Be More Specific: Evaluating Object-centric Realism in Synthetic Images

Supplementary Material

A. Data

A.1. Data Collection Details

Annotation interface and guidelines We developed detailed annotation guidelines to guide annotators on realism definitions, annotation procedures, and interactions with the annotation user interface (Fig. 11 (a)). We provided visual examples (Fig. 11 (b)) illustrating different levels of realism. The detailed definitions of shape and texture realism are outlined as follows:

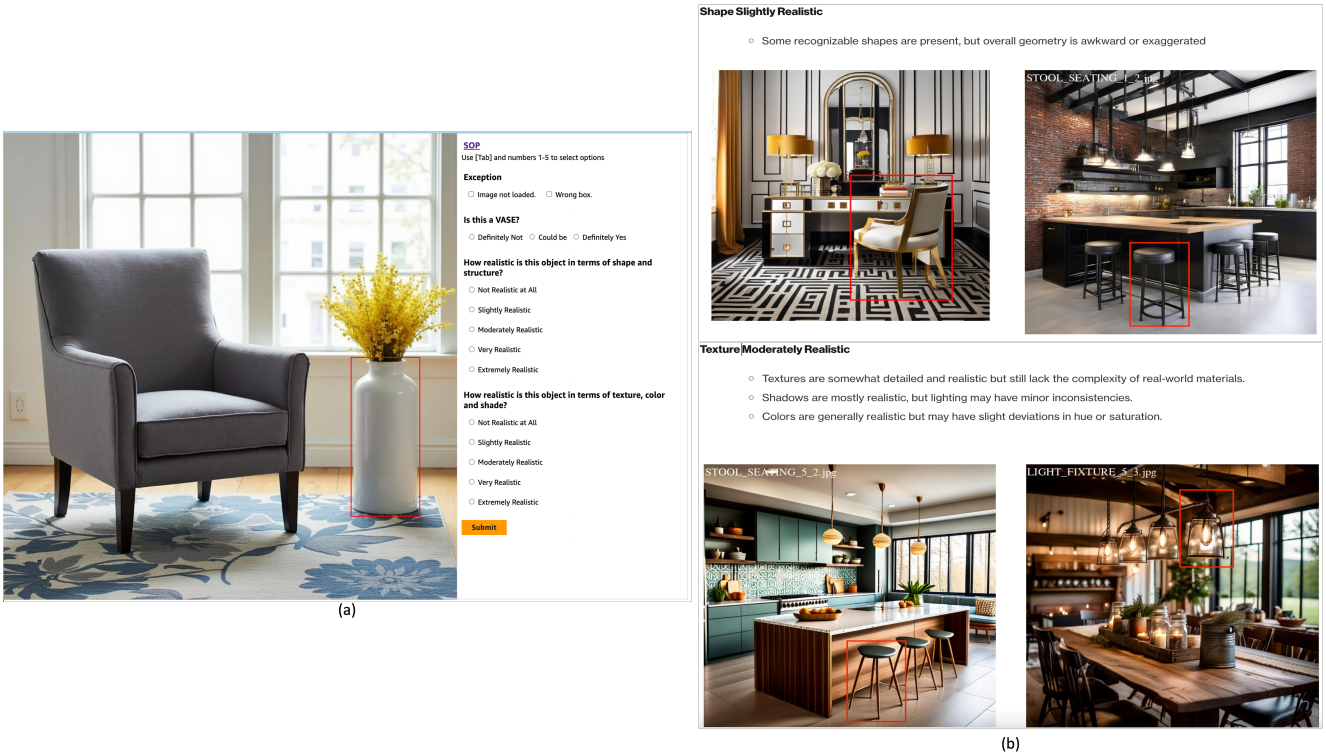


Figure 11. Annotation interface (a) and annotation guideline examples (b).

- **Shape Realism**
  - Not Realistic at All: Object shapes are distorted or nonsensical, with no regard for geometric accuracy.
  - Slightly Realistic: Some recognizable shapes are present, but overall geometry is awkward or exaggerated.
  - Moderately Realistic: Object shapes are mostly accurate with minor to moderate imperfections.
  - Very Realistic: Object shapes closely resemble real-world counterparts with minor imperfections.
  - Extremely Realistic: Object shapes are perfectly defined and closely match real-world counterparts.
- **Texture Realism**
  - Not Realistic at All:
    - Textures are completely artificial, lacking any detail or variation. Material properties are not represented accurately.
    - Harsh, unrealistic shadows and inconsistent lighting that does not mimic real-world conditions.
    - Colors are unnatural, overly saturated, and do not match real-world counterparts.
  - Slightly Realistic:
    - Textures exhibit basic details but lack realism and appropriate variations.
    - Shadows are soft but not entirely realistic, and lighting has some inconsistencies.
    - Colors have a slight resemblance to real-world objects but are still noticeably off in tone or brightness.
  - Moderately Realistic:
    - Textures are somewhat detailed and realistic but still lack the complexity of real-world materials.
    - Shadows are mostly realistic, but lighting may have minor inconsistencies.
    - Colors are generally realistic but may have slight deviations in hue or saturation.

Annotator Type	Alignment		Re-Eval Rate
	Image	Object	
Low expertise	0.36	0.30	27%
High expertise	0.59	0.44	14%

Table 4. Annotation quality among two cohorts of annotators, grouped by level of expertise in synthetic object realism assessment, as measured by Cohen's Kappa score (alignment) and re-evaluation rate.

- Very Realistic:
  - Textures are detailed and contribute to a lifelike appearance, with appropriate variations and reflections.
  - Soft, realistic shadows and consistent lighting that mimics real-world conditions.
  - Colors are accurate and consistent with real-world lighting, with proper shading information.
- Extremely Realistic:
  - Textures are highly detailed, with intricate variations and accurate material properties.
  - Soft, realistic shadows and consistent lighting that is indistinguishable from real-world conditions.
  - Colors are perfectly matched to real-world objects, with realistic shading and highlights.

Annotation management and quality The annotation cohort consisted of nine annotators recruited via a professional platform. They range in age from 25 to 40 with 100% identifying as female. All annotators hold a bachelor's degree or higher. They are all based in United States and fluent in English. The average annotation rate per sample is at 2 minutes. Each annotator completed an average of 4500 annotations over a one-month period.

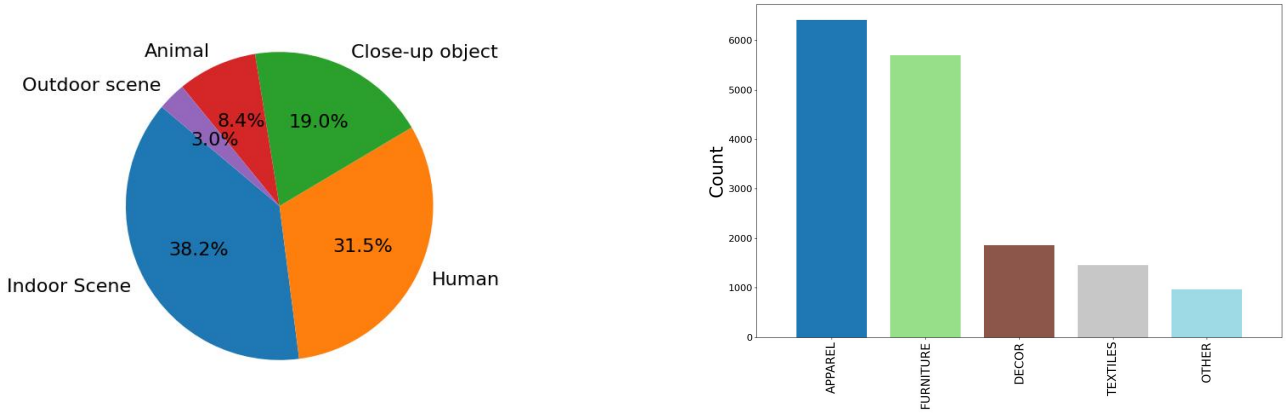
To ensure annotation quality, each annotator has received comprehensive training in task guidelines, visual examples, and tooling interactions. They were also assigned two pilot tasks where their annotations are measured by experts. During the annotation process, we track the Cohen's Kappa score and re-evaluating rate to analyze annotation quality. Overall, the Cohen's Kappa score between two annotators is 0.4 for shape realism assessment and 0.34 for texture realism assessment. A relatively small proportion of objects (10%) with large disagreement between annotators have gone through a re-evaluation. During the annotation process, we observe improving annotation quality. To further analyze the annotation quality enhancement, we conduct an experiment with two cohorts of annotators. One group is the existing annotators with experience in OcR assessment. The other group is completely new to the task. The two groups are asked to evaluate the same set of images at image and object level. As shown in Tab. 4, high-expertise annotators have noticeably higher alignment in both tasks and lower re-evaluation rate than low-expertise annotators. This suggests that evaluating object realism contributes to a better understanding of both image-level and object-level realism, and conducting pilot tasks improves annotation quality.

A.2. Additional Data Analysis

Image attribute distribution Fig. 12(a) shows the distribution of image content types. The synthetic images in our dataset exhibit a relatively balanced distribution of content types, with the exception of the *Outdoor scene* class. This comes from the nature of online product images where outdoor scenes are less common than indoor-scene images and product close-up images. We show sample counts by object group in Fig. 12(b). In addition to the four major object groups defined in Sec. 3.3.1, the grounding model also detects other object types, such as `BOOK` and `TRAY`. These are grouped under the category OTHER for simplicity.

Object shape and texture realism We show the shape and texture realism by all product categories in Fig. 13. The average OcR by product category ranges between 2.0 and 5.0. Realism scores display approximately the same pattern between the two aspects across product categories with a few exceptions such as `SUITCASE`, `SWEATER`, and `BEDDING SET`. Fig. 14(a) shows the heatmap to illustrate the number of samples per shape-texture score combination labeled by human annotators. Off diagonal values represent number of objects with discrepancy between the two realism aspects. Notably, 8% objects show considerate shape and texture realism discrepancy (absolute difference ≥ 2). We show that the realism score difference by the 17 initial object category in Fig. 14(b). In general, `APPAREL` and `FURNITURE` have higher shape and texture discrepancy than `DECOR` and `TEXTILES`.

Object recognizability As a control in the annotation task, we assess the recognizability of the objects. The annotators are asked to evaluate whether the object is recognizable as the reference object category predicted by the grounding model.



((a)) Image type distribution in annotation data.

((b)) Object group distribution in annotation data.

Figure 12. Image type and object type distribution.

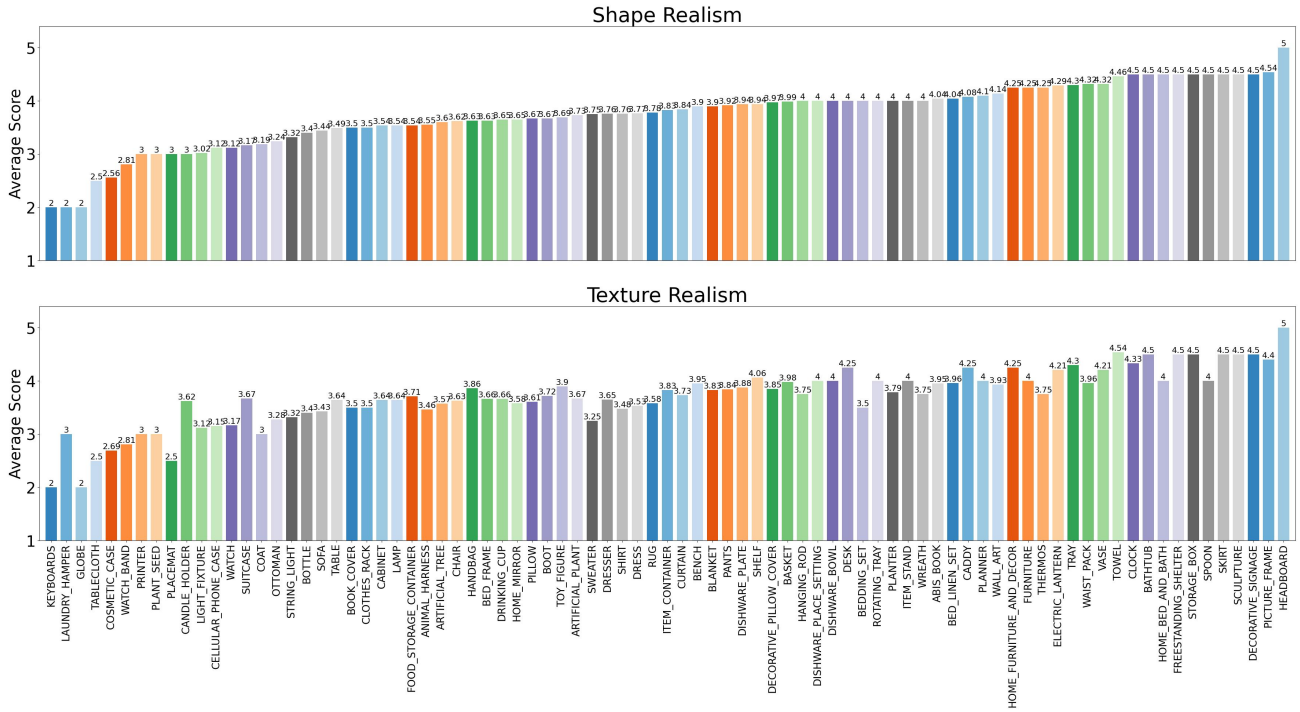
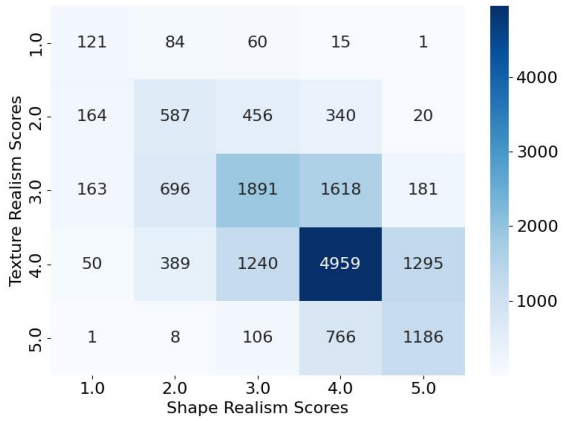


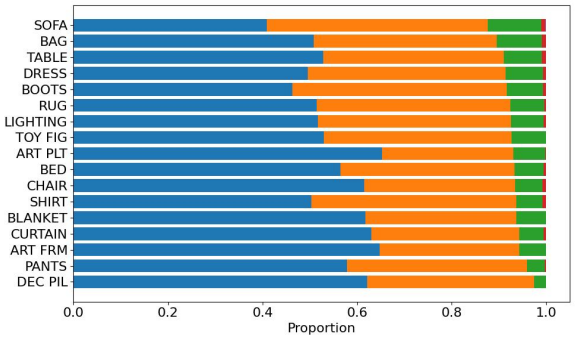
Figure 13. Realism score by object category (complete list).

We define three levels of object recognizability, i.e. `Definitely Not` the reference object type, `Could be` the reference object type and `Definitely is` the reference object type. As shown in Fig. 15(a), the majority (95.1%) of the objects are definitely recognizable as the reference category. There is no significant difference among T2I models, all reaching above 90% recognizability (Fig. 15(b)). The ranks of OcR (Fig. 4) and recognizability across models follow a similar pattern but do not closely align with each other. Compared to model breakdowns, the recognition scores show slightly greater diversity across object category breakdowns (Fig. 15(c)) with recognizable rate ranging from 78% to 100%. Apart from the least recognizable category `LIGHTING`, the rest object categories have at least 93% recognizable rate.

We remove 1.9% objects which are definitely not recognizable from all analysis for two reasons. 1. without a clear reference object category, OcR cannot be evaluated in a relatively subjective manner; 2. to avoid penalizing object distortion errors caused by grounding errors.

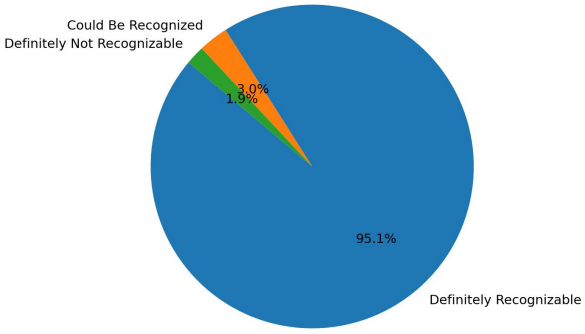


((a)) Number of samples per shape and texture realism bucket.

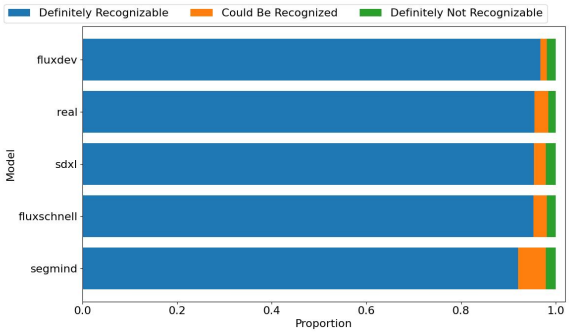


((b)) Shape-texture score discrepancy by product category. : 0, : 1 : 2, ≥ 3.

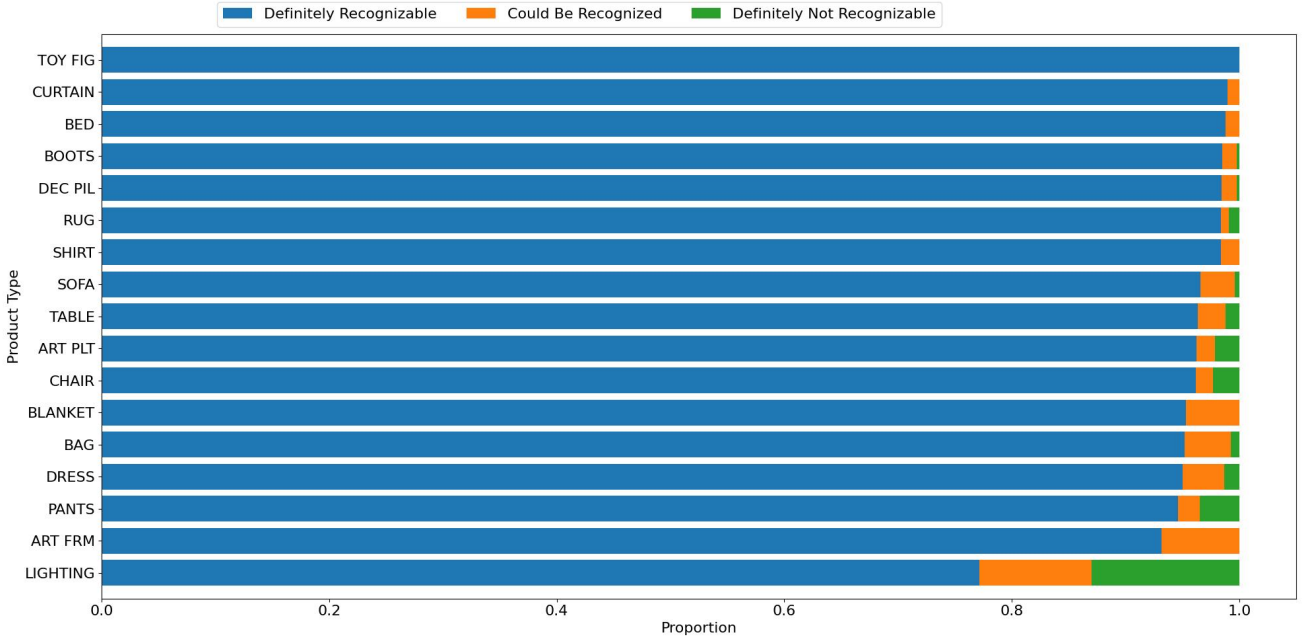
Figure 14. Plots illustrate relationship between shape and texture realism scores.



((a)) Overall.



((b)) By T2I model.



((c)) By object category.

Figure 15. Data distribution for recognizable scores.

B. Further Details on OLIP

Bias to T2I models We analyze potential OLIP model bias toward T2I models (see Tab. 5) by evaluating OLIP's performance across T2I models and comparing it to the overall results presented in in Tab. 3. The results show no significant discrepancies in OLIP's alignment with human annotators among T2I models, indicating its robustness to variations in models.

	All models	fluxdev	fluxschnell	sdxl	segmind
MAE ↓	0.118	0.113	0.113	0.116	0.136
ACC ↑	0.539	0.500	0.491	0.533	0.574

Table 5. OLIP performance consistent across T2I models.

Implementation Details OLIP model is trained with batch size 64 on one A100 GPU card. We train for maximum 30 epochs. Early stopping is used if loss on 10% validation set shows no improvement over 5 epochs. The training is typically finished in around 7 hours.

C. LLaVA Ablations

In Tab. 6, we present an ablation study of the LLaVA baseline, examining four key components:

(i) model architecture, comparing a generalist (LLaVA-OV) and a specialist (LLaVA-Critic) model;

(ii) input format, using either the full image with an overlaid bounding box (img) or just the cropped region (crop);

(iii) in-context sample size, testing both zero-shot and five-shot settings; and

(iv) query format, employing either the original question posed to annotators ('base') or a rephrased version (s5). Refer to Appendix D for more details on the prompts used.

We evaluate these baselines using Mean Absolute Error (MAE) and Accuracy (ACC) metrics for Shape, Texture, and their combination. As expected, increasing the number of in-context examples generally improves performance. In the zero-shot setting, the generalist LLaVA-OV model performs the best. However, in the few-shot scenario, the specialist LLaVA-Critic model achieves superior results, suggesting its enhanced capability in handling complex, specific evaluation tasks. Interestingly, the input format and query phrasing appear to have less impact on performance in the zero-shot setting.

Model	Size	Input	Shots	Query	Shape		Texture		Joint	
					MAE ↓	ACC ↑	MAE ↓	ACC ↑	MAE ↓	ACC ↑
LLaVA-OV	7b	img	0	base	**0.1780**	0.2729	0.1721	0.3169	0.1746	0.3732
LLaVA-OV	7b	img	0	s5	0.2824	0.2940	0.2474	0.2799	0.2521	0.2641
LLaVA-OV	7b	crop	0	base	0.1780	0.2729	0.1712	0.3134	0.1742	0.3750
LLaVA-OV	7b	crop	0	s5	0.2797	0.2975	0.2478	0.2835	0.2518	0.2658
LLaVA-Cr	7b	img	0	base	0.1851	0.2588	0.1840	0.3926	0.1830	0.3644
LLaVA-Cr	7b	img	0	s5	0.2907	0.3099	0.2390	0.4102	0.2281	0.3539
LLaVA-Cr	7b	crop	0	base	0.1838	0.2623	0.1831	0.3908	0.1836	0.3627
LLaVA-Cr	7b	crop	0	s5	0.2925	0.3099	0.2403	0.4084	0.2276	0.3539
LLaVA-OV	7b	img	5	base	0.1829	0.2764	0.1655	0.2958	0.1662	0.3891
LLaVA-OV	7b	img	5	s5	0.2273	0.3099	0.1967	0.3187	0.2122	0.3556
LLaVA-OV	7b	crop	5	base	0.1829	0.2782	0.1659	0.2940	**0.1660**	0.3891
LLaVA-OV	7b	crop	5	s5	0.2287	0.3116	0.1981	0.3169	0.2127	0.3556
LLaVA-Cr	7b	img	5	base	0.1860	0.2993	**0.1611**	0.2957	0.1708	0.3855
LLaVA-Cr	7b	img	5	s5	0.2040	**0.3433**	0.1681	0.4542	0.1784	**0.4137**
LLaVA-Cr	7b	crop	5	base	0.1860	0.2993	0.1615	0.3662	0.1710	0.3843
LLaVA-Cr	7b	crop	5	s5	0.2058	0.3415	0.1686	**0.4595**	0.1797	0.4067

Table 6. Comparison of Mean Absolute Error (MAE) and Accuracy (ACC) metrics across LLaVA variants. The analysis considers model architecture (OV [19] and Critic [36]), input type (full image with bounding box (img) vs. cropped region (crop)), sample size (zero-shot vs. five-shot), and query format (base vs. re-phrased (s5)). Cr: Critic.

D. LLaVA-OV and LLaVA-Critic Prompting

System Prompt - img variant

You are an AI visual-language assistant that analyzes objects in images and helps evaluating the quality of objects in AI generated images. You will receive an image with a red bounding box limiting the object of interest. You will receive a question and your task is to answer with the most appropriate score.
If you are provided with a `<context></context>` tag, use the content as examples of how to answer the question given an image.
`<instruction>`
We will follow a shape-surface mental model. We want to make assessments on shape realism and on surface realism of synthetic objects. The two assessments should be treated independently. When judging shape, focus on the actual shape of the object and ignore textures, color and lighting. Similarly, when judging the surface, ignore the shape of the object. For each you will be asked a question and you are supposed to input a quality assessment.
Respond using only the number corresponding with the correct choice.
`</instruction>`

System Prompt - crop variant

You are an AI visual-language assistant that analyzes objects in images and helps evaluating the quality of objects in AI generated images. You will receive a question and your task is to answer with the most appropriate score.
If you are provided with a `<context></context>` tag, use the content as examples of how to answer the question given an image.
`<instruction>`
We will follow a shape-surface mental model. We want to make assessments on shape realism and on surface realism of synthetic objects. The two assessments should be treated independently. When judging shape, focus on the actual shape of the object and ignore textures, color and lighting. Similarly, when judging the surface, ignore the shape of the object. For each you will be asked a question and you are supposed to input a quality assessment.
Respond using only the number corresponding with the correct choice.
`</instruction>`

Queries - base variant

Q1 : Is this a {`object-class`}?
1. No.
2. Maybe.
3. Yes.
Answer:
Q2 : How realistic is this object in terms of shape and structure? Object: {`object-class`}.
1. Not realistic at all.
2. Slightly realistic.
3. Moderately realistic.
4. Very realistic.
5. Extremely realistic.
Score:
Q3 : How realistic is this object in terms of texture, color and shade? Object: {`object-class`}.
1. Not realistic at all.
2. Slightly realistic.
3. Moderately realistic.
4. Very realistic.
5. Extremely realistic.
Score:

Queries - s5 variant

Q1 : Is this a {`object-class`}?
1. No.
2. Maybe.
3. Yes.
Answer:
Q2 : From 1 (not realistic at all) to 5 (extremely realistic), how realistic is this object in terms of shape and structure? Object: {`object-class`}.
Score:
Q3 : From 1 (not realistic at all) to 5 (extremely realistic), how realistic is this object in terms of texture, color and shade? Object: {`object-class`}.
Score:

Queries - joint variant

Q1 : Is this a {`object-class`}?
1. No.
2. Maybe.
3. Yes.
Answer:
Q2 : From 1 (not realistic at all) to 5 (extremely realistic), how realistic is this object in terms of shape, structure, texture, color and shade? Object: {`object-class`}.
Score:

E. More Qualitative Examples

Examples in Fig. 16 illustrate the difference between human and baseline model evaluation. The boots examples considered by humans to have the worst Shape (left), Texture (center) and Joint realism (right) received relatively high realism scores from the three baseline scoring models (ImageReward, QualiCLIP, VQAScore). Fig. 17 shows examples from the full $5 * 5$ shape-texture realism score matrix. Samples closer to the off-diagonal corners indicate a higher shape-texture realism discrepancy. We show examples with holistic image realism and dense OcR in Fig. 18 and Fig. 19, illustrating that holistic image realism is not determined by realism of any single object. Fig. 20 contains qualitative examples of OcR rankings across scoring methods including human, baseline models and our OLIP. Comparing to baseline models, OLIP has a higher consistency with human preference in OcR.

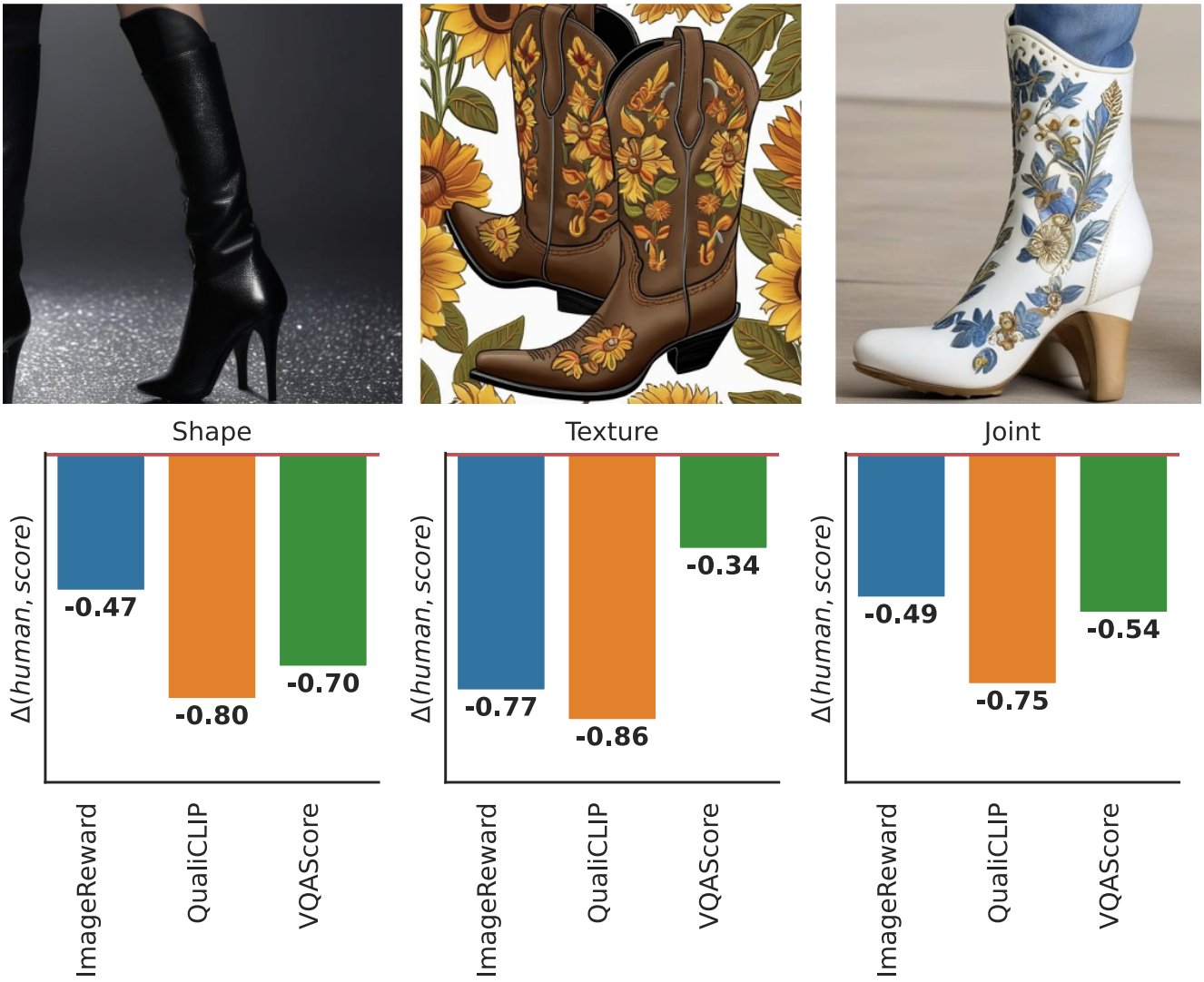


Figure 16. On the top we show the three samples in our test set considered by humans to have the worst Shape (left), Texture (center) and Joint realism (right). Below we show the Δ between the human score and the scores produced by three existing scoring models (ImageReward [37], QualiCLIP [3], VQAScore [22]) arranged to follow the images above. Score range between 0 (worst realism) to 1 (real). Notice how current scoring models align poorly with human perception of realism in synthetic images.

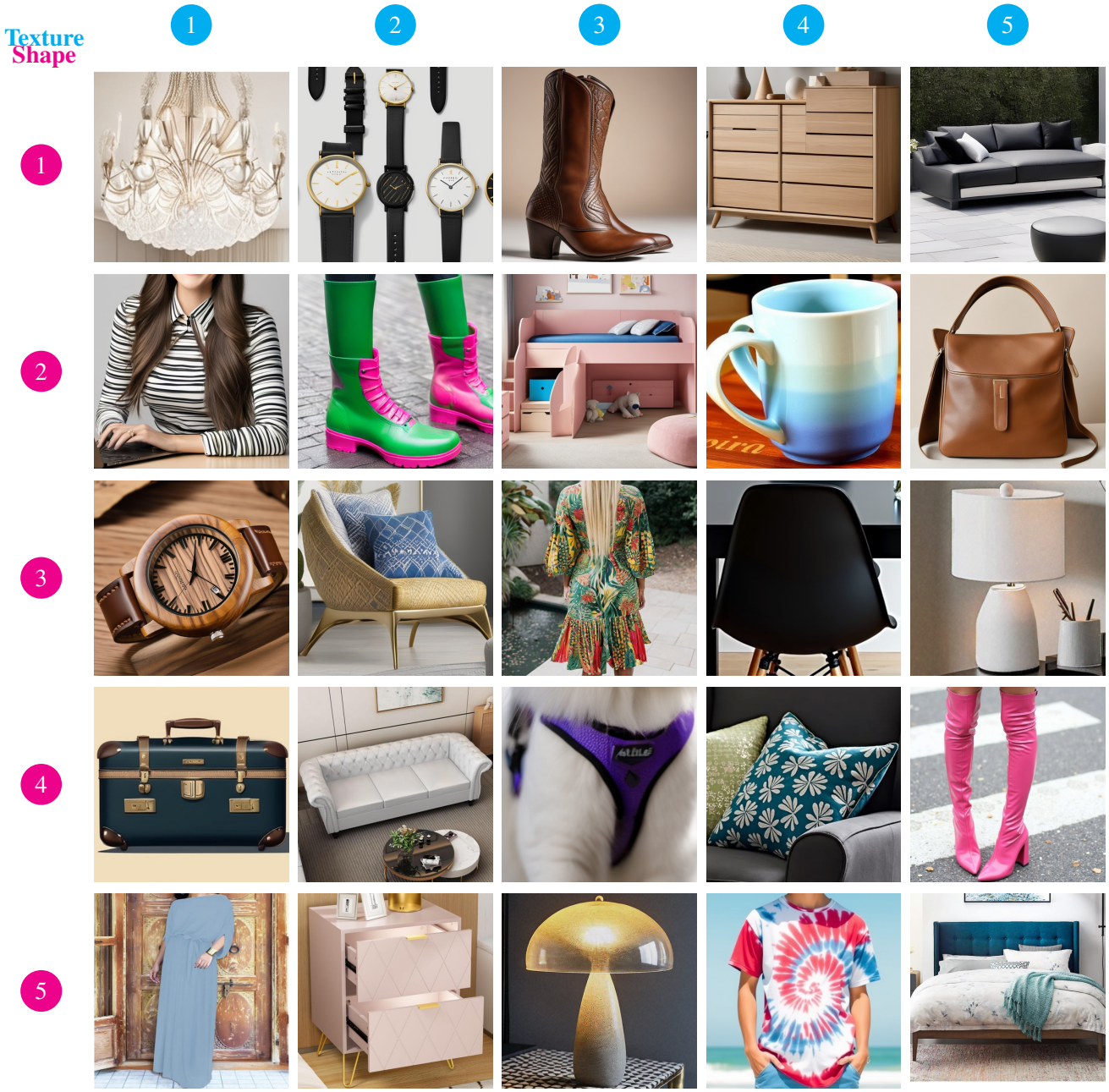
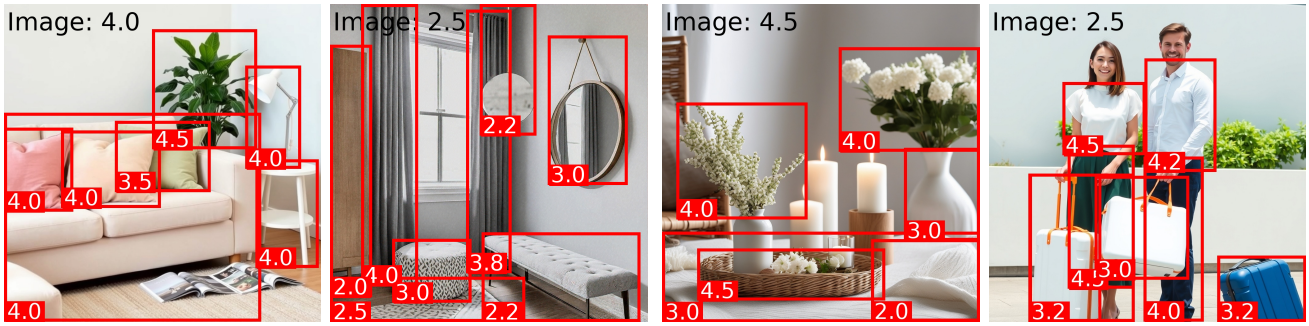


Figure 17. Dataset examples ordered by shape and texture realism. Off diagonal images present discrepancy between the two aspects labelled. Realism increases towards the bottom-right corner.



((a)) Img.: 4.0, Avg. Obj.: 4.0 ((b)) Img.: 2.5, Avg. Obj.: 2.8 ((c)) Img.: 4.5, Avg. Obj.: 3.4 ((d)) Img.: 2.5, Avg. Obj.: 3.8

Figure 18. Object vs. holistic realism. (a) consistently high holistic and object scores; (b) consistently low holistic and object scores; (c) high holistic score but relatively low object scores; (d) low holistic score but relatively high object scores.

Figure 19. Object vs. holistic realism. For better illustration, we use 3 colors and 2 box types to represent different levels of holistic and object realism. ■: low realism (score $<$ 2.33), ■: moderate realism (2.34 $<$score $<$ 3.67) ■: high realism (score $>$ 3.68). Dash box: object; solid box: image. The first row shows images with consistent holistic and object realism. The second row shows images containing objects from all realism levels.

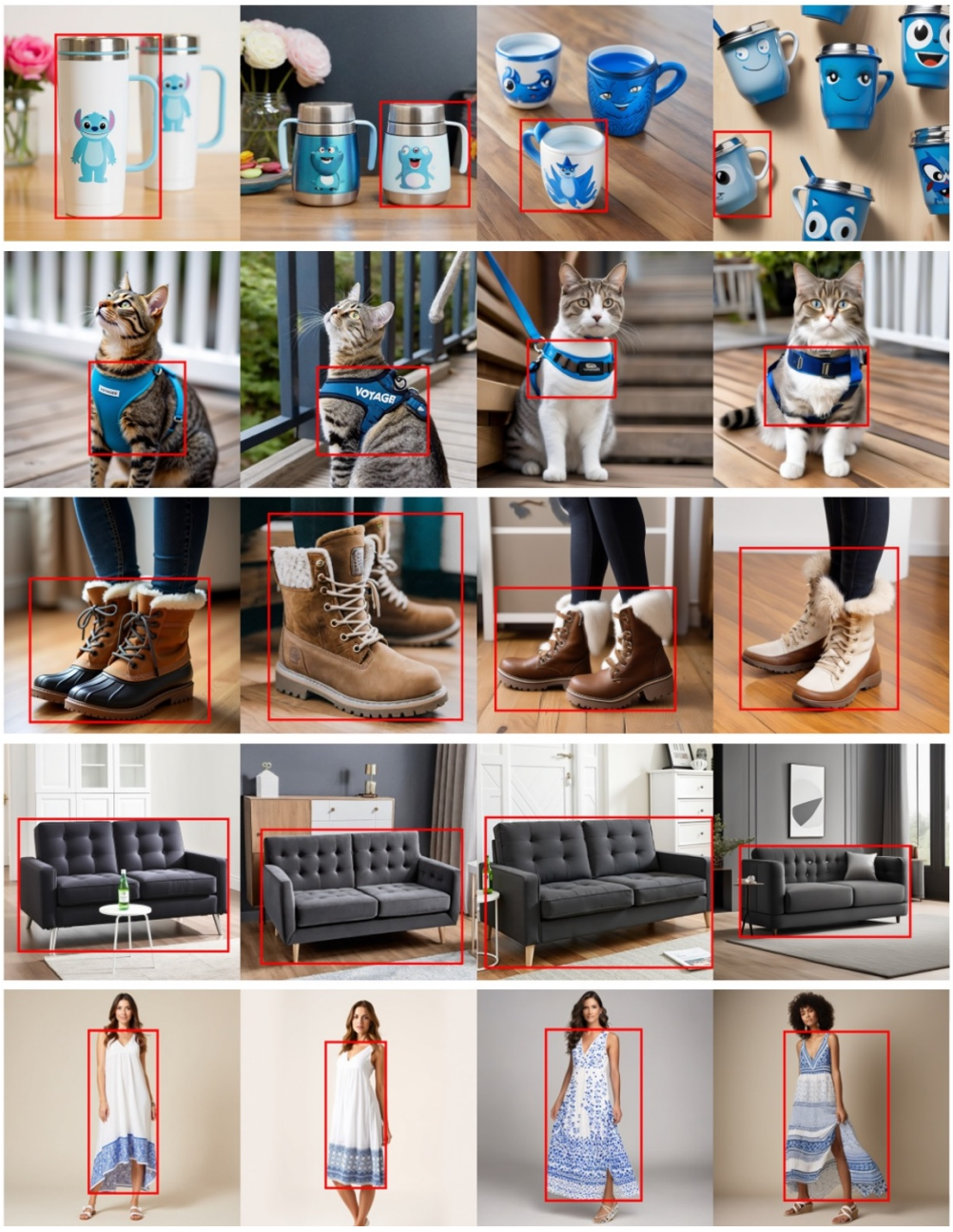


Figure 20. Object realism rank by human and scoring models. Each row shows four images (marked as A, B, C, D from left to right) generated from the same prompt by different T2I models. ImgRwd: Image Reward. Aesthetic: AestheticScore. Our model ranks OcR consistently closer to human annotators while baseline models often fail the task.