The hard negatives should closely mirror
  | the vocabulary of the correct
  | interpretation but must imply an opposite
  | or distinctly different meaning. Only the
  | hard negative statement is needed, without
  | additional explanations.

```

Listing 18. Prompt for generating Reason Alter hard negatives. {Blue} denotes elements added dynamically, based on the correct option.

```

1 | Create a hard negative statement that
  | presents semantically incorrect or
  | opposite reasons compared to the provided
  | correct statement while keeping the main
  | action unchanged. These hard negatives
  | should seem plausible at a glance but must
  | convey a reason that contradicts the
  | correct one. The intention is to maintain
  | a surface-level similarity in wording with
  | the original statement but to invert the
  | underlying rationale.
2 |
3 | Example:

```

4 - Correct Statement: ``I should get  
involved with artistic expression  
because dressing in style is a type of  
art.``  
5 - Generated Hard Negative: ``I should  
get involved with artistic expression  
because dressing in style lacks  
artistic value.``

6  
7 In this example, the action phrase ``I  
should get involved with artistic  
expression`` remains the same across both  
statements. The original reason, ``because  
dressing in style is a type of art`` is  
transformed to imply the opposite meaning,  
``because dressing in style lacks artistic  
value,`` for the hard negative.

8 **Guidelines:**

- 9  
10 1. Retain the action statement unchanged.  
11 2. Invert the logic or reasoning of the  
correct statement to formulate the hard  
negative.  
12 3. Ensure the hard negative retains  
similar wording to the original, but  
clearly communicates a contradictory  
reason.

13  
14 **Correct Interpretation:** {correct option}

15  
16 Provide only the hard negative statement,  
ensuring it closely mirrors the correct  
interpretation in structure and vocabulary  
but distinctly opposes it in meaning.

Listing 19. Prompt for generating Statement Alter hard negatives. {Blue} denotes elements added dynamically, based on the correct option.

1 Generate a hard negative statement that is  
semantically unrelated and incorrect  
compared to a given correct statement.  
These hard negatives should be coherent  
statements on their own but must diverge  
completely in meaning from the original  
statement. The challenge is to craft a  
statement that, while maintaining  
superficial word similarity to the correct  
statement, introduces a concept or  
reasoning that is entirely irrelevant and  
incorrect.

2  
3 **Example:**

4 - Correct Statement: ``I should use  
5-hour energy because it will keep me  
focused.``  
5 - Generated Hard Negative: ``I should  
use 5-hour stress drink because it

promotes relaxation.``

6  
7 **Guidelines:**

- 8 1. Keep a superficial structural  
similarity to the correct statement in  
terms of wording.  
9 2. Change the concept or reasoning to  
something totally irrelevant or even  
diametrically opposed to the original  
statement.  
10 3. The hard negative should be plausible  
as a standalone statement but should not  
accurately reflect the logic or purpose of  
the correct interpretation.

11  
12 **Correct Interpretation:** {correct option}

13  
14 Provide only the hard negative statement.  
It should closely mimic the correct  
statement in form but must diverge  
significantly in semantic content or  
meaning, introducing a totally different  
concept.

Listing 20. Prompt for generating Object Swap hard negatives. {Blue} denotes elements added dynamically, based on the correct option.

1 Please generate a hard negative statement  
that has semantically incorrect (e.g.,  
opposite) meaning to the one in the  
following correct statement by changing at  
least one object in the statement. Each  
hard negative should be a plausible option  
but must convey the incorrect meaning as  
the correct one.

2  
3 **Example:**

4 - Correct statement: I should get  
involved with artistic expression  
Because dressing in style is a type of  
art  
5 - Generated Incorrect statement: I  
should get involved with sports  
Because professional soccer is a type  
of sport

6  
7  
8 **Correct Interpretation:** {correct option}

9  
10 Ensure that the hard negatives maintain a  
degree of similarity to the correct  
interpretation in terms of words but imply  
incorrect meaning and include incorrect  
objects.  
11 Only return the hard negative.

Listing 21. **Prompt for generating Adjective Alter hard negatives.** {Blue} denotes elements added dynamically, based on the correct option.

```
1 Given a correct statement, your task is to
2 generate a hard negative statement. A hard
3 negative statement should closely resemble
4 the original statement in structure but
5 convey a totally different meaning. This
6 can be achieved by either changing an
7 adjective to its antonym or by adding a
8 qualifying adjective that totally changes
9 the statement's sentiment. The goal is to
10 create a plausible, yet semantically
11 different version of the original
12 statement.
13 The resulting hard negative should:
14 - Only change or add an adjective
15 - Keep the core structure of the
16 original statement intact.
17 - Alter the meaning to be totally
18 different or even opposite by focusing
19 on the modification of adjectives.
20 - Ensure that the new statement is
21 plausible and grammatically correct,
22 but clearly wrong when compared to the
23 original correct interpretation.
24 Example:
25 - Correct Statement: ``I should use
26 5-hour energy because it will keep me
27 focused.''
28 - Hard Negative: ``I should use 5-hour
29 energy because it will keep me
30 sleepy.''
31 Correct Interpretation: {correct option}
32 Please generate a hard negative based on
33 the provided correct interpretation,
34 focusing on the inversion of adjectives to
35 create a totally different meaning.
```

## References

- [1] Gpt-4v(ision) system card. 2023. 1
- [2] Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2616–2627, 2023. 1, 3, 4
- [3] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024. 2
- [4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198, June 2024. 2
- [5] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality, March 2023. 2, 3
- [6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1
- [7] Zaem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1705–1715, 2017. 3
- [8] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022. 2
- [9] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 1, 3
- [10] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 2, 3
- [11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1
- [12] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023. 2
- [13] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 1