What Makes a Good Generated Image? Investigating Human and Multimodal LLM Image Preference Alignment

Rishab Parthasarathy[1], **Jasmine Collins**[2], **Cory Stephenson**[2]
[1]MIT
[2]Databricks Mosaic AI Research
rpartha@mit.edu, jasmine.collins@databricks.com, cory.stephenson@databricks.com

Abstract

Automated evaluation of generative text-to-image models remains a challenging problem. Recent works have proposed using multimodal LLMs to judge the quality of images, but these works offer little insight into how multimodal LLMs make use of concepts relevant to humans, such as image style or composition, to generate their overall assessment. In this work, we study what attributes of an image–specifically aesthetics, lack of artifacts, anatomical accuracy, compositional correctness, object adherence, and style–are important for both LLMs and humans to make judgments on image quality. We first curate a dataset of human preferences using synthetically generated image pairs. We use inter-task correlation between each pair of image quality attributes to understand which attributes are related in making human judgments. Repeating the same analysis with LLMs, we find that the relationships between image quality attributes are much weaker. Finally, we study individual image quality attributes by generating synthetic datasets with a high degree of control for each axis. Humans are able to easily judge the quality of an image with respect to all of the specific image quality attributes (e.g. high vs. low aesthetic image), however we find that some attributes, such as anatomical accuracy, are much more difficult for multimodal LLMs to learn to judge. Taken together, these findings reveal interesting differences between how humans and multimodal LLMs perceive images.

Extended version — https://arxiv.org/abs/2509.12750

1 Introduction

Recent advancements in diffusion models have led to the rapid spread of text-to-image models, such as Stable Diffusion (Rombach et al. 2021; Podell et al. 2023) and FLUX (Black Forest Labs 2024), which have far greater capabilities than image generation models of the past. However, these more advanced models have failure modes that can be difficult to detect during evaluation. In some cases, the images may contain artifacts, not follow the prompt precisely, or may not suit the aesthetic desires of the user (Cao et al. 2024; Jiao et al. 2024; Huang et al. 2024; Zhou et al. 2024). To this end, recent work has increasingly focused on improving evaluation, which can be divided into two categories: human and automated (Lee et al. 2023; Zheng et al. 2023). Although human evaluation is still considered the gold standard, undertaking human studies has become challenging with the rapid speed of model iteration.

As a result, recent works prioritize automated evaluation, often through LLM-as-a-Judge, where a large pretrained or finetuned language model is used to judge the quality of model outputs as a proxy for human preferences (Zheng et al. 2023). In the case of evaluating generative diffusion models, this process typically involves either prompting a small contrastively trained multimodal LLM or a large frontier LLM with the generated image, and using the model to provide judgments on the quality of outputs. While these approaches have great potential for evaluating diffusion-generated images, recent works focus primarily on their alignment with humans on separately evaluated tasks like bias, prompt alignment, safety, and perceptual quality, with little focus on the interaction of these tasks in producing an overall image judgment (Chen et al. 2024b; Wu, Huang, and Wei 2024; Huang et al. 2023; Lee et al. 2024).

In this work, we study which aspects of an image are important to humans and LLMs for judging pairwise image quality. Specifically, we curate a dataset of pairs of images generated by open weight diffusion models, which are then evaluated by human raters on six attributes typically implicated in the human evaluation of images: aesthetic quality, artifacts, anatomical issues, correctness of compositional relationships, object adherence, and style. Since this dataset uniquely has human ratings on all of these attributes for the same image pairs, we can construct inter-attribute correlations for both LLMs and humans, allowing us to study whether LLMs and humans share a standardized paradigm for evaluating images. We find that even though LLMs may seem to have outputs aligned with humans, their internal image evaluation metrics may not.

We then generate synthetic datasets based on four of our image quality attributes–aesthetic, anatomy, compositionality, and style–to further study which aspects of image evaluation are more or less challenging to multimodal LLMs. Each of these tasks involves a single-image task instead of pairwise comparison, where the LLM is either prompted to output a Yes/No rating or a numerical rating from 1-10. Using these synthetic single-axis datasets, we find that certain tasks like evaluating compositional correctness and style prove easy for LLMs to handle, but a significant gap still

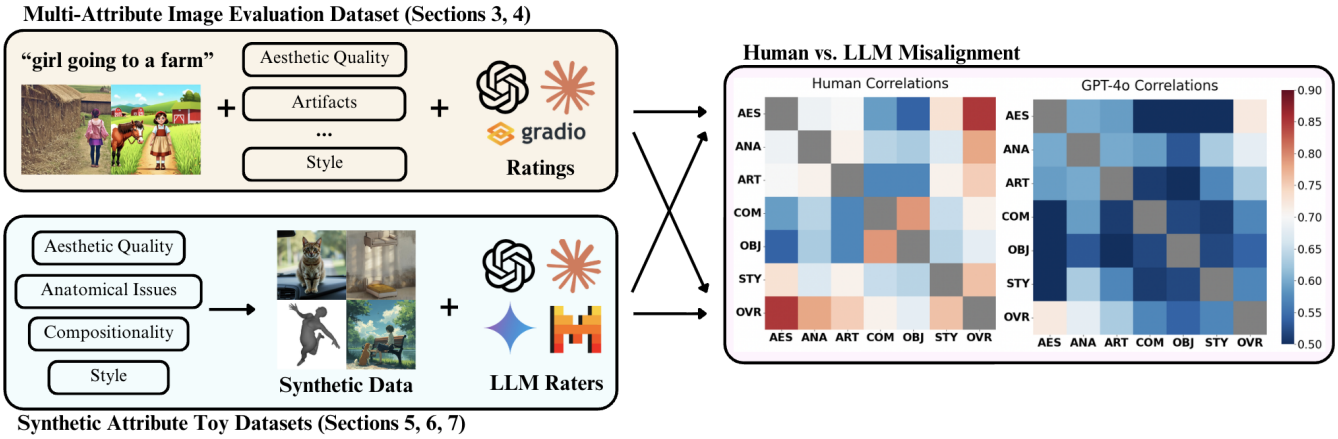

Figure 1: We collect pairwise preference data across multiple axes of image quality and compute correlations between different axes of image quality. We also generate synthetic toy datasets for four of the image quality axes. Our approach enables us to determine which image quality attributes are most relevant to how humans and multimodal LLMs judge images, as well as which quality attributes are most difficult for LLMs to learn and judge, determining alignment gaps between human and LLMs.

remains in judging anatomy validity, where humans outperform LLMs.

In summary, we investigate whether multimodal LLMs share the same attribute values as humans when evaluating diffusion-generated images using a custom curated dataset where each image pair is rated by both multimodal LLMs and humans along seven different axes. We also conduct a fine grained evaluation of these same frontier LLMs on toy versions of the tasks evaluated before, comparing them to small finetuned models and human raters.

Our key findings are as follows:

- LLMs and finetuned reward models perform similarly as overall image judges, and align well with humans.
- LLMs and humans show significant differences in how specific image attributes influence their overall rating.
- LLMs struggle to generalize when evaluated on basic tasks including judging anatomical correctness, image style, and aesthetics, which are easy for humans.

These findings imply that even though small models and frontier LLMs can match human performance on the overall image rating task, they process the contents of the image quite differently, and can fail to generalize in surprising ways. Our work suggests that despite impressive overall benchmark performance, there is further work to be done in aligning multimodal LLMs with human preferences.

2 Related Work

2.1 Multimodal Modeling

Recent research in multimodality has focused on adding visual input to transformer models originally designed for language. These models can be fully pretrained from scratch on interleaved multimodal data, like Chameleon (Team 2024; Bai et al. 2023), or take pretrained LLMs and add additional modalities by training vision adapters, like LLaVA (Liu et al. 2023b,a) and Flamingo (Alayrac et al. 2022). These vision adapters transform visual image inputs into the embedding space of the pretrained LLM, often using a small contrastive vision-language model (VLM) like CLIP (Dai et al. 2023; Radford et al. 2021; Li et al. 2022). In this work, due to compute and data constraints, we focus on frontier multimodal LLMs and LLaVA-based models that we train with a LLaMA-3-8B backbone (Dubey et al. 2024).

2.2 Reward Modeling and Image Generation

Current works on generative models focus on aligning them with human preferences, which directly matches real world use cases. Reinforcement learning-based approaches like RLHF and DDPO have shown that with adequate reward signal, both LLMs and generative text-to-image models can be tuned to align with arbitrary human preferences, highlighting the need for powerful reward modeling (Ouyang et al. 2022; Black et al. 2024; Wallace et al. 2024; Lambert et al. 2024; Yuan et al. 2024; Wang et al. 2024a; Stiennon et al. 2020).

The first class of such reward models are based off of backbones like CLIP or BLIP, which are pretrained using a contrastive training loss. Here, groups of labeled images are ingested simultaneously to align the model along a general multimodal reward axis (Li et al. 2022; Radford et al. 2021; Xu et al. 2023). Finetuned reward models like ImageReward (Xu et al. 2023) and PickScore (Kirstain et al. 2023) instead focus on the task of human preferences, and finetune BLIP (Li et al. 2022) and CLIP (Radford et al. 2021) on curated datasets of image pairs, labeled with human preferences. These classifiers lose the general image labeling capabilities of the BLIP/CLIP backbones, but are found to be significantly more aligned with human preferences (Xu et al. 2023, 2024; Kirstain et al. 2023; Wang et al. 2024b; Dzabraev, Kunitsyn, and Ivaniuta 2024). On top of that, with recent advancements in scaling multimodal LLMs (Bai et al. 2023; Dubey et al. 2024), recent research

has also focused using large frontier LLMs as generalist reward models, where a finetuning-free approach allows for plug-and-play usage with various evaluation and reinforcement learning pipelines (Jiao et al. 2024; Huang et al. 2023; Cao et al. 2024). This pipeline typically involves using large models like GPT-4V as a rater that iteratively improves generated data, such as in the GORS finetuning pipeline from T2I-CompBench (Huang et al. 2023).

2.3 LLM-as-a-Judge

In sync with research on frontier LLMs as reward models, recent work has proposed a new evaluation paradigm termed "LLM-as-a-Judge". Due to the high costs of human evaluation, LLM-as-a-Judge provides a platform for rapid evaluation of subjective tasks, using a frontier LLM as a proxy for human preferences. In NLP, LLM-as-a-Judge has seen rapid adoption in platforms like MT-Bench (Zheng et al. 2023) because frontier LLMs are highly aligned with human preferences on text (Son et al. 2024; Wei et al. 2024). However, research in multimodal LLM-as-a-Judge for generated images remains a more open area, as multimodal frontier models are relatively recent advancements. The majority of multimodal LLM evaluations still focus on generating text, while a few works like MJ-Bench and VisionPrefer instead focus on directly augmenting diffusion model capability (Chen et al. 2024b; Wu, Huang, and Wei 2024; Chen et al. 2024a; Wang et al. 2025). These existing pairwise evaluations investigate a number of different image axes, but none contain multiple ratings and evaluations for the same pair of images (Kirstain et al. 2023; Jiao et al. 2024; Chen et al. 2024b; Huang et al. 2024, 2023). Instead, they feature different challenging image pairs for each task, such as safety, image difference captioning, or image quality rating (Wang et al. 2025; Chen et al. 2024a; Jiao et al. 2024; Zhang et al. 2025; Han et al. 2024). While these datasets can thus provide insight into individual tasks, they struggle to provide insight into how LLM decisions are made on the same set of images. This paper hence focuses on a small set of axes, which are all evaluated on the same set of images, to evaluate how the learned preferences of multimodal LLMs align with humans in the task of image evaluation, and highlighting where multimodal LLMs lag behind human capabilities.

3 Human Curated Evaluation Dataset

We curate an evaluation dataset where each data point has seven human-generated annotations: aesthetic quality, presence of artifacts, issues with anatomy, compositional correctness, object adherence, style, and overall rating. Compositional correctness refers to whether the image correctly describes compositional aspects in the prompt like direction and location of objects. Each image pair is sampled using a prompt from PartiPrompts and generated from a sweep of 14 top open models from the ImgSys leaderboard (see Appendix Section 1 for the list of models and generation settings) (Yu et al. 2022; fal 2025). These images are provided to 40+ human raters (AI researchers who rated 30+ images each) pairwise in a UI (detailed in Appendix Section 1) where users are asked to rate the images on each

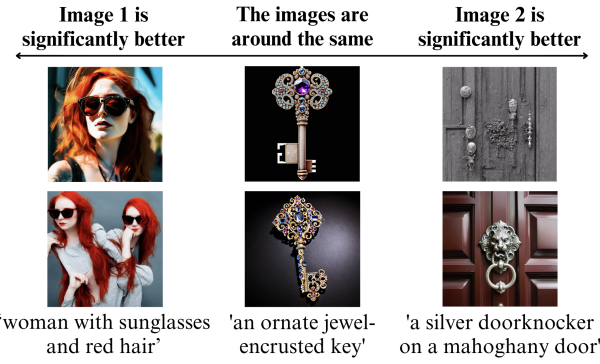

Figure 2: A sample of image pairs across the spectrum of the dataset from similar image quality to each of image 1 and image 2 being significantly better.

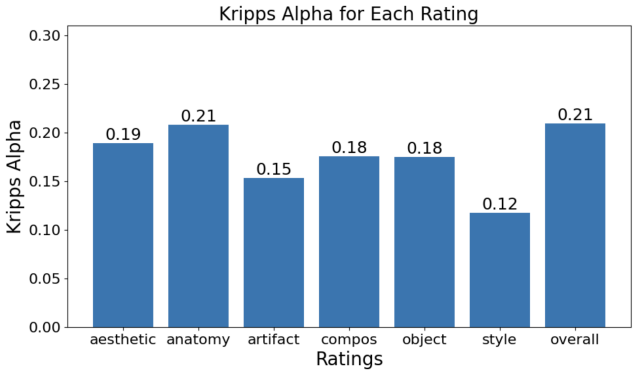

Figure 3: Kripp's Alpha for all seven of the ratings axes evaluated in this study. We find that all of the ratings have Kripp's Alpha significantly above 0, indicating that there is statistically significant inter-rater agreement.

of the discussed rating axes using a five-level Likert scale. For special cases, a "not relevant" rating is also provided. A sample set of ratings is provided in Fig. 2. We provide additional example ratings in Appendix Section 3 and specific task definitions in Appendix Section 2.

3.1 Verifying Dataset Quality

After data collection, we verify that the generated dataset has both inter-rater alignment and aligns with existing human datasets. Since each pair of images is only evaluated by one rater, we evaluate inter-rater alignment using Krippendorff's Alpha (Kripp's Alpha), which is a statistical value measuring whether ratings agree more or less than random (Nassar et al. 2019). Kripp's Alpha is calculated as

$$\alpha = \frac{p_a - p_e}{1 - p_e} \tag{1}$$

where p_a is the observed probability of agreement and p_e is the expected probability of agreement. We calculate Kripp's alpha over the ratings for each model rather than each image pair as a proxy for evaluating inter-rater agreement.

A Kripp's alpha value of 0 indicates negligible agreement, whereas ratings of 1 and -1 indicate perfect agreement and

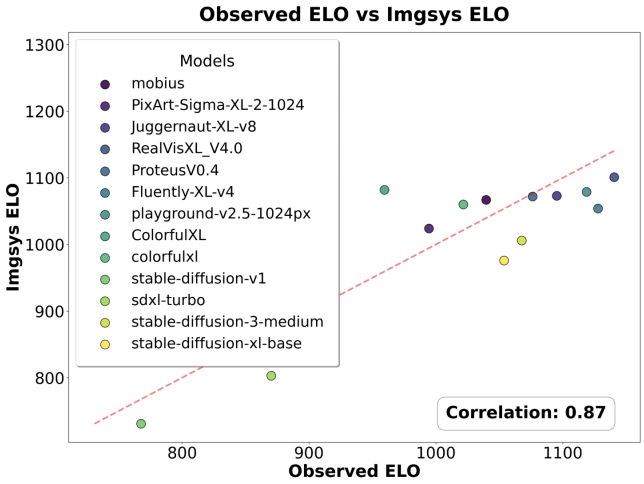

Figure 4: We find that ImgSys Elo scores are strongly correlated with Elo scores calculated from dataset.

disagreement, respectively. As seen in Fig. 3, each evaluation axis has Kripp's alpha far above 0, indicating good inter-rater agreement. We note that style and artifacts have lower Kripp's alpha than the other categories – we attribute this to raters frequently saying that these categories were not relevant, reducing sample size and increasing noise. For exaple, when certain prompts did not denote a style like "in the style of Picasso", users frequently selected "Not Relevant".

We also compare model Elo on our dataset against the larger ImgSys dataset in Fig 4, where ImgSys is a standard arena for rating generative image models (Zheng et al. 2023; fal 2025). This serves as a sanity check to validate our rater pool and data collection methodology. We expect that the ELO rankings in the two datasets should be correlated if our raters are similar to the much larger ImgSys pool. Indeed, we find that ImgSys Elos are highly correlated (Pearson's correlation coefficient $= 0.87$) with the overall Elos calculated on our multi-image dataset, suggesting our pool of raters has similar preferences to ImgSys raters. Finally, we quantify bias in our rating scheme (due to the rater pool consisting only of AI researchers) by taking a subset of the dataset and comparing ratings with a dataset of non-AI researcher ratings (college students). We find that the correlation ranges from 0.75 and 0.88 across categories and Cohen's Kappa ranges from 0.34 to 0.7, indicating significant agreement.

4 Results on Frontier Models

In this section, we aim to answer these two guiding questions: (1) how do large frontier models perform on an ensemble of image evaluation tasks, (2) what kind of relationships do large frontier models discover between image quality attributes and how do those compare to humans? We also compare to several smaller models finetuned image reward models on the overall rating task, but these models are unable to provide finer grained results.

4.1 Prompting Strategy

LLMs and specialty image reward models require different prompting strategies due to their different training and architectures. Large frontier models can accept multiple images as inputs, so we prompt these models with the two images and request them to output results on a five-item Likert scale specified using special tokens (identical to the dataset collection). These special tokens can then easily be parsed into an identical Likert scale as that used by the dataset.

For single-image finetuned reward models, we instead pass each image into the model separately. The image with the higher rating is considered to be better. We detail the specific prompts used in Appendix Section 5.

4.2 Frontier Model Evaluation

We evaluate models using two metrics: the Pearson correlation coefficient (Tab. 2) and Cohen's Kappa (Tab. 1). Cohen's Kappa reflects rater agreement, whereas Pearson correlation describes the similarity of the two distributions.

We find that across all tasks, GPT-4o (OpenAI 2024b) outperforms the other foundation models. However, the specially trained PickScore (Kirstain et al. 2023), HPSv2 (Wu et al. 2023), VQAScore (Lin et al. 2024b), and ImageReward (Xu et al. 2023) outperform Gemini-1.5-Flash (Google 2024), GPT-4o-mini (OpenAI 2024a), and Pixtral (Agrawal et al. 2024) on the overall task, showing that fine-tuned reward models can still outperform foundation models given enough human preference data. However, unlike foundation models, these small models cannot perform tasks outside of overall judgement, as these are out of the models' domain.

We also find that object adherence is the easiest of the six subtasks, as the Pearson correlation and Cohen's Kappa are the highest. We also find that style, artifacts, and aesthetic are consistently more difficult judgment subtasks.

4.3 Correlating Image Quality Attributes

To evaluate correlations between image quality attributes, we calculate the Pearson correlation coefficient between tasks for both human ratings and the best performing general LLMs: GPT-4o and Claude-3.5-Sonnet.

As shown in Fig. 5, humans exhibit a strong correlation between overall ratings and all attributes other than object adherence. This possibly signifies that object adherence is relatively unimportant for human image preferences, which is interesting given that object adherence is the easiest task for LLMs to evaluate. On the other hand, aesthetics are highly correlated with overall rating. We also find correlations between individual tasks. For example, human raters associate composition with object adherence, and we observe weak associations between aesthetic and style, as well as between anatomy and diffusion artifacts.

Interestingly, LLMs lack a similar correlation structure. The preference for aesthetic is still present in both GPT-4o and Claude-3.5-Sonnet (Claude) as exhibited by Fig. 6 and Fig. 7, but Claude shows a high correlation between object adherence and the overall rating, in contrast to humans, who show a higher overall correlation for style or artifacts. On

Model	Aesthetic	Composition	Style	Artifacts	Anatomy	Objects	Overall
gpt4o	**0.127**	**0.147**	**0.110**	**0.117**	**0.152**	**0.229**	**0.176**
pickscore	-	-	-	-	-	-	0.164
claude	0.119	0.114	0.095	0.104	0.147	0.207	0.140
hpsv2	-	-	-	-	-	-	0.133
imagereward	-	-	-	-	-	-	0.128
vqascore	-	-	-	-	-	-	0.110
gemini-1.5-flash	0.051	0.047	0.101	0.062	0.075	0.138	0.071
gpt4o-mini	0.042	0.091	0.055	0.056	0.046	0.170	0.063
pixtral	0.030	0.023	0.028	0.013	0.017	0.041	0.045

Table 1: Cohen's Kappa agreement between human and LLM ratings across different aspects (sorted by Overall).

Model	Aesthetic	Composition	Style	Artifacts	Anatomy	Objects	Overall
pickscore	-	-	-	-	-	-	**0.498**
gpt4o	**0.376**	**0.467**	**0.337**	0.349	**0.398**	**0.501**	0.461
claude	0.338	0.397	0.336	**0.361**	0.390	0.467	0.407
hpsv2	-	-	-	-	-	-	0.398
imagereward	-	-	-	-	-	-	0.398
vqascore	-	-	-	-	-	-	0.356
gemini-1.5-flash	0.260	0.339	0.282	0.307	0.337	0.360	0.362
gpt4o-mini	0.211	0.259	0.152	0.175	0.167	0.350	0.286
pixtral	0.052	0.119	0.080	0.053	0.074	0.154	0.140

Table 2: Pearson correlation between human and LLM ratings across different aspects (sorted by Overall).

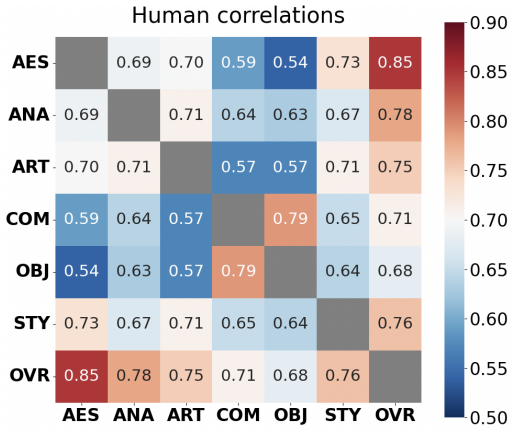

Figure 5: Correlation coefficients for human ratings, demonstrating various strong linkages between pairs of tasks.

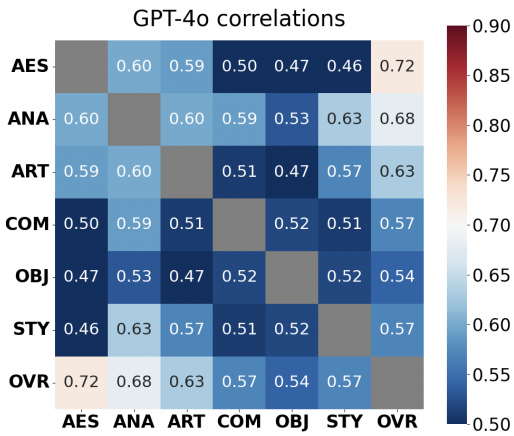

Figure 6: Correlation coefficients for GPT-4o ratings, demonstrating a lack of strong link between tasks.

top of that, GPT-4o seems to have extremely weak correlations between each pair of non-overall tasks, which is in stark contrast to the human case. These differences highlight that while humans rate images based on complex image quality attribute relationships, LLMs have not yet achieved similar ways of reasoning about these attributes.

5 Synthetic Attribute Datasets

As we have seen in Section 4.3, LLMs struggle to understand the relationship humans display between image quality attributes and overall image quality. This raises questions about the degree to which LLMs understand each individual image quality attribute in isolation. To explore this, we construct subtasks for four of the image attributes previously discussed: aesthetic, composition, style, and anatomy. For each task, we build large, but simple synthetic training and evaluation datasets with high quality labels. Each of these subtasks is intended to be a task where humans can easily perform well (e.g. identifying correct and incorrect human geometry) but LLMs may struggle. We describe each of the datasets below and provide examples of sample data in Fig. 8. Further examples are presented in Appendix Section 4.

5.1 Synthetic Aesthetic Dataset

Purpose: To study whether LLMs can easily model the judgement of aesthetic quality, using the AesExpert model

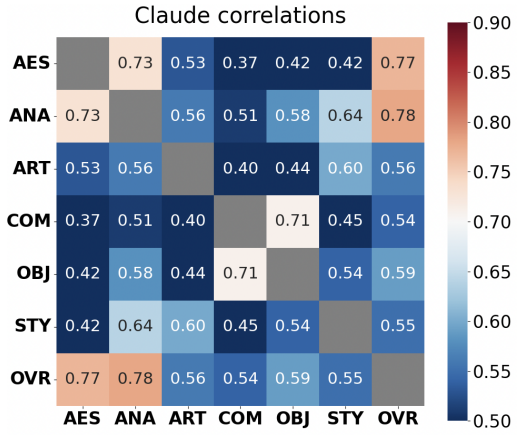

Figure 7: Correlation coefficients for Claude ratings, demonstrating a strong link between aesthetic and anatomy, and compositionality and object adherence.

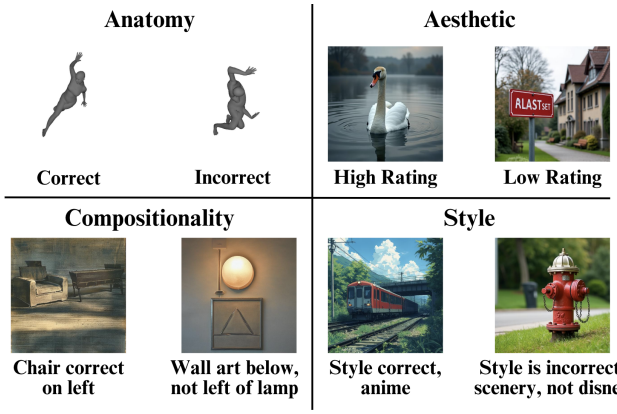

Figure 8: Sample correct and incorrect images from each of the synthetic datasets.

as a proxy (Huang et al. 2024).
Data Generation: We sample 50K prompts from COCO prompts and generate images using FLUX-1.0[dev] (Lin et al. 2014; Black Forest Labs 2024). Each of these images is rated by the AesExpert model (Huang et al. 2024).
Prompt Format: We provide the model with the image and original image prompt, and ask the model to output a floating point rating between 1 and 10.

5.2 Synthetic Anatomy Dataset

Purpose: To study whether LLMs can detect simple distortions in human anatomy.
Data Generation: We use the parameterized articulated SKEL 3D model of a human body (Keller et al. 2023). By varying joint parameters, 50k samples are generated in a 50/50 split where half of the samples match normal human joint parameters, while half of the samples do not. These samples are rendered using OpenGL with a fixed camera.
Prompt Format: This task provides the model with an image of a generated human body, and asks the model whether the human is anatomically correct or distorted.

5.3 Synthetic Composition Dataset

Purpose: To study whether LLMs can identify the four cardinal directions, a fundamental compositional task.
Data Generation: We sample 50k image pairs of household objects taken from the Amazon Object (ABO) dataset (Collins et al. 2022). These pairs are converted into Canny edge maps, which are combined based on the cardinal direction of choice. We convert the combined Canny edges into images using a FLUX-1.0[dev] ControlNet (Black Forest Labs 2024; Zhang, Rao, and Agrawala 2023). The image is screened using Florence-2 to evaluate whether the original objects taken from the ABO dataset match those generated in the image, ensuring output quality (Xiao et al. 2024).
Prompt Format: This task provides the model with the image and a prompt which may or may not have the correct direction, and asks the model if the direction is properly displayed in the image.

5.4 Synthetic Style Dataset

Purpose: With this dataset, we evaluate whether LLMs can distinguish six styles that are distinguishable for humans.
Data Generation: We sample 10k images each from these six styles: anime, art, disney, midjourney, realism, and scenery. Images are generated using prompts from COCO prompts and a LoRA version of FLUX-1.0[dev] (Lin et al. 2014; Black Forest Labs 2024; Hu et al. 2022).
Prompt Format: This task provides the model with a prompt that may or may not have the correct style, asking the model whether or not the style direction is properly displayed in the image.

5.5 Validating Synthetic Data Quality

We perform an independent human evaluation on a random subset of the datasets to validate the quality of our synthetic data. For each of the datasets, we align based on the metric used to evaluate the models, which is accuracy for the Yes/No tasks of anatomy, compositionality, and style, and Pearson correlation for the aesthetic task. For each of the tasks, as presented in Table 3, the human evaluator does significantly better than random guessing, indicating that the synthetic data has a strong signal for each the task.

6 Shared Training and Evaluation Details

For training, we use LLaVA (LLaMA-3-8B LM and 384x384 shape-optimized SigLIP VLM) trained with a 1e-5 learning rate and Adam (Liu et al. 2023b; Kingma and Ba 2015; Zhai et al. 2023). Models were trained on A100-80GB for around 160 A100-hours each on Transformers 4.44.

To teach LLaVA multi-image understanding, we use a custom data mix, mixing the multi-image Commonalities and Differences (CaD) dataset with baseline LLaVA data (Lin et al. 2024a). We use this data in a third phase of LLaVA training, where the LLM and vision projector are both unfrozen. On top of CaD, we also mix in the training split of our synthetically generated datasets as a small proportion of the training dataset. All models are trained for

Model	Anatomy Accuracy	Compo Accuracy	Style Accuracy	Aesthetic Correlation
Human Evaluator	**78.2%**	**99%**	**79%**	**0.671**
Best Finetuned Model	**57.6%**	**91.2%**	**81.8%**	**0.847**
claude-3-5-sonnet-20241022	52.0%	71.0%	60.4%	0.257
gpt-4o	51.6%	75.2%	61.8%	0.583
gpt-4o-mini	47.7%	58.6%	61.0%	0.586
gemini-1.5-flash-002	47.8%	57.4%	62.7%	0.037
pixtral-12b	51.2%	46.6%	55.7%	-0.053
baseline LLaVA	50.0%	52.0%	57.2%	-0.010

Table 3: Merged accuracy results across all datasets, demonstrating that anatomy is the most challenging, while compositionality can be easily learned, despite a somewhat lack of capability in large language models.

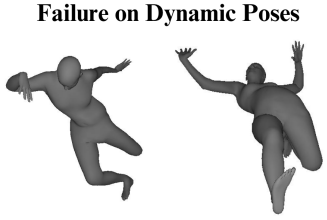

Figure 9: GPT-4o and other LLMs label dynamic poses as anatomically incorrect, an intriguing failure mode.

one epoch, and we sweep across 200k, 400k, and 800k as the number of multi-image samples seen.

We find that including the CaD and LLaVA data creates a push-and-pull effect, where CaD data helps the model acquire multi-image capabilities and LLaVA data prevents the model from losing baseline VQA capabilities. Without LLaVA data, the model catastrophically overfits on either CaD data or the synthetic data provided, even when the synthetic data is at most 10% of the data mix. In-depth ablations are provided in Appendix Section 6.

7 Results on Individual Tasks

7.1 Individual Model Performance

On these tasks, we evaluate the five large LLMs evaluated in the previous section (GPT-4o, GPT-4o-mini, Claude-3.5-Sonnet, Gemini-1.5-Flash, and Pixtral-12B). As in Table 3 we find that these models perform better than baseline LLaVA, which performs near random on these simple tasks, but do not reach the performance of a human evaluator.

Pixtral and baseline LLaVA lag behind most, with performances mostly near random guessing, compared to the strong performance of frontier scale models, indicating that there may be a fundamental change in the performance of multimodal LLMs as image evaluators at some baseline data and parameter scale. However, models specifically finetuned for each task vastly outperform the frontier models, coming near human accuracy on both compositionality and style, and better matching the aesthetic of AesExpert than humans (Huang et al. 2024). We find that finetuning on a mix of tasks does not increase performance, with ablations provided in Appendix Section 6.

Failure Modes on Anatomy Even with finetuning, model performance lags behind on anatomy, despite human evaluators performing well. No LLM achieves more than random guessing, which is primarily caused by the failure mode in Fig. 9. We observe that in anatomically valid dynamic poses, like the two images provided (kicking a soccer ball and running), LLMs mark these poses as anatomically impossible, leading to poor correlation with human raters.

Mechanistically, multimodal LLM (MLLM) performance can be strongly attributed to the alignment of vision-language encoders like CLIP, which are trained on captioning tasks. MLLMs are thus biased to captioning-related tasks like object adherence, and not spatial reasoning. Table 3 shows this behavior, as while anatomy involves the same spatial reasoning as composition (angle and orientation), LLMs consistently underperform, implying inductive biases in the training procedure.

To confirm this, we conduct a further experiment, where MLLMs are prompted to output joint angles before outputting an anatomical decision. Outputted angles are reasonable, but final accuracy remains low, with Gemini and GPT at 46% and Claude at 54%. This indicates that while models can reason about angles, they lack inductive bias to correctly process these angles into some more complex result.

8 Conclusion

We discover that multimodal LLMs show significant differences from humans in how they rate generated images. Surprisingly, this difference is hidden by their strong performance at providing overall ratings, only becoming apparent on individual image quality attributes. Specifically, through our curated multi-task image rating dataset, we find that although large LLMs and smaller reward models are aligned with human preferences, LLMs do not learn the same similarities between pairs of image quality attributes that standard human raters do. Even on simplified versions of image rating tasks, LLMs do not generalize in the same fashion as humans, performing well on tasks like evaluating compositional correctness, but falling far behind in tasks like detecting incorrect anatomy. These findings unveil potential gaps in the overall alignment of multimodal LLMs as image rating judges. The more we rely on automated evaluation for generative image models, the more the need to fill these gaps and ensure that our automated judges are actually aligned with human preferences on an ensemble of tasks.

References

Agrawal, P.; Antoniak, S.; Hanna, E. B.; Bout, B.; Chaplot, D. S.; Chudnovsky, J.; Costa, D.; Monicault, B. D.; Garg, S.; Gervet, T.; Ghosh, S.; Héliou, A.; Jacob, P.; Jiang, A. Q.; Khandelwal, K.; Lacroix, T.; Lample, G.; de Las Casas, D.; Lavril, T.; Scao, T. L.; Lo, A.; Marshall, W.; Martin, L.; Mensch, A.; Muddireddy, P.; Nemychnikova, V.; Pellat, M.; von Platen, P.; Raghuraman, N.; Rozière, B.; Sablayrolles, A.; Saulnier, L.; Sauvestre, R.; Shang, W.; Soletskyi, R.; Stewart, L.; Stock, P.; Studnia, J.; Subramanian, S.; Vaze, S.; Wang, T.; and Yang, S. 2024. Pixtral 12B. *CoRR*, abs/2410.07073.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities. *CoRR*, abs/2308.12966.

Black, K.; Janner, M.; Du, Y.; Kostrikov, I.; and Levine, S. 2024. Training Diffusion Models with Reinforcement Learning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Black Forest Labs. 2024. FLUX.1-dev: A Vision-Language Model by Black Forest Labs. Hugging Face Model Hub. Accessed: 2025-03-06.

Cao, B.; Yuan, J.; Liu, Y.; Li, J.; Sun, S.; Liu, J.; and Zhao, B. 2024. SynArtifact: Classifying and Alleviating Artifacts in Synthetic Images via Vision-Language Model. *CoRR*, abs/2402.18068.

Chen, D.; Chen, R.; Zhang, S.; Wang, Y.; Liu, Y.; Zhou, H.; Zhang, Q.; Wan, Y.; Zhou, P.; and Sun, L. 2024a. MLLM-as-a-Judge: Assessing Multimodal LLM-as-a-Judge with Vision-Language Benchmark. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Chen, Z.; Du, Y.; Wen, Z.; Zhou, Y.; Cui, C.; Weng, Z.; Tu, H.; Wang, C.; Tong, Z.; Huang, Q.; Chen, C.; Ye, Q.; Zhu, Z.; Zhang, Y.; Zhou, J.; Zhao, Z.; Rafailov, R.; Finn, C.; and Yao, H. 2024b. MJ-Bench: Is Your Multimodal Reward Model Really a Good Judge for Text-to-Image Generation? *CoRR*, abs/2407.04842.

Collins, J.; Goel, S.; Deng, K.; Luthra, A.; Xu, L.; Gundogdu, E.; Zhang, X.; Yago Vicente, T. F.; Dideriksen, T.; Arora, H.; Guillaumin, M.; and Malik, J. 2022. ABO: Dataset and Benchmarks for Real-World 3D Object Understanding. *CVPR*.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. C. H. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; Goyal, A.; Hartshorn, A.; Yang, A.; Mitra, A.; Sravankumar, A.; Korenev, A.; Hinsvark, A.; Rao, A.; Zhang, A.; Rodriguez, A.; Gregerson, A.; Spataru, A.; Rozière, B.; Biron, B.; Tang, B.; Chern, B.; Caucheteux, C.; Nayak, C.; Bi, C.; Marra, C.; McConnell, C.; Keller, C.; Touret, C.; Wu, C.; Wong, C.; Ferrer, C. C.; Nikolaidis, C.; Allonsius, D.; Song, D.; Pintz, D.; Livshits, D.; Esiobu, D.; Choudhary, D.; Mahajan, D.; Garcia-Olano, D.; Perino, D.; Hupkes, D.; Lakomkin, E.; AlBadawy, E.; Lobanova, E.; Dinan, E.; Smith, E. M.; Radenovic, F.; Zhang, F.; Synnaeve, G.; Lee, G.; Anderson, G. L.; Nail, G.; Mialon, G.; Pang, G.; Cucurell, G.; Nguyen, H.; Korevaar, H.; Xu, H.; Touvron, H.; Zarov, I.; Ibarra, I. A.; Kloumann, I. M.; Misra, I.; Evtimov, I.; Copet, J.; Lee, J.; Geffert, J.; Vranes, J.; Park, J.; Mahadeokar, J.; Shah, J.; van der Linde, J.; Billock, J.; Hong, J.; Lee, J.; Fu, J.; Chi, J.; Huang, J.; Liu, J.; Wang, J.; Yu, J.; Bitton, J.; Spisak, J.; Park, J.; Rocca, J.; Johnstun, J.; Saxe, J.; Jia, J.; Alwala, K. V.; Upasani, K.; Plawiak, K.; Li, K.; Heafield, K.; Stone, K.; and et al. 2024. The Llama 3 Herd of Models. *CoRR*, abs/2407.21783.

Dzabraev, M.; Kunitsyn, A.; and Ivaniuta, A. 2024. VLRM: Vision-Language Models act as Reward Models for Image Captioning. *CoRR*, abs/2404.01911.

fal. 2025. Image Systems Arena. https://imgsys.org/. An open source platform for evaluating text-guided image generation models. Sister project to Chatbot Arena by lmsys.org and Image Arena by artificialanalysis.ai. Collected preference data released under CC BY-SA 4.0 DEED license. Accessed: 2025-03-06.

Google. 2024. Our next-generation model: Gemini 1.5. https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/. Accessed: 2025-02-15.

Han, S.; Fan, H.; Fu, J.; Li, L.; Li, T.; Cui, J.; Wang, Y.; Tai, Y.; Sun, J.; Guo, C.; and Li, C. 2024. EvalMuse-40K: A Reliable and Fine-Grained Benchmark with Comprehensive Human Annotations for Text-to-Image Generation Model Evaluation. arXiv:2412.18150.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Huang, K.; Sun, K.; Xie, E.; Li, Z.; and Liu, X. 2023. T2I-CompBench: A Comprehensive Benchmark for Open-world Compositional Text-to-image Generation. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Huang, Y.; Sheng, X.; Yang, Z.; Yuan, Q.; Duan, Z.; Chen, P.; Li, L.; Lin, W.; and Shi, G. 2024. AesExpert: Towards Multi-modality Foundation Model for Image Aesthetics Perception. In Cai, J.; Kankanhalli, M. S.; Prabhakaran, B.;

Boll, S.; Subramanian, R.; Zheng, L.; Singh, V. K.; César, P.; Xie, L.; and Xu, D., eds., *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, 5911–5920. ACM.

Jiao, Q.; Chen, D.; Huang, Y.; Li, Y.; and Shen, Y. 2024. Img-Diff: Contrastive Data Synthesis for Multimodal Large Language Models. *CoRR*, abs/2408.04594.

Keller, M.; Werling, K.; Shin, S.; Delp, S.; Pujades, S.; Liu, C. K.; and Black, M. J. 2023. From Skin to Skeleton: Towards Biomechanically Accurate 3D Digital Humans. *ACM Transaction on Graphics (ToG)*, 42(6): 253:1–253:15.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 36652–36663.

Lambert, N.; Pyatkin, V.; Morrison, J.; Miranda, L.; Lin, B. Y.; Chandu, K. R.; Dziri, N.; Kumar, S.; Zick, T.; Choi, Y.; Smith, N. A.; and Hajishirzi, H. 2024. RewardBench: Evaluating Reward Models for Language Modeling. *CoRR*, abs/2403.13787.

Lee, S.; Kim, S.; Park, S. H.; Kim, G.; and Seo, M. 2024. Prometheus-Vision: Vision-Language Model as a Judge for Fine-Grained Evaluation. In Ku, L.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, 11286–11315. Association for Computational Linguistics.

Lee, T.; Yasunaga, M.; Meng, C.; Mai, Y.; Park, J. S.; Gupta, A.; Zhang, Y.; Narayanan, D.; Teufel, H.; Bellagente, M.; Kang, M.; Park, T.; Leskovec, J.; Zhu, J.; Li, F.; Wu, J.; Ermon, S.; and Liang, P. 2023. Holistic Evaluation of Text-to-Image Models. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In Fleet, D. J.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, 740–755. Springer.

Lin, W.; Mirza, M. J.; Doveh, S.; Feris, R.; Giryes, R.; Hochreiter, S.; and Karlinsky, L. 2024a. Comparison Visual Instruction Tuning. *arXiv preprint*.

Lin, Z.; Pathak, D.; Li, B.; Li, J.; Xia, X.; Neubig, G.; Zhang, P.; and Ramanan, D. 2024b. Evaluating Text-to-Visual Generation with Image-to-Text Generation. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part IX*, volume 15067 of *Lecture Notes in Computer Science*, 366–384. Springer.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023a. Improved baselines with visual instruction tuning, 2024. *URL https://arxiv.org/abs/2310.03744*, 3(4): 5.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Nassar, J.; Pavon-Harr, V.; Bosch, M.; and McCulloh, I. 2019. Assessing Data Quality of Annotations with Krippendorff Alpha For Applications in Computer Vision. *CoRR*, abs/1912.10107.

OpenAI. 2024a. GPT-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Accessed: 2025-02-15.

OpenAI. 2024b. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2025-02-15.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P. F.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-resolution image synthesis with latent diffusion models, 2021.

Son, G.; Ko, H.; Lee, H.; Kim, Y.; and Hong, S. 2024. LLM-as-a-Judge & Reward Model: What They Can and Cannot Do. *CoRR*, abs/2409.11239.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D. M.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and

Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Team, C. 2024. Chameleon: Mixed-Modal Early-Fusion Foundation Models. *CoRR*, abs/2405.09818.

Wallace, B.; Dang, M.; Rafailov, R.; Zhou, L.; Lou, A.; Purushwalkam, S.; Ermon, S.; Xiong, C.; Joty, S.; and Naik, N. 2024. Diffusion Model Alignment Using Direct Preference Optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, 8228–8238. IEEE.

Wang, B.; Zheng, R.; Chen, L.; Liu, Y.; Dou, S.; Huang, C.; Shen, W.; Jin, S.; Zhou, E.; Shi, C.; Gao, S.; Xu, N.; Zhou, Y.; Fan, X.; Xi, Z.; Zhao, J.; Wang, X.; Ji, T.; Yan, H.; Shen, L.; Chen, Z.; Gui, T.; Zhang, Q.; Qiu, X.; Huang, X.; Wu, Z.; and Jiang, Y. 2024a. Secrets of RLHF in Large Language Models Part II: Reward Modeling. *CoRR*, abs/2401.06080.

Wang, M.; Mao, J.; Wang, X.; and Yamasaki, T. 2024b. Reward Incremental Learning in Text-to-Image Generation. *CoRR*, abs/2411.17310.

Wang, Z.; Hu, S.; Zhao, S.; Lin, X.; Juefei-Xu, F.; Li, Z.; Han, L.; Subramanyam, H.; Chen, L.; Chen, J.; Jiang, N.; Lyu, L.; Ma, S.; Metaxas, D. N.; and Jain, A. 2025. MLLM-as-a-Judge for Image Safety without Human Labeling. *CoRR*, abs/2501.00192.

Wei, H.; He, S.; Xia, T.; Wong, A.; Lin, J.; and Han, M. 2024. Systematic Evaluation of LLM-as-a-Judge in LLM Alignment Tasks: Explainable Metrics and Diverse Prompt Templates. *CoRR*, abs/2408.13006.

Wu, X.; Hao, Y.; Sun, K.; Chen, Y.; Zhu, F.; Zhao, R.; and Li, H. 2023. Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis. *arXiv preprint arXiv:2306.09341*.

Wu, X.; Huang, S.; and Wei, F. 2024. Multimodal Large Language Model is a Human-Aligned Annotator for Text-to-Image Generation. *CoRR*, abs/2404.15100.

Xiao, B.; Wu, H.; Xu, W.; Dai, X.; Hu, H.; Lu, Y.; Zeng, M.; Liu, C.; and Yuan, L. 2024. Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, 4818–4829. IEEE.

Xu, J.; Huang, Y.; Cheng, J.; Yang, Y.; Xu, J.; Wang, Y.; Duan, W.; Yang, S.; Jin, Q.; Li, S.; Teng, J.; Yang, Z.; Zheng, W.; Liu, X.; Ding, M.; Zhang, X.; Gu, X.; Huang, S.; Huang, M.; Tang, J.; and Dong, Y. 2024. VisionReward: Fine-Grained Multi-Dimensional Human Preference Learning for Image and Video Generation. *CoRR*, abs/2412.21059.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 15903–15935.

Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; Hutchinson, B.; Han, W.; Parekh, Z.; Li, X.; Zhang, H.; Baldridge, J.; and Wu, Y. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *Trans. Mach. Learn. Res.*, 2022.

Yuan, W.; Pang, R. Y.; Cho, K.; Li, X.; Sukhbaatar, S.; Xu, J.; and Weston, J. 2024. Self-Rewarding Language Models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid Loss for Language Image Pre-Training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 11941–11952. IEEE.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 3813–3824. IEEE.

Zhang, Z.; Kou, T.; Wang, S.; Li, C.; Sun, W.; Wang, W.; Li, X.; Wang, Z.; Cao, X.; Min, X.; Liu, X.; and Zhai, G. 2025. Q-Eval-100K: Evaluating Visual Quality and Alignment Level for Text-to-Vision Content. arXiv:2503.02357.

Zheng, L.; Chiang, W.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Zhou, Z.; Wang, Q.; Lin, B.; Su, Y.; Chen, R.; Tao, X.; Zheng, A.; Yuan, L.; Wan, P.; and Zhang, D. 2024. UNIAA: A Unified Multi-modal Image Aesthetic Assessment Baseline and Benchmark. *CoRR*, abs/2404.09619.