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

Anqi Liang Ciprian Corneanu Qianli Feng Giorgio Giannone Aleix Martinez

Abstract

Evaluation of synthetic images is important for both model development and selection. An ideal evaluation should be specific, accurate and aligned with human perception. This paper addresses the problem of evaluating realism of objects in synthetic images. Although methods have been proposed to evaluate holistic realism, there are no methods tailored towards object-centric realism evaluation. In this work, we define a new standard for assessing object-centric realism that follows a shape-texture breakdown and proposes the first object-centric realism evaluation dataset for synthetic images. The dataset contains images generated from state-of-the-art image generative models and is richly annotated at object level across a diverse set of object categories. We then design and train the OLIP model, an architecture that considerably outperforms any existing baseline on object-centric realism evaluation.

1. Introduction

Generative models have become excellent at synthesizing high-quality diverse images from text [28–31]. Diffusion models, a particularly effective type, can create impressive images that closely follow complex prompts [17].

Evaluating these models is critical, not only for further improvement but also for selecting best models for specific applications. To this end, several model-level evaluation metrics have been introduced, including FID [10], FID_∞ [7], and CMMD [13]. These metrics primarily measure the distribution divergence of generated vs. reference images, offering a general perspective on model quality. However, they *do not align well with human perception* and *they lack the granularity necessary to detect specific quality errors essential for subsequent improvement*.

To address these limitations, new evaluation methods and benchmarks increasingly focus on alignment with human perception and a finer taxonomy of image defects. The field has evolved from broad metrics like Aesthetic Score [1], PickScore [16], and CLIPScore [9], towards more nuanced assessments, such as dividing overall image quality into realism (fidelity) and text-image alignment [37]. More recently, even finer-grained text-image alignment evaluation systems like TIFA [12], VQAScore [22] and RichHF [21] have been introduced.

Despite this trend there remains surprisingly limited focus on decomposing the fidelity/realism aspect of generated images (Tab. 1). Unlike text-image alignment, which can anchor evaluation around each semantic concept in the text prompt [12, 22], image realism does not benefit from such straightforward anchoring. This makes guiding annotators to assess specific aspects of realism challenging.

We propose an evaluation framework, a dataset and an algorithm to *be more specific* on image evaluation. Inspired by anchoring text-image alignment evaluation to prompts, we ground our realism evaluation on individual *objects*, introducing what we term *object-centric realism*, or OcR.

For each image, OcR is assessed at the object level, with each evaluation comprising a shape score and a texture score. This separation aligns with human perception and reflects the image generation process itself, where shape (mesh) and texture (material) are typically defined independently. We developed an annotation framework that facilitates efficient and consistent collection of OcR data from human annotators, resulting in an object-centric evaluation dataset. Sec. 3 provides further details.

Additionally, we develop a set of algorithms to assess OcR, ranging from low-effort prompt engineering and zero-shot/few-shot learning approaches to our own OLIP model, which is specifically designed and trained for this task. Results demonstrate that existing image realism algorithms are largely ineffective in this context—some performing almost randomly—yet we identify promising directions for adaptation, with our model achieving *state-of-the-art* performance. This work is detailed in Sec. 4.

Our contributions are: 1. Introducing an object-level image realism evaluation framework with a systematic assessment method. 2. Compiling the first dataset dedicated to object realism assessment in synthetic images. 3. Proposing a suite of automated evaluation methods and present extensive experimental results for estimating object realism in synthetic images.

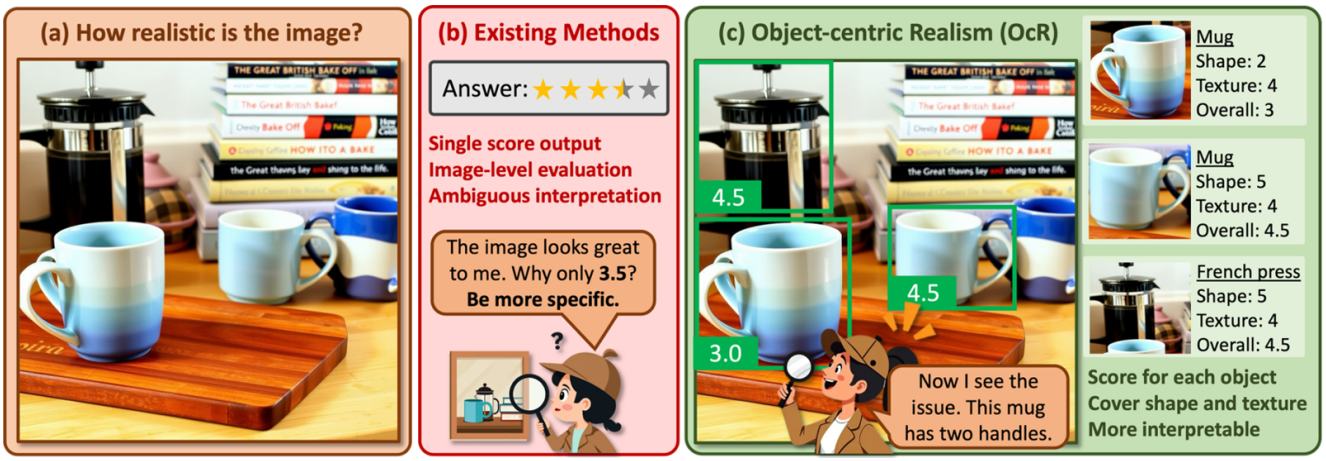

Figure 1. Overview of our method. (a) In this work, we are focusing on image realism evaluation task. (b) Existing methods, like TIFA, VQAScore, ImageReward, etc., give realism evaluation on the image-level, without interpretable details on their realism score. (c) Our method, Object-centric Realism evaluation, not only provide detailed scoring per object, but also evaluating in both shape and texture.

Realism Eval	FID	CLIP Score	Img. Rwd	TIFA	VQA Score	Rich-HF	Ours
Generic	✓	✗	✓	✗	✗	✓	✓
Obj-level	✗	✗	✗	✗	✗	✗	✓

Table 1. Comparing image realism granularity across methods.

2. Related Work

The two most commonly evaluated synthetic image attributes are: a. *image realism*, a.k.a. fidelity, which measures the overall quality and realism of generated images and b. *text-image alignment* which assesses how well the image aligns with the text prompt. Other metrics like model bias (e.g., race, gender), aesthetics, reasoning capacity, toxicity, and multilingualism [4, 18] have been also studied but not yet commonly applied.

Model-level image realism metrics like Fréchet Inception Distance (FID) [11], CMMD [13] and Learned Perceptual Image Patch Similarity (LPIPS) [34] use pre-trained neural networks to assess the quality of generated imagery. Although useful for model selection, these metrics rely on a reference dataset, operate at the model level rather than image level, and lack generalization to text-image alignment.

To address these limitations, CLIPScore [27] was proposed. It measures the cosine similarity between the embedded image and text prompt. Although impactful, it struggles on complex prompt and evaluate compositionality [35, 38].

To further enhance alignment with human perception, approaches like ImageReward [37] and PickScore [16] fine-tune vision-language models on large-scale human ratings. To get more granularity and specificity, Divide-and-conquer approaches [25, 33] use large language models (LLMs) to decompose prompts into simpler components for analysis. A notable technique within this framework is Question Generation and Answering (QG/A), exemplified by TIFA [12], DSG [6] and more recently VQA [22] where questions are extracted from prompts, and the alignment score is computed based on the accuracy of the answers by a visual question-answering (VQA) model. These works significantly pushed the field forward, but they primarily focus on granular text-image alignment. When focusing on image realism axes, these algorithms produce a holistic, generic score instead of being more specific, which we want to solve in this study by object-level realism.

The closest work to ours is RichHF-18K [21] published recently. It collects a dataset of human feedback for T2I generation including plausibility, alignment, aesthetics, and overall quality, each with a scalar rating. They also collect a heatmap annotations of regions that are implausible and misaligned to text prompt. This work is close to ours as both trying to collect more granular ratings for image realism, us with object-level realism and RichHF with artifacts heatmaps. However, the two are significantly different as we focus explicitly on object-level shape and texture quality for every single object in the image, rather than localizing artifacts in the image.

3. Object-centric Realism Dataset

3.1. Data Preparation

We illustrate the data preparation and annotation pipeline in Fig. 2. We start from a set of 2500 images of 17 fashion and home product categories sold online. Each image is captioned using LLaVA-1.6-7b with 1000 tokens [23]. Captions are passed to recent T2I models to synthesize semantically similar images. We apply four *state-of-the-art* models representing a variety of model architectures and training strategies, i.e. `FLUX-dev` [17], `FLUX-schnell` [17], `StableDiffusion-XL` (`SDXL` [2]) and `Segmind` [32]. We generate 3500 images per model and add 10% real im-

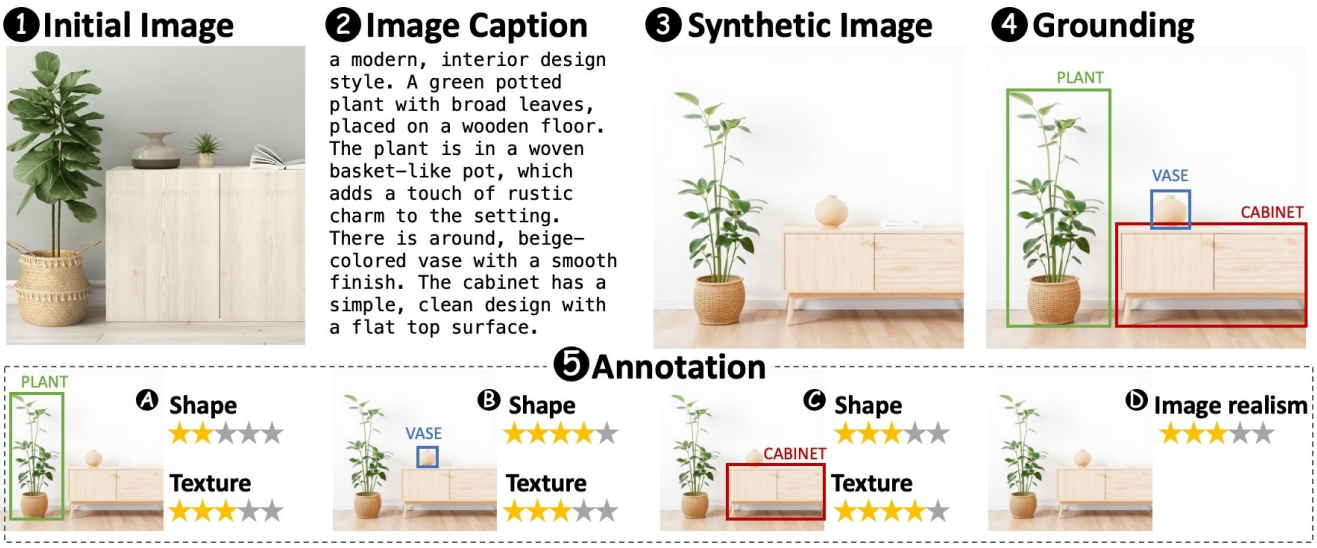

Figure 2. Data preparation and annotation pipeline. We collect a set of real images of products sold online ❶. Each real image is captioned ❷. Resulting text is used to generated synthetic images ❸ on which we ground target objects ❹. Objects are shown to humans for annotation ❺. In the annotation pipeline, objects are assessed individually for shape and texture realism. Here we illustrate the hypothetical annotation of a PLANT Ⓐ, VASE Ⓑ and CABINET Ⓒ. Additionally, we collect holistic realism labels on a small subset Ⓓ.

ages. We use a fine-tuned DETR model [5] to detect object bounding boxes. This results in $16K$ bounding boxes. We filter small objects ($< 1\%$ of the image size) and objects with low detection confidence (mAP[1] < 0.5). We further conduct a manual cleanup to remove boxes where an object is not presented regardless of the predicted object category. This removes 646 (4%) wrong boxes.

3.2. Collecting Human Assessment

Our annotation pipeline involves assessing OcR of synthetic objects from two aspects, i.e. object shape and object texture. Annotators are asked to evaluate shape and texture photo-realism of the objects independently. This two-aspect assessment follows the shape-surface mental model in 3D graphic rendering [24] and the practice of measuring the perception of visual realism in images [26]. For each object, we present its global visual context, i.e full image, and the reference category predicted by the grounding model to the annotators. Annotators are asked to judge OcR on a 1 to 5 point Likert scale [14]. We recruit nine annotators with experience in evaluating synthetic images. Each object is first evaluated by two annotators. If the two annotations diverge largely, i.e. difference ≥ 3 for any of the 5-scale questions, the object is included in a re-evaluation task to collect a third annotation. We use the Cohen's Kappa score [8] and the rate for re-evaluation requirements to keep track of the annotators' performance.

We collected valid annotations for $15K$ objects from $14K$ images across 17 object categories. We obtain two photo-realism scores, i.e. shape realism and texture realism, for each object by averaging annotators' annotation. We further derive an average OcR score by averaging its shape and texture scores. Additionally, we collect dense object-level realism and holistic image-level realism on a subset of $1.5K$ objects from 210 images with 83 object categories being grounded [2]. The dataset is balanced across models and object categories We randomly split the dataset into train (95%) and test (5%).

The annotation alignment (Cohen's Kappa score) between two annotators is 0.4 for shape realism assessment and 0.34 for texture realism assessment, and the re-evaluation rate 10%. During the process, we observe a gradual improvement in annotation quality. This emphasizes the importance of conducting pilot tasks at the beginning of the annotation task. The interested reader should refer to Appendix A for further details of the data collection and annotation quality.

3.3. Dataset Analysis

In this section, we detail general data attributes, discuss the relationship between shape and texture photo-realism, and investigate how OcR contributes to holistic image realism.

3.3.1. Data Overview

Fig. 3 shows the distribution of the object shape and texture realism scores. Both scores are skewed to the left with relative fewer counts in the lower score range.

OcR breakdowns by generative model are shown in Fig. 4. A few observations can be made. First, the score

[1]Mean of the average precision scores for all classes.

[2]Refer to Fig. 13 in Appendix A.2 for a complete category list

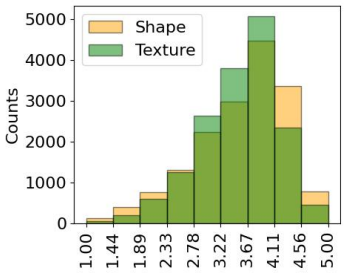

Figure 3. Overall OcR distribution.

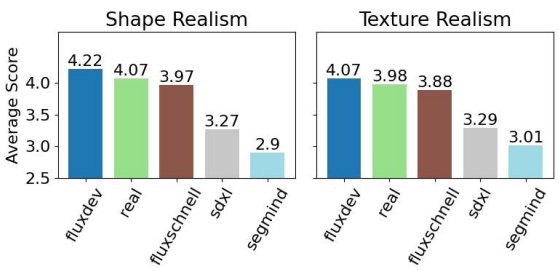

Figure 4. Shape and texture realism by model.

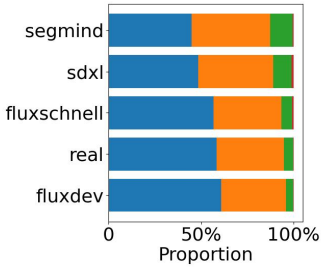

Figure 5. Shape-texture score discrepancy by model. ■: 0, ■: 1 ■: 2, ■ ≥ 3.

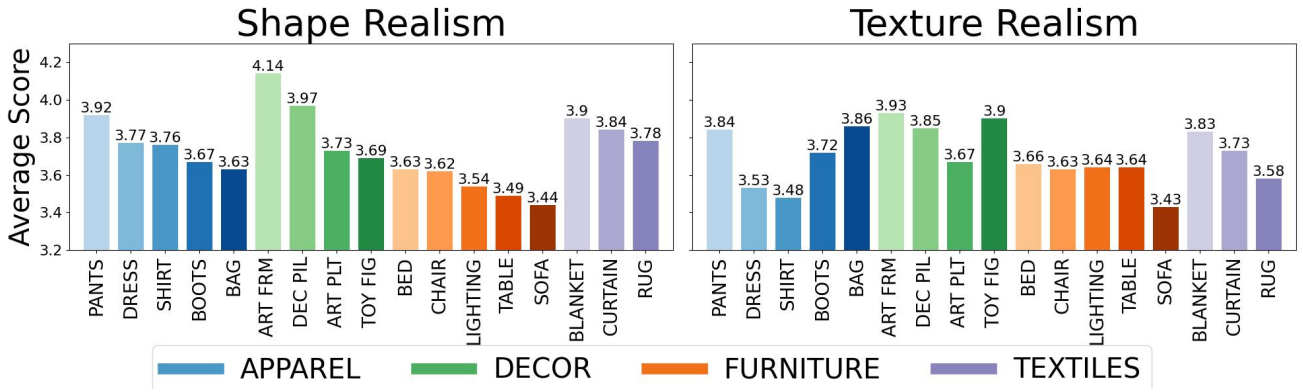

Figure 6. Shape and texture realism by object category.
ART FRM: ART FRAME, DEC PIL: DECORATIVE PILLOW, TOY FIG: TOY FIGURE.

ranks stay consistently across models for shape and texture. Besides overall expected differences coming from training protocols, data and model architecture[3], one can also observe: (i) `FLUX-dev` ranks higher than real images, which while surprising might also show capability of the model to synthesize not only artifact-free images but also images that have standard lighting, contrast and composition which humans tend to prefer. (ii) Time-step distillation barely hurts performance (`FLUX-schnell` vs `FLUX-dev`) contrary to model distillation (`SDXL` vs. `Segmind`) that considerably degrades performance.

We split the object categories into four groups, i.e. `APPAREL`, `DECOR`, `FURNITURE` and `TEXTILES` (Fig. 6). `APPAREL` categories have higher shape realism than texture realism whereas `FURNITURE` categories follow the opposite trend. This suggests that synthetic models have more challenges generating realistic shapes than texture for rigid objects (`SOFA`, `BED`) and vice versa for non-rigid objects (`SHIRT`, `PANTS`). Such discrepancy does not exist in real objects. For example, the average texture realism score of `FURNITURE` categories is 0.14 greater than their shape score for synthetic objects, but for real objects, such difference is only 0.05. `TEXTILES` categories also have consistently higher shape realism than texture realism mainly due to their relatively simple and unified structure. Additionally, decoration-related categories (`DECOR`, `TEXTILES`) have high OcR in general. This is likely because decoration objects, e.g. `WALL ART` and `CURTAIN`, usually have simple geometry, and they are typically highly stylized which makes minor distortions negligible. However, categories like `TABLE` and `BOOTS` follow specific real-world standards for structure, material and functionality. In Fig. 7 we show representative samples from assessment buckets.

3.3.2. Shape vs. Texture Realism

The correlation between shape and texture realism of an object is 0.52 [4]. This indicates that while moderately correlated, the two factors preserve independent information. There are 8% objects with high shape-texture realism discrepancy. [5] Model ranking for overall OcR (Fig. 4) are the same as their ranking for shape-texture realism alignment (Fig. 5). This suggests that models which synthesize less realistic objects also synthesize objects with higher misalignment in terms of shape and texture realism. As for more extreme cases where shape and texture difference ≥ 3, real

[3] `FLUX` models have 4× more parameters than `SDXL` and 8× more than Segmind.

[4] Spearman correlation.

[5] Absolute assessment scores between shape and texture have a difference ≥ 2. In green and red in Fig. 5.

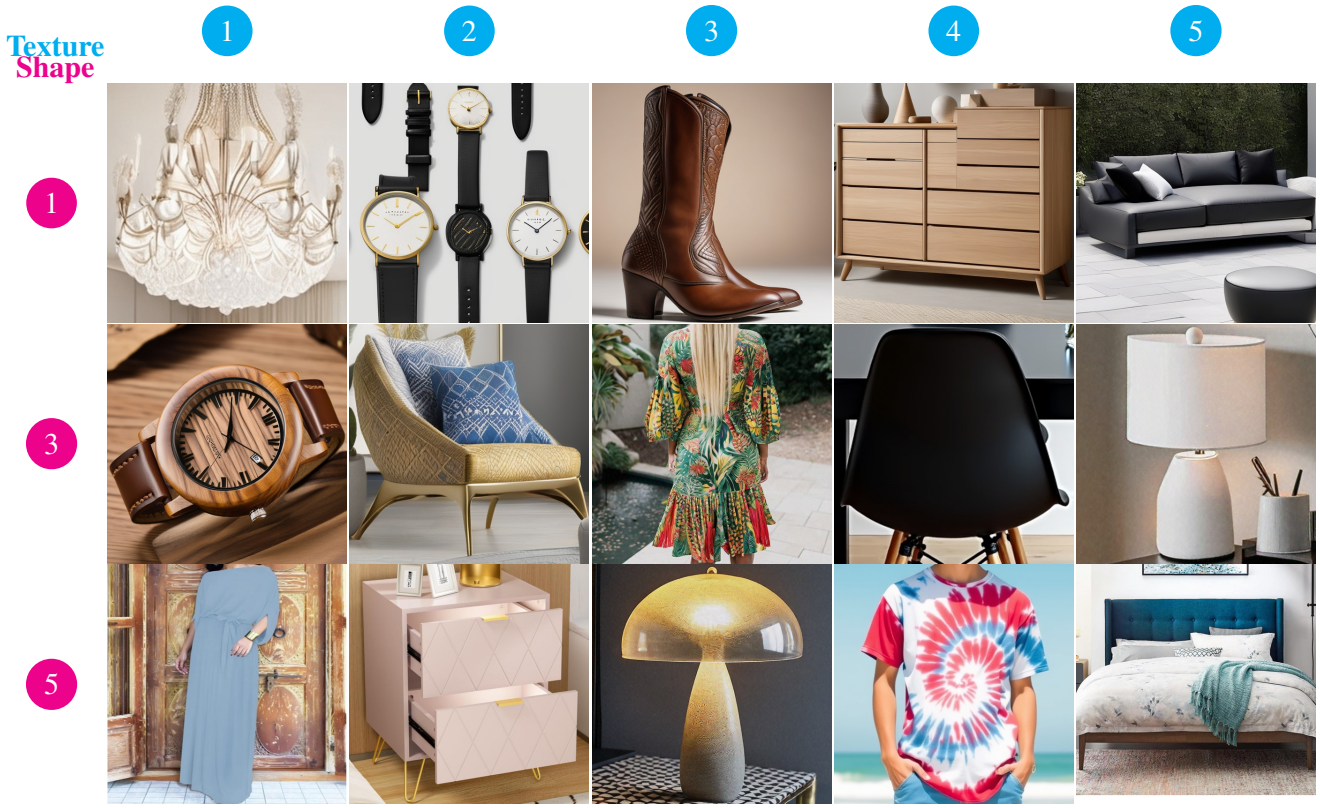

Figure 7. Dataset examples ordered by Shape and Texture realism. Off diagonal images present discrepancy between the two aspects labelled. Realism increases towards the bottom-right corner.

objects have the least prevalence in this bucket than any of the synthetic models. For object category breakdowns, we observe the same trend that categories with higher OcR have lower misalignment in shape and texture realism. Specifically, APPAREL and FURNITURE have higher shape and texture discrepancy than DECOR and TEXTILES For example, SOFA, BED have the largest shape-texture realism discrepancy whereas DECORATIVE PILLOW and ART FRAME have the least. (Refer to Fig. 14(b) in Appendix A.2).

3.3.3. Object vs. Image Realism

We ask ourselves: *"What does object realism tell us about image realism?"*. Tab. 2 shows that image realism has a moderately positive correlation with object shape and texture realism (see "All Objects"). We were further interested if this relationship depends on object relative size. It seems that, when evaluating image realism, annotators tend to focus more on larger objects. This is shown by larger correlation between scores of the largest-sized object and image scores (see "Largest" vs "Smallest").

Most practitioners would still be interested in evaluating images not objects in images. In the same table (see row "Average") we show that by averaging over all object scores in an image, one obtains a very good predictor (correlation 0.79) of the image holistic realism. This strengthens the point that image realism is mostly perceived in terms of object realism.

Type	Shape	Texture	Joint
All	0.57	0.56	0.59
Largest	0.63	0.56	0.63
Smallest	0.57	0.53	0.58
Average	0.78	0.77	0.79

Table 2. Spearman correlation between object and image realism. All: all objects; Largest: object with largest relative size (pixels from total); Smallest: object with smallest relative size; Average; correlation between average score of all objects and holistic score.

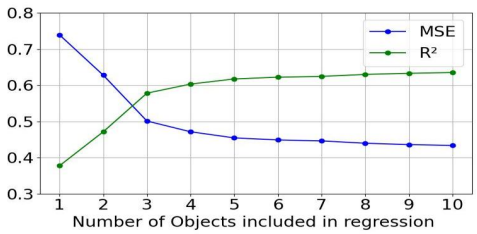

Figure 8. Feature selection for holistic and object realism score regression. Here features represent objects in the image. Most salient 5 are most informative for overall image quality.

Does one need to estimate realism over all objects in an image for a good prediction of image realism? It appears that most salient five objects are enough. We perform a linear regression analysis between holistic image realism and object realism. We remove 13 (6%) images with too few (< 5) or many (> 10) objects from the analysis. We perform a 5-fold cross-validation for conducting a Recursive Feature Elimination (RFE) by iterative removing of the least important object realism scores. Feature selection results are shown in Fig. 8. The conclusion is clear. *Object realism scores of the largest five objects of the image predict most of the holistic image realism* ($\text{MSE} = 0.45, \mathbb{R}^2 = 0.62$).

4. Estimating Object Realism

In this section we explore several automatic methods for estimating OcR in synthetic images. The approaches considered roughly fall in three main categories. First, we explore *how well do existing methods estimate realism*. We consider existing methods from four main categories: a. methods that estimate human preference, b. methods that perform image quality assessment, c. methods that compute image-to-text alignment and d. generic vision-language models tested in a zero-shot fashion. Second, we discuss and analyse strategies for *adapting vision-language models for realism estimation through few-shot in-context learning*. Finally, we design and train OLIP, our own architecture for object realism estimation. For all the experiments we keep 5% of the entire data for testing and use the rest for training. Out of the 17 object categories in the dataset, only 10 are part of the test dataset[6].

4.1. Existing Scoring Methods

We consider the following methods: (a) `ImageReward (IR)` [37] and `AestheticScore (AS)` [1] estimate human preference of synthetic images, (b) `QualiCLIP (QC)` [3] performs image quality assessment, (c) `VQAScore (VS)` [22] computes text-to-image alignment and (d) `LLaVA-OV` [19] a generalist Vision-Language Model (VLM) and `LLaVA-Critic` [36], a variant fine-tuned for scoring and ranking.

For `IR`, `AS`, `VS` and `QC` we pass a square crop of the target object. Both `IR` and `VS` compute alignment to a text prompt. The text prompts follow the questions asked to human annotators[7](see also Sec. 3). The Mean Absolute Error (MAE) is computed against the average normalized OcR score among all annotators while Accuracy (ACC) is computed over a 5-class classification problem (each 1-5 Likert assessment is a class). When multiple labels available the score is averaged and then rounded to closest integer.

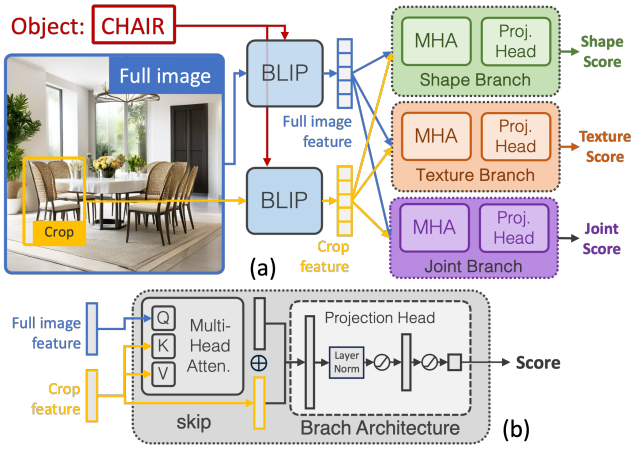

Figure 9. OLIP, the proposed model architecture. (a) The overall architecture of OLIP. (b) Detailed architecture inside each branch.

In the case of the `LLaVA` VLMs, we test several inference strategies. More specifically, when passing the image we: (i) use the entire image with an overlaid bounding box (`img`) and (ii) crop the image to only include the bounding box content (`crop`). When prompting we use: (i) same questions posed to human annotators (`base`) and (ii) employ a rephrased version that allows more flexibility for the model (`s5`). The combination of these input and prompting variations results in four inference strategies.

4.2. Adapting VLMs for OcR Estimation

To address the limitations of zero-shot approaches with general-purpose VLMs, we introduce curated few-shot in-context samples through carefully selecting diverse examples from the training set for each object category to ensure a range of shape and texture scores are presented to the models. This diversity is crucial, as random example selection tends to bias towards average values, potentially hindering the model's ability to learn scaling mechanisms for varying image and object qualities. We limit our few-shot approach to five samples and apply the same four variants (`img`, `crop`, `base`, `s5`) used in our zero-shot experiment. Best results of zero-shot and few-shot VLMs are shown in Tab. 3(a) and the complete ablation results are in Tab. 6 in Appendix C. By combining these approaches, we aim to harness the strengths of both zero-shot and few-shot learning while overcoming their individual weaknesses, potentially enhancing performance across specialized tasks.[8]

4.3. OLIP: Object-centric Scoring Model

We propose OLIP, an architecture specifically designed for realism estimation. It contains three major components: (1)

[6]SHIRT, SOFA, LAMP, HANDBAG, TABLE, PANTS, BED, DRESS, RUG, BOOTS.

[7]*Shape*: '[OBJ] with realistic shape and structure.' *Texture*: '[OBJ] with realistic texture, color and shade.'; Joint '[OBJ] with realistic shape, structure, texture, color and shade.'. [OBJ] is one of the target object categories.

[8]For an overview of prompting strategies, see Appendix D.

Method	Shape		Texture		Average		Joint	
	MAE ↓	ACC ↑	MAE ↓	ACC ↑	MAE ↓	ACC ↑	MAE ↓	ACC ↑
Random	0.345	0.223	0.337	0.232	0.341	0.227	0.342	0.211
ImageReward [37]	0.287	0.229	0.268	0.297	0.277	0.263	0.263	0.316
AestheticScore [1]	0.247	0.298	0.228	0.356	0.237	0.327	0.228	0.321
QualiCLIP [3]	0.218	0.267	0.191	0.307	0.204	0.287	0.199	0.349
LLaVA-OV (ZS) [19]	0.165	0.273	0.172	0.317	0.168	0.295	0.174	0.375
VQAScore [22]	0.192	0.302	0.159	0.382	0.175	0.342	0.164	0.407
LLaVA-Critic (FS) [36]	0.204	0.343	0.168	0.454	0.186	0.398	0.178	0.414
OLIP (Ours)	**0.136**	**0.439**	**0.123**	**0.489**	**0.130**	**0.464**	**0.118**	**0.539**

((a)) Object realism prediction performance.

Model Design	MAE ↓	ACC ↑
Full OLIP	**0.118**	**0.539**
– Full image	0.132	0.479
– Multi-head	0.130	0.486
– Deep proj.	0.136	0.446
– All	0.148	0.387

((b)) Ablation study of OLIP model variations.

Table 3. (a) Object realism prediction performance. (b) OLIP ablation study. (a) **MAE**: Mean. Absolute Error, **ACC**: Accuracy. **Shape**: '[OBJ] with realistic shape and structure.' **Texture**: '[OBJ] with realistic texture, color and shade.'; **ZS**: Zero-shot; **FS**: Few-shot; **Average**: average of Shape and Texture; **Joint** '[OBJ] with realistic shape, structure, texture, color and shade.'. [OBJ] is one of the target object categories. (b) **– Full image**: no full image to attend. **– Multi-head**: a single head for all tasks. **– Deep proj.**: a linear instead of a deeper projection head. **– All:**: no full image, no multi-head, no deep proj.

a global-local multi-modal encoder, (2) a multi-head attention and (3) a MLP regression projector. The three components combine into a data-efficient regression model for object-level shape, texture realism evaluation. An illustration of the architecture is shown in Fig. 9.

We use pre-trained BLIP [20] as the backbone on extracting the multi-modal embedding. It remains frozen throughout training. The frozen BLIP ensures that the foundational language-image representations remain consistent throughout training. Inspired by human perception on object quality which considers global context and local attention, we extract embedding for both full image and object crop, using object category name as prompt. This practice is different from previous image holistic evaluation method where prompts like "A good image" and "A bad image" are used. This gives us two 512 dimensional embedding vectors, for full image and object crop respectively.

To assess quality dimensions independently, the model applies three quality attention heads: Shape Head, Texture Head, and Joint Head. Each attention head employs multi-head attention with 8 heads and a dropout rate of 0.1. Each attention head takes the full image embedding as query and crop image embedding as key and value and is connected to a dedicated prediction head, which processes attention-enhanced embeddings through a 4-layer MLP with ReLU non-linearity. The MLP gradually reduces its hidden layer dimensions $\{1024 \rightarrow 768 \rightarrow 384 \rightarrow 1\}$ with layer normalization and dropout at rate of 0.2. At the output layer, the 1-dimension scalar is mapped to their corresponding range with a scaled and translated Sigmoid function. We use MSE loss for all regression heads. AdamW optimizer [15] is used with learning rate $= 1e^{-4}$ with cosine annealing scheduler and with weight decay $= 0.01$. See Appendix B for additional implementation details.

OLIP addresses *Object-centric Realism* through several design features: (1) usage of the full image as query in attention to provide more quality context, (2) separated attention/projection heads for texture, shape, and joint realism, as each requires different visual aspects (e.g., boundaries for shape vs appearance for texture) and (3) deeper projection head instead of linear projection to accommodate complex quality definition. We perform ablations on these designs in Tab. 3(b). Independently, each ablation causes 6-10% drop in accuracy and over 15% when ablated together.

4.4. Discussion

All results are documented in Tab. 3. Qualitative examples are shown in Fig. 10. VLM ablations are in Appendix C. We discuss them in the following.

Existing methods are approximate at best and sometimes only slightly better than random in OcR. This is true regardless of their type, i.e. Human Preference, IQA or VLM, and is particularly true for ImageReward, QualiCLIP and zero-shot LLaVA-OV. Interestingly, VQAScore and few-shot LLaVA seem to be best among this set of methods tested in the wild.

Existing methods are especially bad with unrealistic objects. One can see a few examples in Fig. 10. Most often than not, these methods fail badly with heavily distorted objects which can be problematic in practical scenarios.

All methods perform better in evaluating texture realism. This holds regardless of having been trained or not and remains true for the proposed architecture.

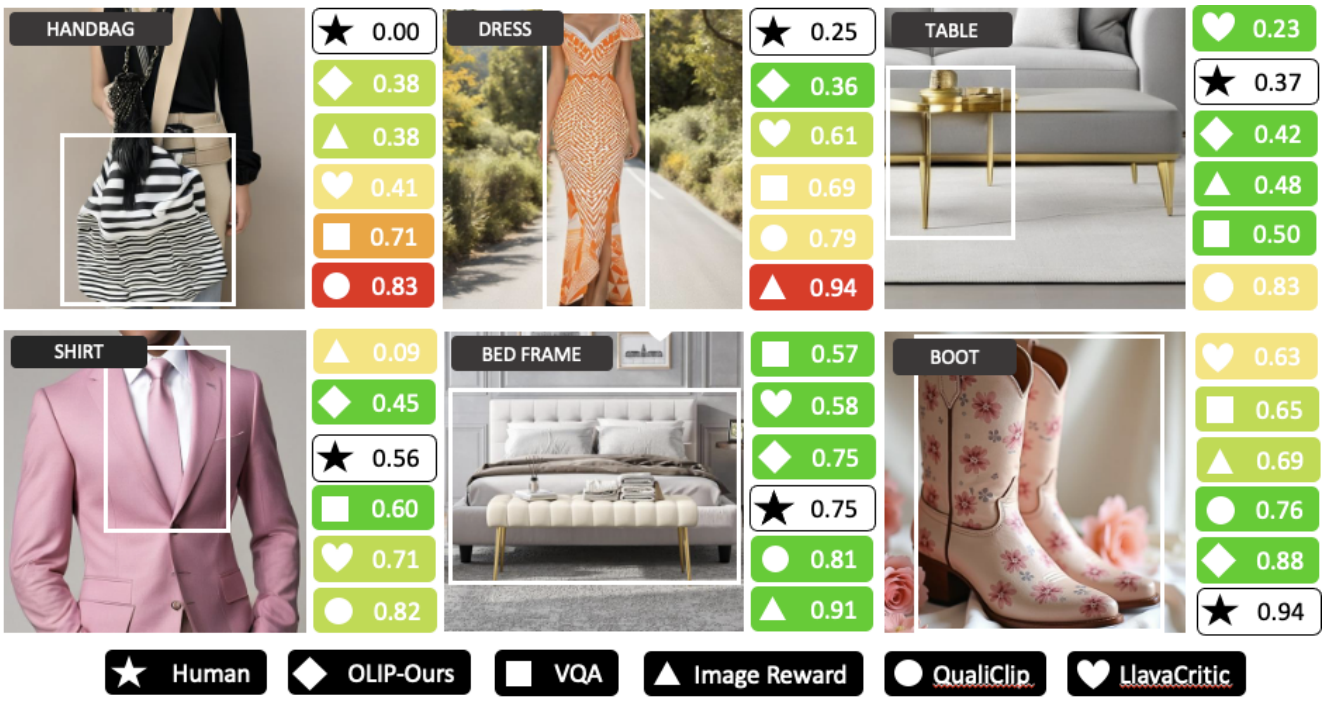

Figure 10. Object realism for a selected set of images from ground-truth (Human), selected baselines and OLIP. Color coding reflects discrepancy to ground-truth from low (green) to high (red).

Methods that need prompting[9] *benefit from aligning to both shape and texture jointly*[10]. This shows that OcR benefits from explicitly specifying both realism aspects in a single comprehensive prompt than using separate prompts.

Specialized in-context learning enhances VLM capabilities. Adapting VLMs using few-shot in-context learning effectively improves OcR compared to the zero-shot setting[11]. Specialist VLMs further outperform the generic VLMs in few-shot learning (Tab. 6 in Appendix C). These indicate that specialized vision-language models with few-shot learning represent a promising direction for developing complex multimodal scoring mechanisms.

Designing and training a dedicated architecture shows considerable improvements over baselines. Particularly effective is the combination between strong pretraining (BLIP backbone) and dedicated design (object and holistic attention). This is confirmed by best results across the board and considerable improvements on unrealistic objects.

5. Conclusion and Limitations

In this paper, we have introduced an object-level realism evaluation framework. Based on it, we have compiled the first dataset devoted to object realism assessment in synthetic images. Finally, we have discussed a suite of existing evaluation methods and proposed a model dedicated to OcR. In additional to OcR evaluation, the proposed pipeline has application opportunities in (1) localized editing for quality enhancement and (2) model improvement via finetuning or human preference alignment. These do not come without limitations. First, we base our study on a diverse but limited number of image types and object categories. Future work would be needed for real-world scale generalization. Then, additional information that object-level specificity brings comes with added computational cost. Several models passes, one per object, would be required for dense realism evaluation with a prerequisite of object grounding. This cost can be significant for certain applications. In addition, adapting to new artifacts and models is persistently unclear and difficult, which remains an on-going complexity until a powerful system that resembles human perception is established. In this work, we propose careful artifacts definition, meticulous data collection, and robust model development, with the hope that it can help our community tackle this problem. Finally, realism estimation remains a challenging problem. While we address the issue of subjectivity in annotation by focusing on objects with broad consensus in general understanding, and we show promising results in modelling, considerably more improvement has to be made until automatic realism assessment closes the gap with human perception of realism.

[9] All methods except AestheticScore

[10] See column "Joint" v.s. column "Average" of Tab. 3(a)

[11] See row "LLaVA-OV (ZS)" v.s. row " LLaVA-Critic (FS)" of Tab. 3(a)

References

[1] https://github.com/christophschuhmann/improved-aesthetic-predictor. 1, 6, 7

[2] https://huggingface.co/stabilityai/stable-diffusion-2-1. Accessed: 2023-11-15. 2

[3] Lorenzo Agnolucci, Leonardo Galteri, and Marco Bertini. Quality-aware image-text alignment for real-world image quality assessment, 2024. 6, 7, 10

[4] Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models, 2023. 2

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3

[6] Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. *arXiv preprint arXiv:2310.18235*, 2023. 2

[7] Min Jin Chong and David Forsyth. Effectively unbiased fid and inception score and where to find them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[8] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960. 3

[9] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 1

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 1

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 2

[12] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417, 2023. 1, 2

[13] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9307–9315, 2024. 1, 2

[14] Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396–403, 2015. 3

[15] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[16] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 36652–36663, 2023. 1, 2

[17] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 1, 2

[18] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Fei-Fei Li, Jiajun Wu, Stefano Ermon, and Percy S Liang. Holistic evaluation of text-to-image models. In *Advances in Neural Information Processing Systems*, pages 69981–70011. Curran Associates, Inc., 2023. 2

[19] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 6, 7

[20] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 7

[21] Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, Junjie Ke, Krishnamurthy Dj Dvijotham, Katie Collins, Yiwen Luo, Yang Li, Kai J Kohlhoff, Deepak Ramachandran, and Vidhya Navalpakkam. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2

[22] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*, 2024. 1, 2, 6, 7, 10

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2

[24] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998. 3

[25] Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[26] Paul Rademacher, Jed Lengyel, Edward Cutrell, and Turner Whitted. Measuring the perception of visual realism in images. In *Rendering Techniques 2001: Proceedings of the Eurographics Workshop in London, United Kingdom, June 25–27, 2001 12*, pages 235–247. Springer, 2001. 3

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[28] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1

[29] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022.

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[31] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1

[32] Segmind. Ssd-1b. https://huggingface.co/segmind/SSD-1B, 2023. Accessed: [11/11/2024]. 2

[33] Jaskirat Singh and Liang Zheng. Divide, evaluate, and refine: Evaluating and improving text-to-image alignment with iterative vqa feedback. *Advances in Neural Information Processing Systems*, 36:70799–70811, 2023. 2

[34] Jake Snell, Karl Ridgeway, Renjie Liao, Brett D Roads, Michael C Mozer, and Richard S Zemel. Learning to generate images with perceptual similarity metrics. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 4277–4281. IEEE, 2017. 2

[35] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 2

[36] Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models. *arXiv preprint arXiv:2410.02712*, 2024. 6, 7

[37] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023. 1, 2, 6, 7, 10

[38] 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 *The Eleventh International Conference on Learning Representations*, 2023. 2