HumanBench: Two Heads, No Legs, But Mostly Human, the State of Generative Capabilities in T2I Models - Supplementary Material

Anubhooti Jain, Mayank Vatsa, Richa Singh
Indian Institute of Technology Jodhpur, India
{jain.44,mvatsa,richa}@iitj.ac.in

1. Experimental Setup and Details

Models used for evaluation. As discussed in the main paper, different models have been used for generating captions (LLaVA-1.5 and BLIP-2, detecting humans and objects in the images (OWL-ViT), and for evaluating similarity (CLIP). The pretrained models are used with their configurations reported in Tab. 1.

Model	#Params	Backbone
OWL-ViT	153M	ViT/B-32
CLIP	151M	ViT/B-32
BLIP-2	3.94B	Flan-T5
LLaVA-1.5	7B	LLaMA

Table 1. Number of Parameters and Backbone details for models used in the experiments.

Counterfactuals. We create a set of counterfactuals to semantically change the prompt. The logic for creating them is reported in Tab. 2

Template	Replacements
<spatial>	(above, below), (on, in), (on the right of, on the left of), (in front of, behind), beside, among → above, between → in, around → below
<action>	walking → cycling, standing → eating, sitting → jumping, lying down → walking, cycling → walking, eating → writing, laughing → crying, crying → laughing, playing → lying down, jumping → eating, writing → standing
<texture>	rubber → wooden, plastic → metallic, metallic → glass, wooden → fluffy, fluffy → rubber, glass → leather, leather → plastic
<number>	(two, seven) (three, six) (four, five)

Table 2. Logic used for creating counterfactuals. (a, b) means that a was replaced by b and vice versa. a → b means that a was replaced by b.

These changes help create a meaningful gap between the counterfactual and the image. <spatial> was replaced for prompts following template P5 and P4, <action> for P1, P2, P7, and P8, <texture> for P3, and <number> for P6.

Missing Prompts and Seeding. For generating four variations using the Stable Diffusion and FLUX models, four different seeds were used. The seeds are: 21, 42, 84, and 168. The guidance level and number of inference steps for all the seeds are kept constant at 2.5 and 28 for Stable Diffusion, and 2.5 and 30 for FLUX, respectively. Furthermore, for image generation using Midjourney, there are twenty prompts that were marked as inappropriate by the model. The list of prompts is given in Tab. 3.

Inappropriate Prompts
A female lying down and a male eating
A female between rubber chairs
A woman below a leather table
A girl between leather bags
A girl below a leather table
Five men standing and four women lying down
Two men lying down and four women eating
Five men eating and five women lying down
Four females lying down and four males walking
Four females lying down and four males standing
Four females lying down and four males sitting
Four females lying down and four males cycling
Four females lying down and four males eating
Four females lying down and four males laughing
Two females lying down and four males crying
Two females lying down and four males playing
Two females lying down and four males jumping
Two females lying down and five males writing
Two girls lying down and five boys sitting
Two girls lying down and five boys eating

Table 3. Inappropriate prompts as marked by the Midjourney model.

Images for these prompts are not generated. However,

Spatial Attribute	Logic Condition	Equations
above	bottom of person is above the top of object	$P(y_{max}) \leq O(y_{min}) + \tau \cdot (O(y_{max}) - O(y_{min})$
below	top of person is below the bottom of object	$P(y_{min}) \geq O(y_{max}) - \tau \cdot (O(y_{max}) - O(y_{min})$
on	high IoU or very close vertically	IoU $\geq$ IoU$_T$ or $\lvert P_{y_{\min}} - O_{y_{\max}} \rvert \leq \tau \cdot (O_{y_{\max}} - O_{y_{\min}})$
in	person almost completely inside object	$\mathrm{I} = max(0, min(P(x_{max}), O(x_{max})))$ $- max(P(x_{min}), O(x_{min})))\cdot$ $max(0, min(P(y_{max}), O(y_{max}))$ $- max(P(y_{min}), O(y_{min}))),$ $\mathrm{Ar} = (P(x_{max}) - P(x_{min}))\cdot$ $(P(y_{max}) - P(y_{min})), \quad \frac{\mathrm{I}}{\mathrm{Ar}} >= \tau$
on the right of	left side of person is to the right of object	$P(x_{min}) \geq O(x_{max}) - \tau \cdot (O(x_{max}) - O(x_{min})$
on the left of	right side of person is to the left of object	$P(x_{max}) \leq O(x_{min}) + \tau \cdot (O(x_{max}) - O(x_{min})$
beside	significant horizontal overlap	$\text{Overlap} = max(0, min(P(x_{max}), O(x_{max})))$ $- max(P(x_{min}), O(x_{min})),$, $width_P = P(x_{max}) - P(x_{min}),$ $width_O = O(x_{max}) - O(x_{min}),$ $\frac{Overlap}{max(width_P, width_O)} > \tau$
in front of	person is smaller vertically (analyzing in 2D)	$(P(y_{max}) - P(y_{min})) < (O(y_{max}) - O(y_{min}))$
behind	person is larger vertically (analyzing in 2D)	$(P(y_{max}) - P(y_{min})) > (O(y_{max}) - O(y_{min}))$
between	center of person between object 1 and 2	$\text{vertical} = O^{left}_{center} \leq P_{center} \leq O^{right}_{center},$ $\text{horizontal} = O^{top}_{center} \leq P_{center} \leq O^{bottom}_{center}$
among	center of person near average center of other objects	$\lVert \mathbf{c} - \bar{\mathbf{c}} \rVert \leq \tau \cdot \left(\frac{1}{n} \sum_{i=1}^{n} \lvert \mathbf{c}_i - \bar{\mathbf{c}} \rvert \right)$
around	center of person near average center of other objects	$\lVert \mathbf{c} - \bar{\mathbf{c}} \rVert \leq \tau \cdot \left(\frac{1}{n} \sum_{i=1}^{n} \lvert \mathbf{c}_i - \bar{\mathbf{c}} \rvert \right)$

Table 4. Different Spatial Attributes with the logical conditions and equations used to analyze the attributes. $P(.)$ and $O(.)$ are the box coordinates for Person and Object in the images, τ is the threshold for different conditions, IoU is the Intersection over Union, $\mathbf{c}$ and $\bar{\mathbf{c}}$ are the centers of person and different objects in the image respectively.

as the Stable Diffusion and FLUX models were not used in any NSFW setting, the latter two generate images for these prompts.

2. Spatial Relation Analysis

As discussed in the section on Spatial Relation Analysis in the main paper, there are different spatial attributes used for generating and evaluating the synthetic images. The logic and formulations used for evaluating them are reported in Tab. 4. Most of the attributes depend on threshold values as per the logic used for evaluation. The thresholds for different attributes are reported in Tab. 5.

3. Extra Results and Discussion

Focused Texture Analysis. For a better texture analysis and to focus on the intended region, we crop the detected person (for prompt template type P3) or object (for prompt template type P5) via OWL-ViT before classification. The accuracy is reported in Tab. 6. Although 4515 images lack a detectable target and are discarded, the low-performing

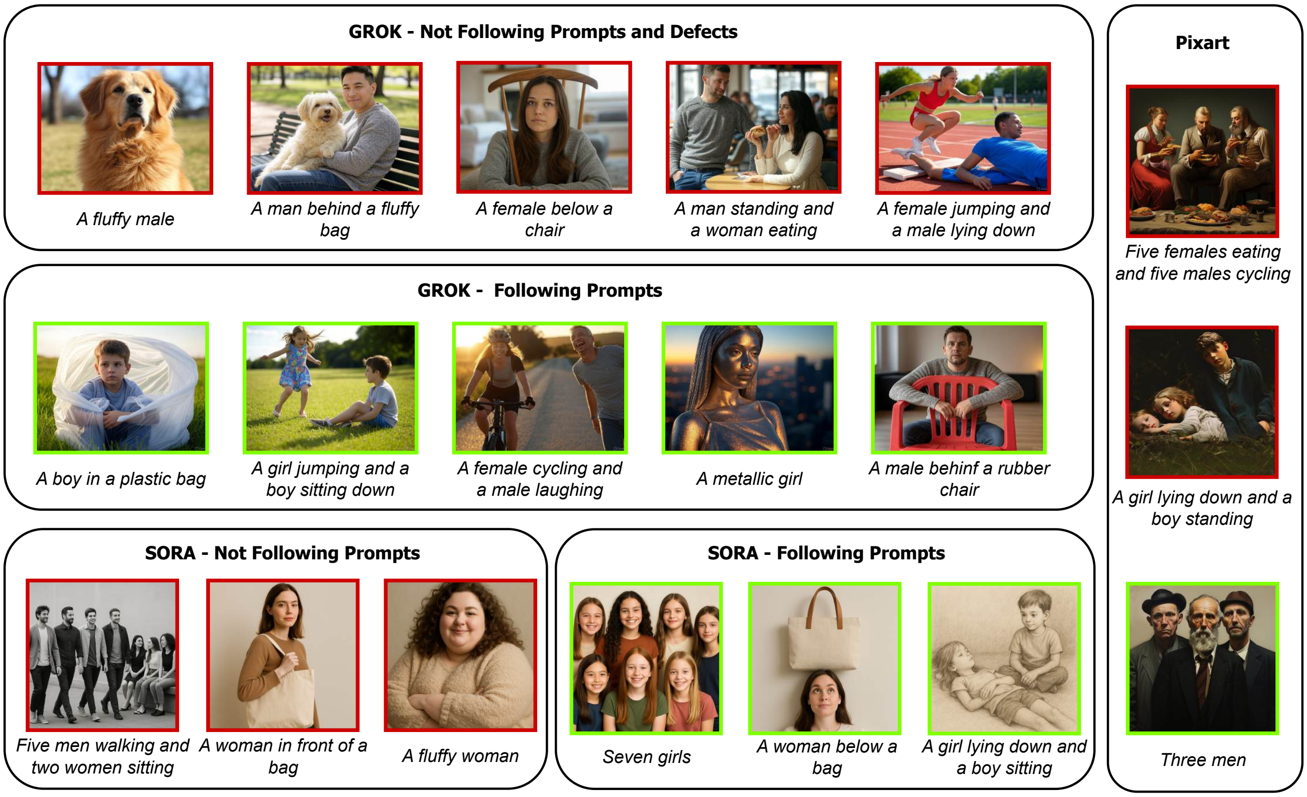
GROK - Not Following Prompts and Defects
A fluffy male
A man behind a fluffy bag
A female below a chair
A man standing and a woman eating
A female jumping and a male lying down
GROK - Following Prompts
A boy in a plastic bag
A girl jumping and a boy sitting down
A female cycling and a male laughing
A metallic girl
A male behinf a rubber chair
SORA - Not Following Prompts
Five men walking and two women sitting
A woman in front of a bag
A fluffy woman
SORA - Following Prompts
Seven girls
A woman below a bag
A girl lying down and a boy sitting
Pixart
Five females eating and five males cycling
A girl lying down and a boy standing
Three men

Figure 1. Examples (green border: positive and red border: negative) for images generated using GROK, SORA, and Pixart-α.

Spatial Relations	Threshold
above/below	0.1
on	0.3
in	0.9
on the right of/ on the left of	0.1
beside	0.2
among/ around	0.5

Table 5. Thresholds used for evaluating different spatial attributes.

textures,*rubber* and *metallic*, improved significantly.

In several cases, we observe that the texture is generated, but in different objects. Especially for *fluffy*, where the texture is often mistaken for a fluffy animal. Particularly, in SORA, we observed that 'fluffy' generated images of a female with a larger body size. In terms of accuracy, across all the tasks, we find that Midjourney performs well. However, all the models show comparable performances and struggle equally with compounded instructions.

Other Models. As already stated, we generated images using GROK, SORA, and Pixart-α. Pixart-α is an open-source model that generated several defects in the images for our prompts. It was thus not included in the main set of the benchmark. Further, both GROK and SORA are closed-source models and generate images that have fewer defects, especially SORA. However, they struggle like every other model. Some examples are shown in the Fig. 1. Just like FLUX, Pixart-α, and SORA have a tendency to generate animated images, especially when involving children.

Texture	MJ	SD	Flux	#N	Averaged
rubber	37.66	49.89	37.75	2628	41.77
plastic	37.97	34.09	29.95	2722	34.00
metallic	59.56	40.79	19.93	2769	40.09
wooden	93.27	81.53	69.00	2728	81.27
fluffy	66.67	66.35	65.63	1394	66.22
glass	63.84	59.30	57.34	2047	60.16
leather	77.06	71.37	67.19	2937	71.87
All	62.29	57.62	49.54	17225	56.48

Table 6. Number of Instances and Accuracy (Acc) (in %) for the task of Texture Recognition (on Cropped images). MJ: Midjourney, SD: Stable Diffusion.

Regression Analysis. As mentioned in the main paper, we reported that accuracy declines with increasing complexity. For the regression analysis, we used the accuracy from our binary check between the prompt and the image as the de-

Aspect	Question	Choices
Consistency	Considering the prompt, do you think the given image follows the associated prompt?	Fully, Partially, Not at all
Defect	Do all humans look physically plausible (e.g., no extra limbs)	Yes, No
Realism	Rate Image Realism (how plausible and real the image looks to you)	Numeric Rating between 1 to 5

Table 7. Questions as presented to evaluators for the human study for evaluating the concept of Consistency, Defect, and Realism.

Prompt Type	Aspect	Question	Choices
P1, P7	AR (Single)	Do you think that the action in the mentioned prompt is appropriately depicted in the given image?	P1: Yes{1}, No{0}; P7: Fully (All people follow the same action){1}, Partially (Some people follow the given action){0.5}, Not at all{0}
P2, P8	AR (Multiple)	Do you think that the actions in the mentioned prompt is appropriately depicted in the given image?	P2: Both actions are correct{1}, Only one correct action{0.5}, No correct actions{0}; P8: Fully (Each group performs the action assigned to them){1}, Correct actions but mismatch in terms of gender group{1}, Partially (Only one group performs the action assigned to them){0.5}, Not at all{0}
P3, P5	Texture	In the given image, do you see the texture mentioned in the prompt?	P3: Texture seen in the person{1}, Texture seen around/ in the background of the person{0.5}, Texture doesn't match the one mentioned in the prompt at all{0}; P5: Texture seen in the object{1}, Texture seen in the person{0.5}, Texture seen around/ in the background of the person{0.5}, Texture doesn't match the one mentioned in the prompt at all{0}
P4, P5	Spatial	Does the image follow the spatial relationship mentioned in the prompt?	Yes{1}, No{0}
P6, P7, P8	Count	In the given image, does the number of people match the given count?	P6, P7: Yes{1}, No{0}; P8: Yes{1}, The count matches but there is a mismatch in gender count{1}, Partially (one set of gender matches the given count and the other doesn't){0.5}, Not at all{0}

Table 8. Questions as presented to evaluators, specific to the prompt type, for evaluating Action Recognition (AR), Texture Recognition, Spatial Recognition, and Counting. Numbers within {} indicate the score assigned to choices presented to the evaluator.

pendent variable, and complexity as the independent variable. Controlled for model differences, the analysis is performed on the averaged template-level accuracies (Table 3 in the main paper). We obtain a significant negative slope ($\hat{\beta} = -8.15$, $p < 0.001$), with $R^2 = 0.67$. This aligns with our observations that higher compositional complexity reduces prompt adherence. With no control for the models, we separately performed a regression analysis and observed a negative slope across the board: $\hat{\beta} = -7.19$, $R^2 = 0.67$ for Midjourney, $\hat{\beta} = -8.48$, $R^2 = 0.73$ for Stable Diffusion, and $\hat{\beta} = -8.78$, $R^2 = 0.63$ for FLUX; all for $p < 0.05$.

4. Human Grounding in Evaluation Pipeline

Incorporating human judgments can be crucial, especially since the evaluation pipeline so far is mainly automated. We conducted human validation (outside the authors) of the generated images. We received 38 submissions, leading to human annotation of 300 images, assigned to the participants randomly. The evaluators are all researchers between the ages of 18 and 40. The images are selected uniformly at random from each template set to have a true representative set. Out of the 300 images, we included 50 images from each model. That is, 50 images from all six models, Midjourney, Stable Diffusion, FLUX, GROK, SORA, and Pixart-α, contributed to the evaluation set. Each evaluator was provided with a unique compressed archive containing the evaluation images. Within this archive, images were named according to the format "prompt-type_prompt," allowing the evaluators to identify the originating prompt condition. Responses were collected through a questionnaire implemented using Google Forms.

To evaluate template consistency, the evaluators are asked if a given image follows the corresponding prompt. We also ask them to mark if they see any defects in the image. For realism, we ask them to rate the image out of 5 based on how realistic the image looks. The exact ques-

Prompt Type	Consistency (↑)	Defect (↓)	Realism (↑)
P1	85.52	15.78	3.92
P2	81.42	11.11	3.52
P3	76.38	27.78	3.63
P4	65.27	**5.71**	4.05
P5	74.32	22.22	3.73
P6	72.97	16.21	3.24
P7	65.79	28.95	3.26
P8	**40.27**	47.22	**2.80**
Average	70.25	21.87	3.52

Table 9. Human Study evaluation for template consistency, defect detection, and image realism.

Spatial	AR (Single)	AR (Multiple)	Texture	Count
57.43	84.86	59.02	80.22	45.73

Table 10. Human Study evaluation for spatial relation, action recognition (AR), and texture recognition along with counting.

tions along with the choices presented to the evaluators are reported in Tab. 7. The choices are mapped to numerical scores and are reported in Tab. 9. It can be observed that as the prompt gets complex, consistency and realism go down, and the defects in the images increase.

The next questions evaluate spatial relations, action recognition, texture accuracy, and counting. For compounded prompts, we assess whether concepts (spatial relations, actions, or textures) are generated correctly for both groups, one group, or neither. For counting, we verify whether the generated images match the number stated in the prompt for each group. We also consider the partial fulfillment of the various concepts. The exact questions with the choices are given in Tab. 8. The scores assigned to the choices are not revealed to the evaluators. Results (Tab. 10) show: (a) action recognition aligns well with automated evaluations, outperforming spatial relations, indicating poor understanding of the concept, and (b) compounded prompts often introduce more errors, particularly in counting. Although limited to 300 responses here, we found consistent human assessments.

5. Complexity and Confusion

As discussed with the formulation in the main paper, we define complexity as the compositional load of a prompt. For example, for P1, there are two attributes and one attribute (action) bound to another attribute (person), so the total complexity becomes 3. P3 and P6 follow the same logic and have a complexity level of 3. Further, for P2 we have 4 attributes, with 1 binding between two attributes, so two bindings in total, along with a conjunction binding, that is $4 + 2 + 1 = 7$. For P4, we have 3 attributes, along with 1 binding through a spatial attribute and 1 relation between person and object. Collectively, it becomes 5. With P5, we have 4 attributes along with 1 relation between person and object, 1 binding between texture and object, and 1 binding through spatial relation. Thus, giving a complexity of 7. For P7, there are 3 attributes, and 2 bindings, 1 between number and person, and 1 binding between action and person, giving a complexity of 5. And finally, for P8, there are 6 attributes, 1 conjunction, and 2 bindings between each group, giving a complexity of 11.

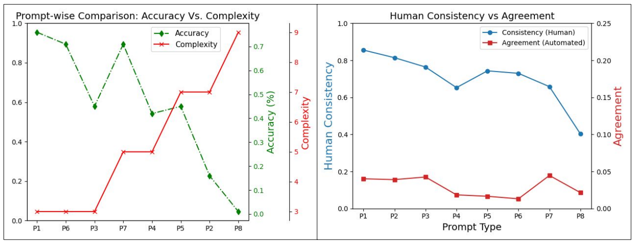

Figure 2. Left: Accuracy Vs. Complexity (on the entire image set); Right: Human-based consistency Vs. CLIPScore-based Agreement metric (on the 300 samples used for the human study).

We also report the correlation chart between accuracy (from the binary check using the LLaVA model) and complexity level as seen in the left plot of Figure 2. We can see that as complexity increases, the accuracy tends to decrease in most cases.

As for confusion, we define the term as the model struggling with generating the right part. For example, when the <action> part is given as walking, but instead the model generated sitting, we say the model was confused while generating the right action. We do not claim that prompt complexity is the sole cause of generation or fidelity errors, but it is a major factor based on our experiments and findings. There is a constant degradation in the quality and anatomically accurate human generation in the generated images as the complexity load increases in the corresponding prompt.

6. CLIPScore and Other Metrics

CLIPScore has been known to, sometimes, miss finer-grained details. As seen in the main paper, we found that even then, we are able to correlate with human evaluation and truly evaluate the subtleties in counterfactuals and negations. We thus claim that it works well in our evaluation because: (1) CLIPScore provides a strong, scalable baseline for text–image alignment. Here, its primary role is to capture coarse semantic fidelity with counterfactuals and negation analysis. (2) We have complemented the metric with a VQAScore-based evaluation check in Section 3.2,1(c) of the main paper using two models, LLaVA and BLIP-2. (3)

We also plotted a correlation graph between the Agreement metric and the consistency evaluated from the human study. It can be seen as the right plot in Figure 2.

7. Limitations and Future Work

The benchmark evaluation is mostly automated and lacks a human evaluation on the entire set. Getting human annotations is a possible research direction for this work. Also, there is a tradeoff between prioritizing maintaining human anatomy and following the prompt completely in all the tested models. Finding the balance to accurately gauge and maintain human formation is crucial and the natural next step to this work.