Cycle Consistency as Reward: Learning Image-Text Alignment without Human Preferences

Hyojin Bahng* Caroline Chan* Frédo Durand Phillip Isola
MIT CSAIL
{bahng, cmchan, fredo, phillipi}@mit.edu

Abstract

Learning alignment between language and vision is a fundamental challenge especially as multimodal data becomes increasingly detailed and complex. Existing methods often rely on collecting human or AI feedback, which can be costly and time-intensive. We propose an alternative approach that leverages cycle consistency as a supervisory signal. Given an image and generated text, we map the text back to image space with a text-to-image model and compute the similarity between the original image and its reconstruction. Analogously, for text-to-image generation, we measure textual similarity between an input caption and its reconstruction through the cycle. We use the cycle consistency score to rank candidates and construct a preference dataset of 866K comparison pairs. The reward model trained on our dataset outperforms state-of-the-art alignment metrics on detailed captioning, with superior inference-time scalability when used as a verifier for best-of-N sampling. Furthermore, performing DPO [70] and Diffusion-DPO [86] using our dataset enhances performance across a wide range of vision-language tasks and text-to-image generation respectively. Our dataset, model, and code is publicly released here https://carolineec.github.io/cyclereward. **TODO: change link when it is live**

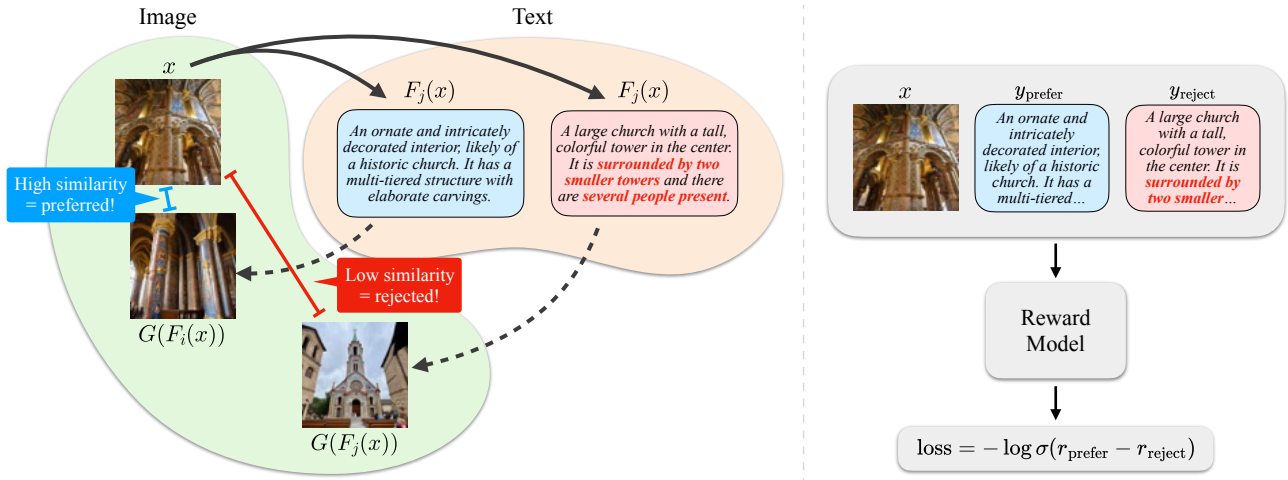
1. Introduction

Aligning language and vision remains a central challenge in multimodal learning. As input data becomes increasingly complex and detailed, generative models often produce *misaligned* outputs — vision-language models often hallucinate descriptions not grounded in the image [50], while diffusion models can generate images that misrepresent quantities, attributes, or object relationships in text prompts [33, 40]. Existing alignment metrics primarily focus on short captions or prompts [40, 90, 92], limiting their effectiveness for evaluating dense alignment. Common strategies for improving alignment include reinforcement learning from human feedback (RLHF) [65] or direct preference optimization (DPO) [70], which requires collecting high-quality human preferences [40, 89, 90, 92] or AI feedback [48] from proprietary models (e.g., GPT-4V [62]), which is prohibitively expensive.

Comparing images and text is inherently challenging, even more so with detailed text. However, the comparison becomes much easier when we map text back into image space (Figure 2). As the generated text becomes more descriptive and accurate, the reconstructed image better resembles the original image. This idea of cycle consistency [39, 81, 107] has been used as a metric to evaluate image-to-text generation [23, 32] and to optmize diffusion models [4]. However, these approaches compute cycle consistency on-the-fly using large pre-trained models, which is prohibitively slow and often not differentiable.

In this paper, we introduce a reward model trained on preferences derived from cycle consistency. Given an image-to-text mapping $F : X \rightarrow Y$ and a backward text-to-image mapping $G : Y \rightarrow X$, we define *cycle consistency score* as the similarity between the original input x and its reconstruction $G(F(x))$. We use the cycle consistency score as a proxy for preferences, where a higher score indicates a preferred output. This provides a more scalable and cheaper signal for learning image-text alignment compared to human supervision. We construct CyclePrefDB, a large-scale preference dataset comprising 866K comparison pairs from 11 image-to-text models and 4 text-to-image models. CyclePrefDB provides significantly denser text than typical text-to-image datasets (Table 1), while staying within the 77-token limit of the text-to-image model to compute cycle consistency. We train a reward model on this dataset, called CycleReward, which proves effective both as an alignment metric and for best-of-N optimization. It achieves state-of-the-art performance for detailed captioning and competitive performance in text-to-image synthesis. Finally, applying direct preference optimization (DPO) [70, 86] using our dataset enhances a wide range of vision-language and

*Equal contribution.



(a) Cycle Consistency Preference Collection

(b) Reward Model Training

Figure 1. **Method overview.** (a) Given an input image x, we generate multiple candidate captions $F_i(x)$, $F_j(x)$ using different captioning models. Each caption is mapped back to the image domain via a text-to-image model G, and compared against the original image. Captions whose reconstructions $G(F(x))$ are more similar to the original image are preferred; those with low similarity are rejected. (b) These comparison pairs are used to train a reward model, which learns to assign higher scores to preferred captions. We apply the same process for text-to-image generation.

text-to-image generation tasks without requiring any human supervision.

In summary, we make the following contributions:

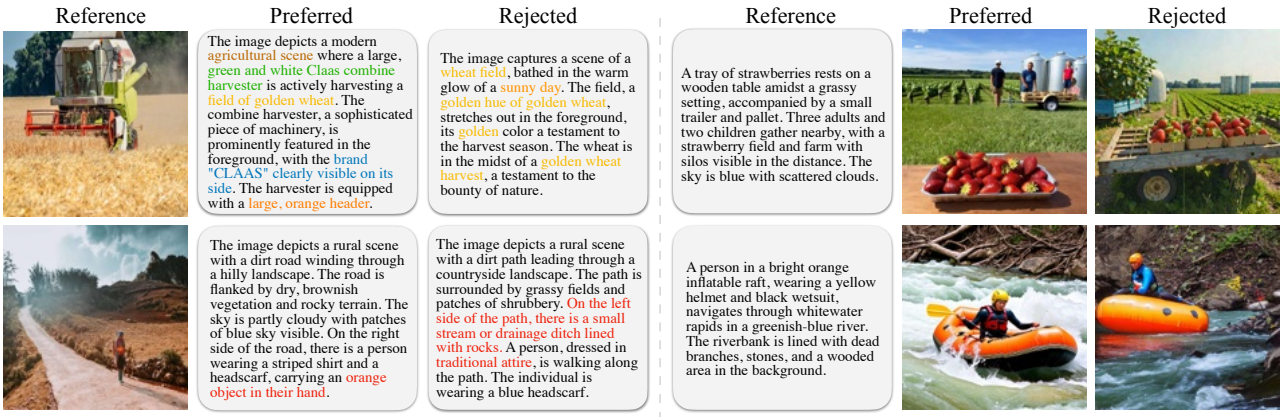
- **CyclePrefDB**, a cycle-consistent preference dataset of 866K comparison pairs spanning image-to-text and text-to-image generation, specifically designed for *dense* captions and prompts.
- **CycleReward**, a reward model trained on our dataset, which is effective both as an alignment metric and a verifier for best-of-N optimization.
- **Hyojin: An ablation study on reward model design choices, which demonstrates the effect of decoders, similarity metrics, objective functions, and dataset scale. Phillip: Don't use the word "extensive". It comes across as weak. Also here it would be better to say what we found rather than saying what you ran.**
- Demonstration of DPO using our CyclePrefDB dataset. This leads to improvements on a wide range of vision-language tasks and text-to-image generation tasks.

2. Related Work

Image-text alignment. Image-text alignment metrics can be classified in two ways - as reference-based, which require comparison with ground truth text, or as reference-free, which compute alignment based solely on the provided image and text. Reference-based metrics include BLEU [66], CIDEr [85], and METEOR [41] which measure linguistic similarity between candidate and reference captions. These methods do not generalize well to texts which vary in style and syntax from the reference caption. Recent approaches such as SPICE [1], CAPTURE [16], and DCScore [93] decompose the candidate text into scene graphs or basic information units which are then compared to ground truth labels. Although recent reference-based metrics are more flexible and thorough, they are limited by lack of differentiability, slow runtime, and, most importantly, they require a reference, which means they are not suitable as an objective function.

Reference-free metrics come in a variety of forms. Many approaches adapt pre-trained CLIP [69] image and text encodings [18, 29, 37, 99] while others collect human preferences to train reward models [40, 89, 90, 92]. Some recent methods query a large pre-trained model to directly evaluate alignment [7, 42, 52, 91]. Although current metrics increasingly align with human preferences for visio-linguistic reasoning and text-to-image evaluation, many of these methods fail to evaluate longer, more descriptive captions effectively. Most similar to our method, Image2Text2Image [32] computes image captioning performance by leveraging text-to-image generation to produce reconstructed images given text captions. The final score is the reconstruction error between the original and generated image's DINOv2 [64] or CLIP [69] features. DDPO [4] also includes a similar text-to-image-to text reconstruction score to optimize diffusion models. These pipelines match our dataset collection process outlined in Section 3. However, our method uses cycle consistency scores to train a reward model with the benefit of inference speed, differentiability for downstream applications, and better performance.

Detailed captioning evaluation. Image-to-text models



(a) Image-to-text Generation (b) Text-to-Image Generation

Figure 2. **What do cycle consistency preferences look like?** We visualize examples from our dataset, where cycle consistency determines preferences over comparison pairs. **(a) Image-to-text**: Preferred and rejected texts vary in levels of density, comprehensiveness, and hallucination. For instance, the preferred text (upper example) provides a fine-grained, dense description of the image, whereas the rejected text only describes the golden wheat field. In the lower example, increased hallucination lowers preference. **(b) Text-to-image**: Images that capture fine-grained details of the prompt produce better text reconstructions, resulting in higher preference. Note that we use dense captions from the DCI dataset [84].

can produce comprehensive descriptions [24, 61] by scaling the language model [53, 54] and training on semantically rich synthetic captions [46, 47, 53, 54, 78]. Despite growing model capabilities and attention to detailed descriptions, few methods and datasets exist for evaluating descriptive captions. Two recently released benchmarks, DetailCaps-4870 [16] and DeCapBench [93], address this issue. DetailCaps-4870 [16] evaluates image-text alignment metrics on their ability to measure alignment for detailed descriptions, whereas DeCapBench evaluates image-to-text model results on detail captioning with their proposed reference-based metric DCSCORE. Our reward model provides a fast, differentiable, and reference-free approach to evaluating alignment for detailed captions.

Cycle consistency. Imposing cycle consistency continuously has been shown to be effective for many tasks in different domains [6, 25, 28, 34, 36, 87, 95, 100, 104, 105], especially for self-supervised training and cases which lack paired ground truth annotations [25, 30, 36, 49, 59, 95, 100, 105, 107], and recently for evaluating VLM and LLM performance [15, 76]. Rapid progression of image-to-text and text-to-models has facilitated exploring cycle consistency between images and texts [26], and models have begun to incorporate cycle consistency at training time by combining text-to-image diffusion models and vision-language models (VLMs) [3, 20, 49, 78].

Preference optimization. There are several techniques to align model outputs with human preferences [65, 79] during training [70, 75, 77] or at test time [38, 60, 79]. These approaches have been applied mostly to large language models and recently in vision-language models [80, 97, 98] and diffusion models [4, 68, 86]. Text-to-image evaluation metrics such as Human Preference Score (HPS) [89, 90], PickScore [40], and ImageReward [92] collect human preferences which are used to train a reward model. Other methods such as VLFeedback [48] substitute human feedback by using foundation models (e.g., GPT-4V) to annotate preferences [48, 93, 98, 103] and perform Direct Preference Optimization (DPO) [70] on their dataset. Our method collects preferences from a new signal: cycle consistency, which is cheaper and more easily scalable. We apply our dataset both to reward modeling and preference learning via DPO exhibiting competitive performance with human labels.

3. Method

3.1. Cycle Consistency as Preferences

Our goal is to learn preferences for image-text alignment without relying on human annotations. Prior approaches often use human [40, 90, 92] or GPT-4V [48] preferences to rank the quality of generated captions or images. Instead, we propose to derive preferences from *cycle consistency*. Given an image-to-text mapping $F : X \to Y$, we measure how well a text $F(x)$ aligns with an image x by measuring how well it can reconstruct x through a backward mapping $G : Y \to X$. We define *cycle consistency score* for $F(x)$ conditioned on x as:

$$s(x \to F(x)) = d_{\text{img}}(x, G(F(x))), \tag{1}$$

where d_{img} measures the similarity between the reconstructed image $G(F(x))$ and the original image x. We use DreamSim [22] to compute this similarity.

Dataset	Task	# Pairs	Supervision	Tokens
ImageRewardDB [92]	T2I	137K	Human	35.73
HPDv2 [89]	T2I	798K	Human	18.89
Pick-A-Pic v2 [40]	T2I	851K	Human	23.74
VLFeedback [48]	VL	399K	GPT-4V [62]	97.03
CyclePrefDB-I2T	I2T	398K	Cycle consistency	56.82
CyclePrefDB-T2I	T2I	468K	Cycle consistency	55.13

Table 1. **Key differences of preference datasets.** Existing preference datasets often use human or GPT-4V annotations for supervision, whereas we leverage cycle consistency for preference annotation. We provide comparison pairs for both image-to-text (I2T) and text-to-image (T2I) generation tasks. CyclePrefDB features significantly denser text than typical T2I datasets, while remaining within token limits (77 tokens) of text-to-image models. VL denotes vision-language tasks.

Similarly, for a text-to-image mapping $G : Y \to X$, we measure how well an image $G(y)$ aligns with text y by using a backward mapping $F : X \to Y$. We define the cycle consistency score for $G(y)$ conditioned on y as:

$$s(y \to G(y)) = d_{\text{text}}(y, F(G(y))), \tag{2}$$

where d_{text} measures the similarity between the reconstructed text $F(G(y))$ and the original text y. We use SBERT [71] to compute this similarity.

Importantly, these scores generalize to arbitrary image–text pairs (x, y), not just model outputs:

$$\begin{aligned} s(x \to y) &= d_{\text{img}}(x, G(y)), \\ s(y \to x) &= d_{\text{text}}(y, F(x)). \end{aligned} \tag{3}$$

While prior work [23, 32] uses this score directly as an alignment metric, we *learn* preferences from a large pool of comparisons. Given triplets (x, y_i, y_j) and (y, x_i, x_j), we convert the cycle consistency scores into pairwise preferences:

$$\begin{aligned} y_i \succ y_j \;\text{ if }\; s(x \to y_i) > s(x \to y_j), \\ x_i \succ x_j \;\text{ if }\; s(y \to x_i) > s(y \to x_j). \end{aligned} \tag{4}$$

where $\succ$ denotes that y_i is preferred over y_j, vice versa. We establish the connection between cycle consistency score and cycle consistency of mappings in Appendix A.

3.2. Dataset Generation

Image-to-text generation. Given image x, we first obtain multiple candidate text descriptions $\{y_1, ..., y_n\}$ of varying quality. In practice, we use 11 image-to-text models trained on different datasets and scales: BLIP2 (T5-XXL) [47], LLaVA-1.5 (7B, 13B) [55], LLaVA-1.6 (7B, 34B) [54], LLaVA-OneVision (0.5B, 7B) [44], and InternVL2 (2B, 8B, 26B, 40B) [9, 63]. As reward modeling is inherently contrastive, we deliberately include older models that produce short, hallucinated captions as negative examples alongside newer models to maximize text diversity. We specifically instruct the models to generate *rich, descriptive* captions, using the prompt recommended by the model distributor (Appendix C). We use greedy sampling with a maximum token length of 77, i.e., maximum prompt length supported by the text-to-image models. We fix the backward mapping F as LLaVA-1.5-13B to compute $s(x \to y)$.

Text-to-image generation. Given a text prompt y, we generate a set of image candidates $\{x_1, ..., x_n\}$, similarly using 4 text-to-image models: Stable Diffusion 1.5 [72], Stable Diffusion XL [67], Stable Diffusion 3 [20], and FLUX (Timestep-distilled) [5]. Similarly, we select models with varying prompt-following capabilities to maximize diversity of generated images. We use three random seeds to generate the images, creating 12 candidate images per prompt. We fix the backward mapping G as Stable Diffusion 3 to compute $s(x \to y)$. We provide ablation study on decoder choices in Appendix E.

Input source. We design our dataset to capture *dense* alignment between images and text, focusing on captioning images with rich descriptions and generating images from longer, detailed text prompts. To this end, we use the train split of Densely Captioned Images (DCI) dataset [84] for input images and texts. It contains 7.6K image-text pairs featuring high-resolution images annotated with dense, detailed captions. Due to prompt length constraints of text-to-image models, we use sDCI, a summarized version of DCI to fit within 77 tokens.

3.3. Reward Modeling

The generality of cycle-consistent preferences allows us to train a reward model in multiple ways. We explore three variants: (1) **CycleReward-I2T**: trained with image-to-text preferences $s(x \to y)$, (2) **CycleReward-I2T**: trained with text-to-image preferences $s(y \to x)$, and (3) **CycleReward-Combo**: jointly trained with both datasets.

Training details. Given a dataset of image-to-text comparison pairs (x, y_i, y_j), where image x is paired with a preferred text y_i and rejected text y_j, the loss [65, 79] is formulated as:

$$\mathcal{L}_{\text{img}} = -\mathbb{E}_{(x,y_i,y_j)\sim D_X} \left[\log \sigma \left(r_\theta(x, y_i) - r_\theta(x, y_j)\right)\right], \tag{5}$$

where $r_\theta(x, y)$ is the scalar output of the reward model.

Similarly, given a dataset of text-to-image comparison pairs (y, x_i, x_j), where text y is paired with a preferred image x_i and rejected image x_j, the loss is formulated as:

$$\mathcal{L}_{\text{text}} = -\mathbb{E}_{(y,x_i,x_j)\sim D_Y} \left[\log \sigma \left(r_\theta(x_i, y) - r_\theta(x_j, y)\right)\right]. \tag{6}$$

Finally, we also train a reward model on both datasets using the objective. We set $\lambda = 1$ for joint training.

$$\mathcal{L} = \mathcal{L}_{\text{text}} + \lambda \mathcal{L}_{\text{img}}. \tag{7}$$

Network architecture. Similar to ImageReward [92], we adopt BLIP [46] as our backbone. It consists of a ViT-L/16 encoder [17] and a $\text{BERT}_{\text{base}}$ text encoder [14] followed by a 5-layer MLP. Training details are outlined in Appendix D.

Method	*Detailed Captioning* RLHF-V	POVID	*Text-to-Image Generation* HPDv2	PaPv2	IRDB
GPT-4o	61.3%	60.0%	48.1%	45.8%	24.8%
Raw Cycle Consistency	58.6%	61.2%	60.5%	59.8%	54.5%
CycleReward-I2T	63.9%	65.6%	66.5%	65.7%	60.20%
CycleReward-T2I	57.1%	**78.2%**	**68.3%**	**66.2%**	60.2 %
CycleReward-Combo	**66.5%**	63.8%	67.7%	65.8%	**61.3%**

Table 2. Agreement rates between human preferences and those derived from GPT-4o, raw cycle consistency, and CycleReward.

4. Does cycle consistency align with human preferences?

In Table 2 we study how annotating preferences with cycle consistency compares to human labels. We measure the agreement rate between cycle consistency and human preferences across two tasks: detailed captioning and text-to-image generation. For detailed captioning, we compare to preferences from the RLHF-V [97] and POVID [106] datasets. For text-to-image generation, we compare to preferences from Human Preferences Dataset v2 (HPDv2) [89], Pick-a-Pic v2 [40], and ImageRewardDB [92]. For each dataset, we sample 1K random binary comparison pairs. We compare preferences from both "raw cycle consistency" scores $s(x \to y)$, $s(y \to x)$, and preferences using our trained reward models. We also include a comparison with GPT-4o [61] annotations as they have been shown to be useful for preference learning [48].

CycleReward achieves the highest agreement with human annotations, with CycleReward-Combo having the highest average agreement rate of 65%. While GPT-4o annotations align more closely with human preferences on detailed captioning tasks, its agreement drops significantly on text-to-image generation, with as low as 24.84% on ImageRewardDB. In contrast, raw cycle consistency maintains a consistent agreement rate across both tasks. Training a reward model on cycle consistency further improves alignment, which demonstrates the effectiveness of distilling cycle-consistent preferences into a learned reward model.

While we compare with human preferences, our aim is not to mimic them. We aim to learn *image-text alignment*, and demonstrate that cycle consistency is an effective proxy — achieving strong results without collecting *any* human labels, as shown in the following sections.

Method	DetailCaps-4870	GenAI-Bench
Vision-Language Model		
CLIPScore	51.66	49.73
VQAScore (3B)	46.84	59.54
VQAScore (11B)	50.24	**64.13**
Human Preferences		
HPSv2	54.34	56.13
PickScore	51.01	57.05
ImageReward	50.70	56.70
Cycle Consistency		
IRDB-Cycle	49.96	54.58
Raw Cycle Consistency	56.46	52.52
Cycle Consistency (Ours)		
CycleReward-T2I	51.74	55.20
CycleReward-I2T	58.02	53.49
CycleReward-Combo	**60.50**	55.52

Table 3. **Evaluating image-text alignment.** CycleReward-Combo and CycleReward-I2T outperform all approaches on detailed captioning evaluation, even those trained on human preferences. Notably, we outperform VQAScore x24 larger in model size. For text-to-image generation, CycleReward achieves similar performance to models trained on human preferences, while VQAScore outperforms others. Across both tasks, our *learned* reward model outperforms using raw cycle consistency.

5. Reward Model Evaluation

We evaluate the reward model's ability to *assess* and *improve* image-text alignment across two tasks: detailed captioning and text-to-image generation. Specifically, we evaluate CycleReward as an alignment metric (5.1), and then use it to enhance generation quality via best-of-N (5.2).

Comparison Methods. We compare against existing reference-free metrics for image-text alignment. These include: (1) **CLIPScore** [29] which measures cosine similarity between image and text embeddings of CLIP [69], (2) **ImageReward** [92], (3) **HPSv2** [89, 90] and (4) **PickScore** [40], which are trained on large human preference datasets for text-to-image generation, and (5) **VQAScore** [52] which produces alignment scores by querying a VLM with the prompt `"Does this figure show {text}?"`. For VQAScore, we compare two different model sizes: CLIP-T5-xl (3B) and CLIP-T5-xxl (11B). (6) **Raw cycle consistency** directly measures alignment scores $s(x \to y)$ for image-to-text generation and $s(y \to x)$ for text-to-image generation without learning a reward model. For image-to-text generation, this is equivalent to Image2Text2Image [32]. In our experiments, we adopt the same model configurations (i.e., choice of decoders, similarity metrics) for a fair comparison.

5.1. Metric for Image-Text Alignment

We evaluate our reward model's ability to evaluate image-text alignment across two tasks: detailed captioning and text-to-image generation.

Evaluation Benchmarks. While many benchmarks exist for short captions, few target *detailed* image captioning, and those that do often lack human labels or contain limited examples. One exception is DetailCaps-4870 [16], which evaluates captions on accuracy and inclusion across object, attribute, and relation categories. It contains 4,870 image-text pairs from ShareGPT4V [8], LLaVA 1.5, and CogVLM [31, 88], scored by three VLMs: GPT-4V [62], Gemini-1.5 Pro [82], and GPT-4o [61]. We use the mean score as a pseudo-ground truth, and measure agreement with alignment metrics using pairwise accuracy [13]. To evaluate text-to-image generation, we use GenAI-Bench [43, 52], which consists of 1,600 prompts paired with 6 generated images from different models. Each generation is annotated with human ratings based on its fidelity to the text.

Comparison with the Same Data and Backbone. We directly compare to human preference learning by training a reward model on the *same* backbone and dataset, but re-annotated with cycle consistency. We refer to this model as IRDB-Cycle, and compare against ImageReward trained on human annotations. As shown in Table 3, IRDB-Cycle achieves comparable performance, demonstrating that cycle consistency is an effective and cheaply scalable alternative for human labels. Training on *our dataset* yields further improvements, as detailed in the following sections.

Evaluating Detailed Captioning. Table 3 shows that CycleReward outperforms all existing methods by a large margin, including HPSv2, PickScore, and ImageReward, which are trained on *human* preferences. Notably, CycleReward outperforms VQAScore-11B by 10.26%, which is 24x larger in model size. It outperforms raw cycle consistency, i.e., computing cycle consistency on-the-fly as in Image2Text2Image [32], which highlights the effectiveness of distilling cycle consistency into a learned reward model.

Evaluating Text-to-Image Generation. Table 3 reports pairwise accuracy between different methods and human preferences on GenAI-Bench. CycleReward performs comparably to HPS, PickScore, and Imagereward — models trained with human annotations. CycleReward outperforms both raw cycle consistency and IRDB-Cycle, a model trained on ImageRewardDB with cycle-consistent labels. Although VQAScore (11B) aligns most with humans, our model does surprisingly well considering its small scale (477M). See Appendix F.1 for qualitative comparisons.

5.2. Best-of-N Sampling

Best-of-N (BoN) sampling is a simple strategy to improve models at test time [10, 58, 79]. The process involves generating N candidate outputs from a base model which are then ranked according to a reward model which selects the one with the highest score. The selection criterion is entirely based on the reward model, and naturally better models choose higher-quality outputs.

Evaluation Benchmarks. For image-to-text generation, we use two detailed captioning benchmarks: LLaVA-W [53] detailed captioning subset and DeCapBench [93], which assess the correctness and coverage of details (i.e., precision and recall) in generated captions. LLaVA-W^D evaluations are conducted using GPT-4o-mini [61] as the evaluator model, while DeCapBench uses DCScore [93]. For text-to-image generation, we use T2I-Compbench [33] for fine-grained preferences on six compositional categories, and the "Complex" subset of PartiPrompts [96] for complex, detailed prompts.

Improving Detailed Captioning. For each image, we perform BoN selection from a pool of 250 captions obtained from a combination of temperature, nucleus, and prompt sampling LLaVA1.5-13B [44, 55]. For image captioning tasks, BoN sampling with our reward model increases performance significantly over over alignment metrics as seen in 3. Both LLaVA-W and DeCapBench assess captions based on correctness and level of detail, and our reward model yields the largest improvement in the overall evaluation score. For further analysis we plot BoN results for the non-hallucination and comprehensiveness scores from DeCapBench and find that our model excels at describing many things in detail while maintaining correctness (albeit less accurately than VQAScore). In contrast, baselines such as VQAScore and ImageReward highly weigh accuracy to the point of preferring captions with significantly less detail.

Improving Text-to-Image Generation. For all text prompts we use SDXL-Turbo [74] to generate a pool of 100 images with different random seeds to perform BoN sampling. Figure 3 (right) shows relative performance gain using different reward models for BoN sampling. Note that our self-supervised reward models perform similarly to ImageReward which is trained with human preferences, and even outperforms on 'complex' text prompts. For specific categories in T2I-CompBench see F.4

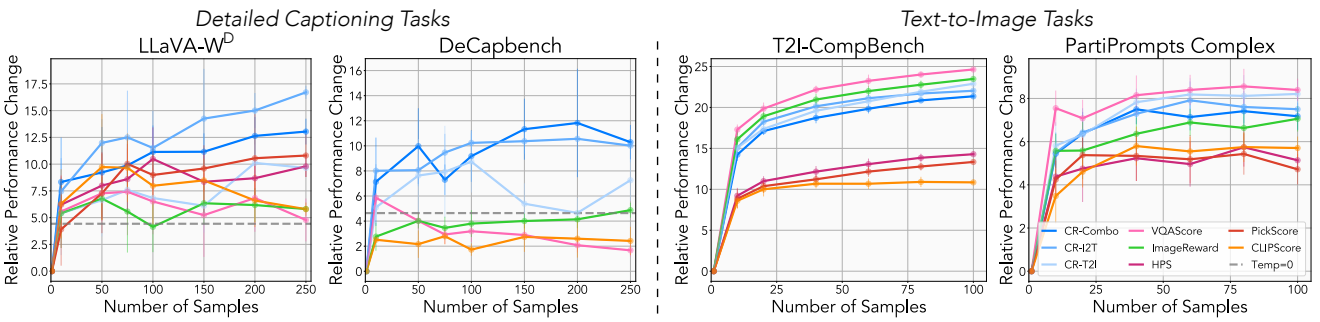


Figure 3. **Best-of-N relative performance gain.** From left to right: LLaVA-W, DeCapBench, T2I-CompBench (mean of 6 categories), and PartiPrompts (complex). In each plot, we show the relative performance gain from BoN sampling with different metrics. Feedback from our reward model lead to the greatest overall improvement for detail captioning tasks, with competitive text-to-image generation. performance to VQA-Score and Imagereward.

Ablation	DetailCaps-4870	GenAI-Bench
Image similarity		
DreamSim	**58.02**	**53.49**
LPIPS	53.16	52.97
CLIP	57.90	53.30
Text similarity		
SBERT	**51.74**	55.20
BERT	47.27	**55.52**
CLIP	49.00	54.92
T2I decoder		
Stable Diffusion3	**58.02**	53.49
FluxSchnell	56.54	53.19
SDXL-Turbo	56.42	**54.83**
I2T decoder		
LLaVA-1.5-13B	51.74	**55.20**
LLaVA-OV-7B	52.80	53.09
InternVL2-26B	**57.21**	54.46

Table 4. **Ablation study** on reward model design choices. We study the effect of different similarity metrics and decoders.**Hyojin: todo**

5.3. Ablation Study

In Table 8, we ablate several design choices for our reward model on DetailCaps-4870 and GenAI-Bench. For both benchmarks we report Pairwise Accuracy and brown entries denote design choices used in CycleReward.

Similarity Metric We study the effect of different image and text similarity metrics for computing $s(x \to y)$ and $s(y \to x)$ respectively. For image similarity we compare our choice metric DreamSim with LPIPS [101] and CLIP [69] image encoding cosine similarity. We compare our chosen text metric SBERT with BERTScore [102] and CLIP text encoding cosine similarity. Using DreamSim leads to the best reward model perform on both benchmarks, while SBERT leads to the most balanced performance on both image-to-text and text-to-image tasks.

Decoders We examine the choice of decoder model for generating image and text reconstructions used to compute $s(x \to y)$ and $s(y \to x)$ respectively. For generating image reconstructions, we compare our choice Stable Diffusion 3 with Flux-Schnell [**?**] and SDXL-Turbo [67, 72], both few step diffusion models. We find that Stable Diffusion 3 leads to the highest performance for detail captioning, while all models perform relatively similarly for text-to-image tasks with SDXL-Turbo having a slight edge. For generating text reconstructions, we compare our choice of LLaVA-1.5 13B to a smaller model LLaVA-OV-7B [44], and a larger, more performant model, InternVL-26B [63]. InternVL2-26B improves detailed captioning performance, however our chosen decoder LLaVA-1.5-13B leads to the highest pairwise accuracy for text-to-image generation.

Hyojin: add dataset size and MSE loss

6. Direct Preference Optimization

We study the alignment effect of cycle-consistent preferences with direct preference optimization (DPO) [70], which optimizes the model to prefer the chosen response over the rejected one without explicit reward modeling. For image-to-text generation, we apply DPO [70] to Qwen-VL-Chat [2] using CyclePrefDB-I2T. For text-to-image generation, we apply Diffusion DPO [86] to Stable Diffusion 1.5 [72] using our CyclePrefDB-T2I dataset. For implementation details see Appendix D.

Comparison Methods. We compare against the base model and models trained on different preference datasets. For image-to-text generation, we compare against VLFeedback [48], a vision-language feedback dataset annotated with GPT-4V. It comprises 82K instructions, including vi-

Model	*Detailed Captioning* DecapBench	LLaVA-W^{D}	*General VQA Tasks* LLaVA-W^{C}	LLaVA-W^{R}	MMHalBench	MME^{P}	MME^{C}
Qwen-VL-Chat	26.47	61.67	73.10	83.71	2.99	1460.2	368.9
DPO w/ VLFeedback	28.03	69.17	**76.39**	**89.50**	**3.32**	**1551.5**	**396.8**
DPO w/ CyclePrefDB-I2T	**30.63**	**70.00**	74.13	84.62	3.11	1485.7	386.4

Table 5. **Direct preference optimization (DPO) for image-to-text generation.** The best results are indicated in **bold**. DPO with CycleReward-VL improves the base model's performance across all tasks – including detailed captioning, perception, reasoning, and hallucination reduction – despite only containing captioning instructions. It achieves comparable or superior results to VLFeedback, a preference dataset annotated with GPT-4V spanning diverse task instructions.

Model	*T2I-CompBench* Spatial	Color	Complex	Numeracy	Shape	Texture	*Short Prompts* DrawBench	PP-Simple Detail	*Long Prompts* PP-FG Detail	PP-Complex
Stable Diffusion 1.5	11.49	36.98	34.49	44.81	37.48	40.39	28.42	7.65	7.13	6.37
Diffusion DPO w/ Pick-A-Pic	14.59	39.12	34.69	**45.88**	37.39	40.66	**30.13**	**7.73**	**7.28**	6.45
Diffusion DPO w/ CyclePrefDB-T2I	**16.55**	**42.35**	**37.75**	45.24	**38.83**	**46.67**	30.04	7.69	7.28	**6.51**

Table 6. **Direct preference optimization (DPO) for text-to-image generation.** For all evaluations, higher scores are better. T2I-Compbench and DrawBench scores range from 0 to 100 while PartiPrompt (PP) scores range from 1 to 10. In all cases, the Diffusion DPO training with CycleReward-VL outperforms the base model. Furthermore, our model often outperforms or is comparable with the Pick-A-Pic Diffusion DPO model, especially for longer text prompts.

sual question answering, image captioning and classification, reasoning, conversation, and red teaming, totaling 399K preference pairs. For text-to-image generation, we compare against Pick-A-Pic v2 [40], a human preference dataset for text-to-image generation comprising 851K comparison pairs for 58,960 unique text prompts. Note that both datasets are larger than CyclePrefDB, which consists of 398K image-to-text pairs and 468K text-to-image pairs.

Evaluation Benchmarks. We evaluate on vision-language benchmarks across diverse tasks. We evaluate on LLaVA-W^{D} [53] and DeCapBench [93] for detailed captioning. Although our dataset only contains captioning instructions, we also test generalization to other tasks: MME [21], consisting of MME^{P} for perception abilities and MME^{C} for cognition abilities such as coding and math problems, and MMHal-Bench [80] for hallucination, and LLaVA-W^{C} for conversation capabilities and LLaVA-W^{R} for reasoning. For text-to-image generation, we use T2I-Compbench [33] for compositionality, DrawBench [73] for general short prompts, and PartiPrompts [96] for dense prompts using the "Simple Detail," "Fine-Grained Detail" and "Complex" categories. For each prompt, we sample 10 images per model. To reduce variance, we repeat GPT-4o evaluations five times and report mean scores.

6.1. Results

Image-to-text generation. To our surprise, DPO fine-tuned with CyclePrefDB-I2T enhances the base model's performance across *all* vision-language tasks – including detailed captioning, perception, reasoning, and hallucination – although our dataset only contains captioning instructions. Despite our narrow task instruction and smaller size, it achieves comparable or superior results to VLFeedback, a preference dataset annotated with proprietary GPT-4V across VQA, captioning, classification, reasoning, conversation, and red teaming instructions.

Text-to-image generation. Table 6 reports evaluation results on T2I-CompBench and DrawBench (scores from 1 to 100) and PartiPrompts (scores from 1 to 10), where higher is better. Across all categories, the model trained on CyclePrefDB-T2I outperforms the base SD1.5 model and is comparable with or outperforms the Pick-A-Pic model on complex prompts, which is particularly a challenge for SD1.5. Qualitative results are shown in the Appendix 8.

7. Discussion

Our findings demonstrate that cycle consistency provides a scalable and effective supervisory signal for image-text alignment, achieving competitive performance without relying on any human-labeled data. By constructing CyclePrefDB, a preference dataset annotated via cycle consistency, we enable training reward models that generalize across both image-to-text and text-to-image tasks. These models outperform or match existing baselines on detailed captioning and compositional text-to-image benchmarks, suggesting that cycle consistency can serve as an effective alternative to human or AI-based preference annotations.

However, cycle consistency is not without limitations. The quality of supervision depends on accurate reconstruc-

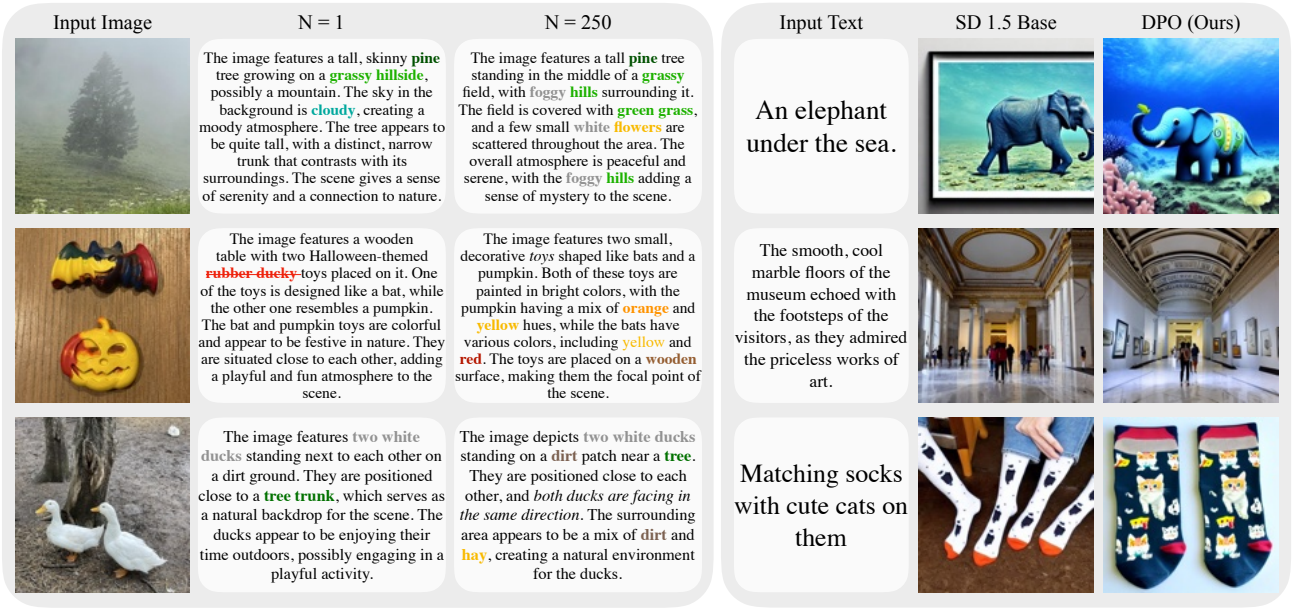


Figure 4. **Results using our method TODO: write captionTODO: change Ours → DPO with dataname** (Right:) Results from using our text-to-image cycle consistency preference dataset in Diffusion-DPO training compared to the base SD1.5 Model. Preference learning on our dataset often leads to generated images which better capture details in the input prompt.

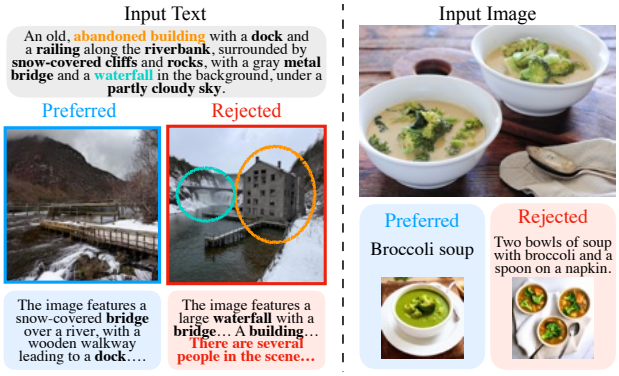


Figure 5. **Failure cases.** (Left): For the input text, cycle consistency prefers the image without a building because the 'rejected' image's text reconstruction contains a hallucination inconsistent with the original prompt. We highlight key components of the text reconstructions under each image with the incorrect text in red. (Right): The shorter caption is preferred over the more descriptive caption due to an error in text-to-image generation. We show the reconstructed image under each caption candidate.

tions from pre-trained decoders, and errors in generation can lead to misleading preference signals (Figure 5). Our method also inherits biases from the underlying models used for similarity measurement: DreamSim often favors images with similar foregrounds over backgrounds [22], and SBERT is sensitive to text style. such as token limits in Stable Diffusion or foreground bias in DreamSim. Furthermore, we observe worse text-to-image performance, which may partially stem from dataset differences. HPSv2, PickScore, and ImageReward are trained on prompts from real users often describing artwork, whereas CycleReward is trained on 6.8K LLM-summarized descriptions for natural images. Future work could address these challenges by improving reconstruction quality and prompt diversity. Furthermore, considering the generality of the framework, we can extend the framework to other modalities or reasoning tasks.

incorporating additional alignment signals, or extending the framework to instruction-following and multimodal reasoning tasks.

on aesthetics-driven tasks and visio-linguistic reasoning benchmarks like Winoground, which may stem from its training focus on semantic descriptions rather than stylistic or logical reasoning. Future work could address these challenges by improving reconstruction quality, incorporating additional alignment signals, or extending the framework to instruction-following and multimodal reasoning tasks.

Our main insight is to leverage cycle consistency scores to determine preferences rather than collecting human annotations. Although theoretically our method can be applied to any two domains, we apply it to learn a reward model for image-text alignment. We collect a dataset of images with long, descriptive image captions, and cycle consistency scores and a symmetric dataset for text cycle consistency. Our model exhibits state-of-the-art performance on detailed caption evaluation. We hope to explore how our method generalizes to new domains in the future.

Acknowledgments

Research was sponsored by the Department of the Air Force Artificial Intelligence Accelerator and was accomplished

under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of the Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. This work was supported in part by a Packard Fellowship and a Sloan Research Fellowship to P.I., by the MIT-IBM Watson AI Lab, by the Sagol Weizmann-MIT Bridge Program, and by ONR MURI grant N00014-22-1-2740.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016. 2

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1 (2):3, 2023. 7

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 3

[4] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2, 3

[5] BlackForestLabs. Announcing black forest labs. https://blackforestlabs.ai/announcing-black-forest-labs/. Accessed: 2024-09-24. 4

[6] Richard W Brislin. Back-translation for cross-cultural research. *Journal of cross-cultural psychology*, 1(3):185–216, 1970. 3

[7] David M Chan, Suzanne Petryk, Joseph Gonzalez, Trevor Darrell, and John F Canny. Clair: Evaluating image captions with large language models. In *EMNLP*, 2023. 2

[8] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2024. 6

[9] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 4

[10] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 6

[11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 20

[12] Xun Deng, Han Zhong, Rui Ai, Fuli Feng, Zheng Wang, and Xiangnan He. Less is more: Improving llm alignment via preference data selection. *arXiv preprint arXiv:2502.14560*, 2025. 16

[13] Daniel Deutsch, George Foster, and Markus Freitag. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 6

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 5

[15] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason E Weston. Chain-of-verification reduces hallucination in large language models. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, 2024. 3

[16] Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. Benchmarking and improving detail image caption. *arXiv preprint arXiv:2405.19092*, 2024. 2, 3, 6

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 5

[18] Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-Bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, et al. Dense and aligned captions (dac) promote compositional reasoning in vl models. *Advances in Neural Information Processing Systems*, 36:76137–76150, 2023. 2

[19] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 18

[20] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 3, 4, 16

[21] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation

benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 8

[22] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. 3, 9, 16

[23] Roopal Garg, Andrea Burns, Burcu Karagol Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Baldridge, and Radu Soricut. Imageinwords: Unlocking hyper-detailed image descriptions. In *EMNLP*, 2024. 1, 4

[24] Gemini. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 3

[25] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017. 3

[26] Satya Krishna Gorti and Jeremy Ma. Text-to-image-to-text translation using cycle consistent adversarial networks. *arXiv preprint arXiv:1808.04538*, 2018. 3

[27] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18135–18143, 2024. 20

[28] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. *Advances in neural information processing systems*, 29, 2016. 3

[29] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 2, 5

[30] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018. 3

[31] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, and Jie Tang. Cogagent: A visual language model for gui agents, 2023. 6

[32] Jia-Hong Huang, Hongyi Zhu, Yixian Shen, Stevan Rudinac, and Evangelos Kanoulas. Image2text2image: A novel framework for label-free evaluation of image-to-text generation with text-to-image diffusion models. In *International Conference on Multimedia Modeling*, pages 413–427, 2025. 1, 2, 4, 5, 6

[33] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023. 1, 6, 8, 18

[34] Qi-Xing Huang and Leonidas Guibas. Consistent shape maps via semidefinite programming. In *Computer graphics forum*, pages 177–186. Wiley Online Library, 2013. 3

[35] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *International Conference on Machine Learning*, 2024. 18

[36] Allan Jabri, Andrew Owens, and Alexei Efros. Space-time correspondence as a contrastive random walk. *Advances in neural information processing systems*, 33:19545–19560, 2020. 3

[37] Lei Jin, Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Annan Shu, and Rongrong Ji. Refclip: A universal teacher for weakly supervised referring expression comprehension. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2681–2690, 2023. 2

[38] Yuu Jinnai, Tetsuro Morimura, Kaito Ariu, and Kenshi Abe. Regularized best-of-n sampling to mitigate reward hacking for language model alignment. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*, 2024. 3

[39] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. Forward-backward error: Automatic detection of tracking failures. In *2010 20th international conference on pattern recognition*, pages 2756–2759. IEEE, 2010. 1

[40] 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, 3, 4, 5, 8

[41] Alon Lavie and Abhaya Agarwal. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, 2007. 2

[42] Yebin Lee, Imseong Park, and Myungjoo Kang. Fleur: An explainable reference-free evaluation metric for image captioning using a large multimodal model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3732–3746, 2024. 2

[43] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024. 6

[44] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 4, 6, 7, 17, 18

[45] Chunyuan Li, Hao Liu, Changyou Chen, Yuchen Pu, Liqun Chen, Ricardo Henao, and Lawrence Carin. Alice: Towards understanding adversarial learning for joint distribution matching. *Advances in neural information processing systems*, 30, 2017. 15

[46] 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. 3, 5, 17

[47] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with

frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3, 4, 17

[48] Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong, and Qi Liu. VLFeedback: A large-scale AI feedback dataset for large vision-language models alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6227–6246, 2024. 1, 3, 4, 5, 7

[49] Tianhong Li, Sangnie Bhardwaj, Yonglong Tian, Han Zhang, Jarred Barber, Dina Katabi, Guillaume Lajoie, Huiwen Chang, and Dilip Krishnan. Leveraging unpaired data for vision-language generative models via cycle consistency. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[50] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore, 2023. Association for Computational Linguistics. 1

[51] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 20

[52] 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. In *European Conference on Computer Vision*, pages 366–384. Springer, 2024. 2, 5, 6

[53] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 3, 6, 8, 18

[54] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 3, 4, 16

[55] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 4, 6

[56] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *7th International Conference on Learning Representations (ICLR)*, 2019. 17

[57] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2019. 16, 17

[58] Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025. 6

[59] Nico Messikommer, Daniel Gehrig, Mathias Gehrig, and Davide Scaramuzza. Bridging the gap between events and frames through unsupervised domain adaptation. *IEEE Robotics and Automation Letters*, 7(2):3515–3522, 2022. 3

[60] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021. 3

[61] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, . Accessed: 2024-09-24. 3, 5, 6, 18, 20

[62] OpenAI. Gpt-4v(ision) system card. https://openai.com/index/gpt-4v-system-card/, . Accessed: 2023-09-31. 1, 4, 6

[63] OpenGVLab. Internvl-2.0. 2024. 4, 7, 17

[64] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 2

[65] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022. 1, 3, 4

[66] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, 2002. 2

[67] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 4, 7, 17

[68] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation, 2023. 3

[69] 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, 5, 7, 17

[70] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023. 1, 3, 7

[71] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019. 4, 16

[72] 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*, pages 10684–10695, 2022. 4, 7, 17, 20

[73] 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. 8, 18

[74] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023. 6, 18

[75] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 3

[76] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6649–6658, 2019. 3

[77] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 3

[78] Sahand Sharifzadeh, Christos Kaplanis, Shreya Pathak, Dharshan Kumaran, Anastasija Ilic, Jovana Mitrovic, Charles Blundell, and Andrea Banino. Synth2: Boosting visual-language models with synthetic captions and image embeddings. *arXiv preprint arXiv:2403.07750*, 2024. 3

[79] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020. 3, 4, 6

[80] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13088–13110, 2024. 3, 8

[81] Narayanan Sundaram, Thomas Brox, and Kurt Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *European conference on computer vision*, pages 438–451. Springer, 2010. 1

[82] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 6

[83] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 17

[84] Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26700–26709, 2024. 3, 4, 18

[85] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 2

[86] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024. 1, 3, 7

[87] Fan Wang, Qixing Huang, and Leonidas J Guibas. Image co-segmentation via consistent functional maps. In *Proceedings of the IEEE international conference on computer vision*, pages 849–856, 2013. 3

[88] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2023. 6

[89] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 1, 2, 3, 4, 5

[90] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023. 1, 2, 3, 5

[91] Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, Jiayan Teng, Zhuoyi Yang, Wendi Zheng, Xiao Liu, Ming Ding, Xiaohan Zhang, Xiaotao Gu, Shiyu Huang, Minlie Huang, Jie Tang, and Yuxiao Dong. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation, 2024. 2

[92] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3, 4, 5, 16, 17

[93] Qinghao Ye, Xianhan Zeng, Fu Li, Chunyuan Li, and Haoqi Fan. Painting with words: Elevating detailed image captioning with benchmark and alignment learning. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 3, 6, 8

[94] Min-Hsuan Yeh, Leitian Tao, Jeffrey Wang, Xuefeng Du, and Yixuan Li. How reliable is human feedback for aligning large language models? *arXiv preprint arXiv:2410.01957*, 2024. 16

[95] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image trans-

lation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017. 3

[96] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022. 6, 8

[97] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816, 2024. 3, 5

[98] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3

[99] 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

[100] Christopher Zach, Manfred Klopschitz, and Marc Pollefeys. Disambiguating visual relations using loop constraints. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1426–1433. IEEE, 2010. 3

[101] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7, 17

[102] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 7, 17

[103] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023. 3

[104] Tinghui Zhou, Yong Jae Lee, Stella X Yu, and Alyosha A Efros. Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2015. 3

[105] Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A Efros. Learning dense correspondence via 3d-guided cycle consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 117–126, 2016. 3

[106] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, 2024. 5

[107] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1, 3

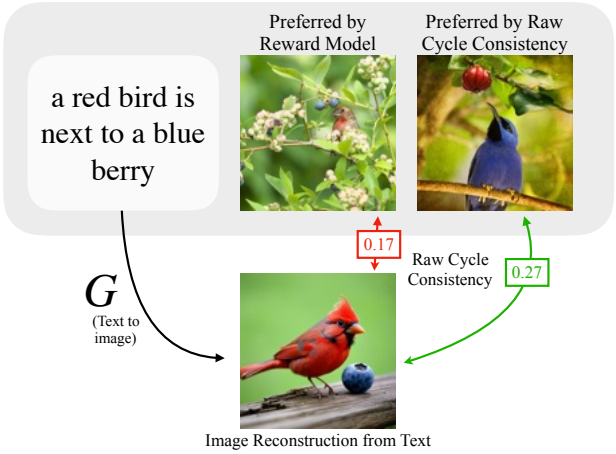


Figure 6. Pointwise consistency $s(x \to y)$ compares the original image and its reconstruction (values indicated in each box). Although the reconstructed image (bottom) perfectly follows the prompt, it is visually more similar to the image of the blue bird, resulting in failed alignment. In contrast, our reward model CycleReward-I2T in Section 3.3 learns the correct alignment.

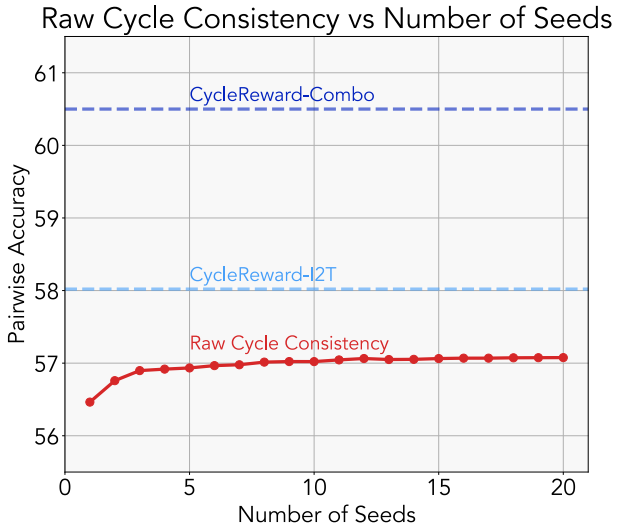


Figure 7. **Raw cycle consistency performance with increasing number of samples.** We plot DetailCaps-4870 benchmark performance (Pairwise Accuracy) for raw cycle consistency calculated over mutliple samples (random seed sampling). Despite increasing number of seeds, raw cycle consistency performance does not come close to reward model performance.

Appendix

A. Cycle Consistency and Pointwise Mutual Information

Let X and Y be random variables that take on realizations x and y, respectively. In Section 3 X and Y represent images and texts respectively, but note how our cycle consistency score (Equation 3) and preference creation (Equations 4, **??**) are general to any X and Y. We now focus on the general case.

In Equation 3, we define $s(x \to y)$ and $s(y \to x)$ with respect to fixed backward mappings $G : Y \to X$ and $F : X \to Y$ respectively. If F, G are stochastic mappings, then we can view G as sampling some $x' = G(y)$ from the distribution $p_G(X|Y = y)$ - a distribution which is determined by G. Symmetrically, we can view F as sampling $y' = F(x)$ from the distribution $p_F(Y|X = x)$ determined by F. We then argue that distributionally,

$$\begin{aligned} s(x \to y)_d &= \log p(x|y) \\ s(y \to x)_d &= \log p(y|x) \end{aligned} \tag{8}$$

caro: add intuition If the two distributions p_F and p_G are sampling from the same underlying distribution, we can define joint distributional cycle consistency score:

$$\begin{aligned} s(x, y)_d &= s(x \to y)_d + s(y \to x)_d \\ &= \log p(x|y) + \log p(y|x) \quad x, y \sim p(X, Y) \end{aligned} \tag{9}$$

Mutual Information Following the connection previous work [45]**caro: cited alice here** has made between cycle consistency and mutual information, we rewrite the joint reward as follows:

$$\begin{aligned} s(x, y)_d &= \log p(x|y) + \log p(y|x) \\ &= \log \frac{p(x, y)}{p(y)} + \log \frac{p(x, y)}{p(x)} \\ &= \log \frac{p(x, y)^2}{p(x)p(y)} \\ &= \log p(x, y) + \text{PMI}(x, y) \end{aligned} \tag{10}$$

Therefore, we can view the joint cycle consistency score as measuring both the likelihood of the pairing $p(x, y)$ and the pointwise mutual information. In turn, CycleReward prefers x, y pairings which are both high probability and informative of each other.

B. Benefits from Reward Modeling

Because our reward model is trained with preferences from cycle consistency, it is natural to assume that the performance of raw cycle consistency scores $s(x \to y)$ and $s(y \to x)$ would be an upper bound for our reward model. In contrast, our trained reward models outperform raw cycle consistency on all benchmarks reported in Section 5 in both mapping directions.

Albeit computationally slow, averaging raw cycle consistency scores over multiple reconstructions as in Equation 11 could provide more accurate alignment measure-

ments than just a single forward pass.

$$s^*(x \to y) = \frac{1}{N}\sum_{n=1}^{N} ||x - g(y, z_n)|| \qquad z_n \sim \mathcal{N}(0, I) \tag{11}$$

Figure 7 plots DetailCaps-4870 benchmark performance against the number of samples used to compute raw cycle consistency alignment with random seed sampling from SD3. Although using more seeds generally benefits raw cycle consistency, improvement tapers off after about 5 seeds and never reaches the performance of CycleReward.

Figure 6 qualitatively compares alignment computed by raw cycle consistency against our reward model. From the rich visual descriptions in our dataset, the reward model has learned that the image of the red bird corresponds best with the text description. In contrast, raw cycle consistency attempts to reconstruct the original input from the input prompt. Due to the lack of fine-grained visual information in the text, the reconstruction is more of a typical, object-centered bird image that happens to be structurally similar to the image of the blue bird over the red bird. This finding highlights additional benefits of distilling cycle consistency to a reward model – beyond speed and differentiability.

C. Dataset Details

Image-to-text comparison pairs Given an image, we first generate a set of corresponding synthetic captions with varying descriptiveness and hallucination. Different captions are created from a selection of image-to-text models $F_1, ..., F_{11}$ described in the model pool.

Given an image x and its corresponding text description $F(x)$, we compute cycle consistency by measuring the distance between the original image and its reconstruction generated by a text-to-image model, formulated as $||x - G(F_i(x))||_{\text{img}}$. We use Stable Diffusion 3 [20] as G and DreamSim [22] as the perceptual distance metric d_{img}. The cycle consistency score is used to determine preferred and rejected captions. Each pair consists of an image and two text descriptions – preferred and rejected – generated from the same image. The final dataset after filtering contains 398K comparison pairs. See Table 7 for exact prompts and Appendix D for the dataset filtering procedure.

Text-to-image comparison pairs We generate corresponding synthetic images for an input texts using 4 image-to-text models $G_1, ..., G_4$ detailed in the model pool. Given a prompt y and its synthesized image $G(y)$, we compute cycle consistency between the original prompt and its reconstruction from an image-to-text model, formulated as $||y - F(G_i(y))||_{\text{img}}$. We use LLaVA-1.5-13B [54] as F and SBERT [71] as $||\cdot||_{\text{text}}$ to measure similarity between the input and reconstructed text. Again, cycle consistency score is used to label preferred and rejected images. The final dataset consists of 468K comparison pairs.

Image and Text Reconstructions We provide examples of reconstructed images and texts used to create comparison pairs in our dataset. Figure 17 shows examples of original image, text, reconstructed image, and DreamSim [22] distance used to create our dataset (note lower DreamSim score corresponds to more similar images). Figure **??** shows the symmetric examples for text cycle consistency (note that higher SBERT score indicates stronger similarity). **caro: add the failure case figures**

Prompt Choice To ensure that all image-to-text models can produce image descriptions to the best of their ability, we use the prompt recommended by the model distributor, as shown in Table 7.

Model	Prompt
BLIP2	"this is a picture of"
LLaVA1.5	"Write a detailed description of the given image."
LLaVA1.6	"Write a detailed description of the given image."
LLaVA-OV	"Write a detailed description of the given image."
InternVL2	"Please describe the image in detail."

Table 7. Prompts used for image-to-text models.

D. Model training details

D.1. Reward Modeling

We use the AdamW optimizer [57] with a batch size of 2048 for 2 epochs. The learning rate is set to 3e-5 with a weight decay of 1e-4 for optimizing $\mathcal{L}_{\text{text}}$, while $\mathcal{L}_{\text{img}}$ and joint training use a learning rate of 2e-5 with no weight decay. We set $\lambda = 1$ for joint training. Following the setup in [92], we fix 70% of the transformer layers during training, which we found to outperform full fine-tuning. All models are trained using 8 H100 GPUs.

Dataset filtering Common strategies for filtering human preferences include: (1) removing duplicate entries, (2) filtering out cases where both responses are harmful or irrelevant [94], and (3) excluding low-margin examples where one response is only marginally better than the other [12]. Following these principles, we adopt a similar filtering strategy by removing duplicate captions, excluding examples where the reward difference is within a certain threshold, i.e., $|r_i - r_j| < \tau_{\text{sim}}$, and discarding comparison pairs where the preferred reward is below a threshold, i.e., $r_i < \tau_{\text{neg}}$. In practice, we use $\tau_{\text{sim}} = 0.005$, $\tau_{\text{neg}} = 0.7$ for DreamSim, and $\tau_{\text{neg}} = 0.4$ for SBERT. In practice, training with this

Ablation	DetailCaps-4870	GenAI-Bench
Image similarity		
DreamSim	**58.02**	**53.49**
LPIPS	53.16	52.97
CLIP	57.90	53.30
Text similarity		
SBERT	**51.74**	55.20
BERT	47.27	**55.52**
CLIP	49.00	54.92
T2I decoder		
Stable Diffusion3	**58.02**	53.49
FluxSchnell	56.54	53.19
SDXL-Turbo	56.42	**54.83**
I2T decoder		
LLaVA-1.5-13B	51.74	**55.20**
LLaVA-OV-7B	52.80	53.09
InternVL2-26B	**57.21**	54.46
RM Backbone		
BLIP	**58.02**	**53.49**
SigLIP2	53.00	52.57

Table 8. Ablation study. We study the effect of different similarity metrics, decoders, and backbones on reward modeling.

dataset filtering leads to a small performance gain on alignment benchmarks and a bigger performance gap in Best-of-N experiments.

E. Ablation Study

In Table 8, we ablate several design choices for our reward model on DetailCaps-4870 and GenAI-Bench. For both benchmarks we report Pairwise Accuracy and brown entries denote design choices used in CycleReward.

Similarity Metric We study the effect of different image and text similarity metrics for computing $s(x \to y)$ and $s(y \to x)$ respectively. For image similarity we compare our choice metric DreamSim with LPIPS [101] and CLIP [69] image encoding cosine similarity. We compare our chosen text metric SBERT with BERTScore [102] and CLIP text encoding cosine similarity. Using DreamSim leads to the best reward model perform on both benchmarks, while SBERT leads to the most balanced performance on both image-to-text and text-to-image tasks.

Decoders We examine the choice of decoder model for generating image and text reconstructions used to compute $s(x \to y)$ and $s(y \to x)$ respectively. For generating image reconstructions, we compare our choice Stable Diffusion 3 with Flux-Schnell [**?**] and SDXL-Turbo [67, 72], both few step diffusion models. We find that Stable Diffusion 3 leads to the highest performance for detail captioning, while all models perform relatively similarly for text-to-image tasks with SDXL-Turbo having a slight edge. For generating text reconstructions, we compare our choice of LLaVA-1.5 13B to a smaller model LLaVA-OV-7B [44], and a larger, more performant model, InternVL-26B [63]. InternVL2-26B improves detailed captioning performance, however our chosen decoder LLaVA-1.5-13B leads to the highest pairwise accuracy for text-to-image generation.

Model Backbone We study different model architectures for reward model training. Note that we use the BLIP [46, 47] model backbone for our reward model as presented in ImageReward [92]. BLIP outperforms the latest model backbone SigLIP2 [83], which may be due to its image-grounded text encoder, unlike SigLIP2's separate image and text encoders, as also evidenced in ImageReward which reports improvements using BLIP over CLIP.

E.1. DPO

We perform DPO to align Qwen-VL-Chat using our preference dataset. The model is trained for 5 epochs with the AdamW optimizer [56] and a weight decay of 0.05. We apply a cosine learning rate schedule with a warmup ratio of 0.1 and a peak learning rate of 1×10^{-5}. Training is performed with a global batch size of 256. To enable more efficient training, we adopt LoRA tuning. The model is trained using 4 H100 GPUs.

E.2. Diffusion-DPO

We use the Diffusion-DPO objective to align Stable Diffusion 1.5 [72] with preferences in our dataset. We use the AdamW optimizer [57] and train with an effective batch size of 512 (batch size 1 with 128 gradient accumulation steps on 4 H100 GPUs). We use learning rate 5×10^{-8} and set $\beta = 1000$ and train for 1500 steps. Similarly to the Diffusion-DPO Pick-A-Pic model, we validate checkpoints with 380 prompts from the **TODO: dataset name** validation set and calculate the mean alignment using the CycleReward-T2I reward model.

F. Additional Results

F.1. Alignment Preferences

Figure 9 shows qualitative examples of our reward model versus other alignment metrics and ground truth preferences in purple. Overall, our reward models are more successful at assessing detailed captions and while performing competitively on evaluating text-to-image generation.

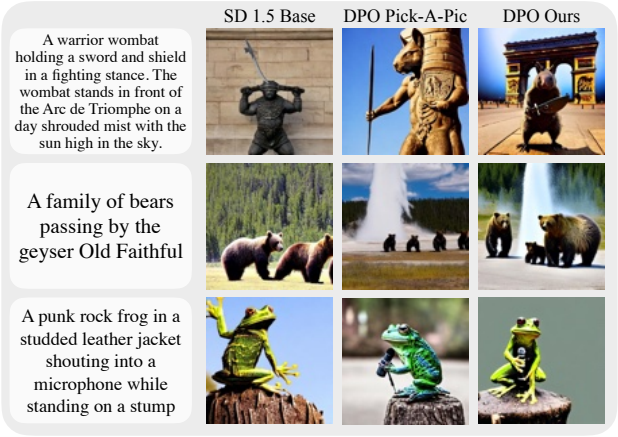


Figure 8. **Generated images from Diffusion DPO training.** From left to right: input text prompts and corresponding image generations from Stable Diffusion 1.5, Diffusion-DPO model trained with Pick-A-Pic, and Diffusion-DPO model trained with our dataset CycleReward-VL (CR-VL). Although all models still contain artifacts, our model can depict complex visual details and often outperforms the Pick-A-Pic model trained with human preferences.**TODO: fix dataset name in figure**

F.2. DrawBench

We additionally add DrawBench [73] as one of our text-to-image Best-of-N sampling tasks. DrawBench contains 200 short prompts designed to generally probe the abilities of text-to-image models. This suite is evaluated with GPT4-o [61] which assigns a binary score indicating whether the generated image is a good representation of the input text description. Similarly to the T2I-Compbench task, we use SDXL-Turbo to generate images [74]. Relative performance gain is shown in Figure 10. VQAScore outperforms all other metrics, and on longer prompts ImageReward leads to better improvement than our method.

F.3. Winoground

Benchmarking We use the Winoground dataset to benchmark performance on visio-linguistic compositional reasoning in Figure 11. We also provide a qualitative examples in Figure 12. The dataset comprises 400 examples, each containing two image-text pairs where the texts use the same words in different orders to convey different meanings. Performance is measured by how often a metric matches the correct image with its corresponding text. Surprisingly, CycleReward variants, trained solely on self-supervised rewards, are competitive with – or even outperform – ImageReward, which relies on expert human annotations. All CycleReward variants are better at selecting text for an image (text score) than selecting images from a given description (image score). While our method outperforms CLIPScore and on-the-fly cycle consistency score, it is outperformed by VQAScore, which benefits from LLM scale (x6 and x24 larger than other methods).

F.4. Best-of-N

Figure 13 and Figure 14 show how different metrics qualitatively affect Best-of-N sampling results for detailed captioning and generated images from text prompts respectively. We show the initial (Best-of-1) output and compare to the final output selected from 250 or 100 samples for captions and generated images respectively.

Sampling Settings To obtain captions for Best-of-N sampling, we used a combination of temperature, nucleus, and prompt sampling with model LLaVA1.5-13B [44, 53]. We set temperature to 1.0, top p to 0.7 respectively, and choose prompts randomly from the original LLaVA dataset prompts [53].

T2I-CompBench Categories We provide Best-of-N plots for individual categories in T2I-CompBench [33].

F.5. DPO

Figure 8 shows comparisons between the base Stable Diffusion 1.5 model, the Diffusion-DPO model trained with Pick-a-Pic v2, and the Diffusion DPO model trained with our dataset. Training with cycle consistency preferences from our text-to-image dataset achieves comparable results as training with Pick-a-Pic model, despite lacking human labels. Furthermore, our dataset is about half the size of Pick-a-Pic v2.

G. Reward Model Trends

We investigate how text and image properties affect different metrics' alignment preferences for the following factors: caption density, object hallucination, image density, and resolution in Figure 15. For each specific factor, we plot the alignment score for individual image, text pairs based on the relevant image or text characteristic. The title of each plot reports the Pearson correlation coefficient between the alignment score and respective factor. We also display the line of best fit. Note that the scale and range of alignment scores are different and not comparable between metrics. Because of this we instead focus on overall trends and correlations between each factor and alignment.

Caption Length To examine which reward models generally prefer long or short captions, we first create a dataset of images paired with captions of various lengths. We utilize the test and validation sets of the DCI [84] dataset for this task, where each image is paired with a long, descripitive text. For each image, we use an LLM (Meta-Llama-3.1-8B-Instruct [19]) to create captions of different lengths but asking for summaries with different numbers of words, similarly to Huh et al. [35]. We ask for summaries of lengths 5, 10, 20, ..., 100 words, and sample 5 different captions

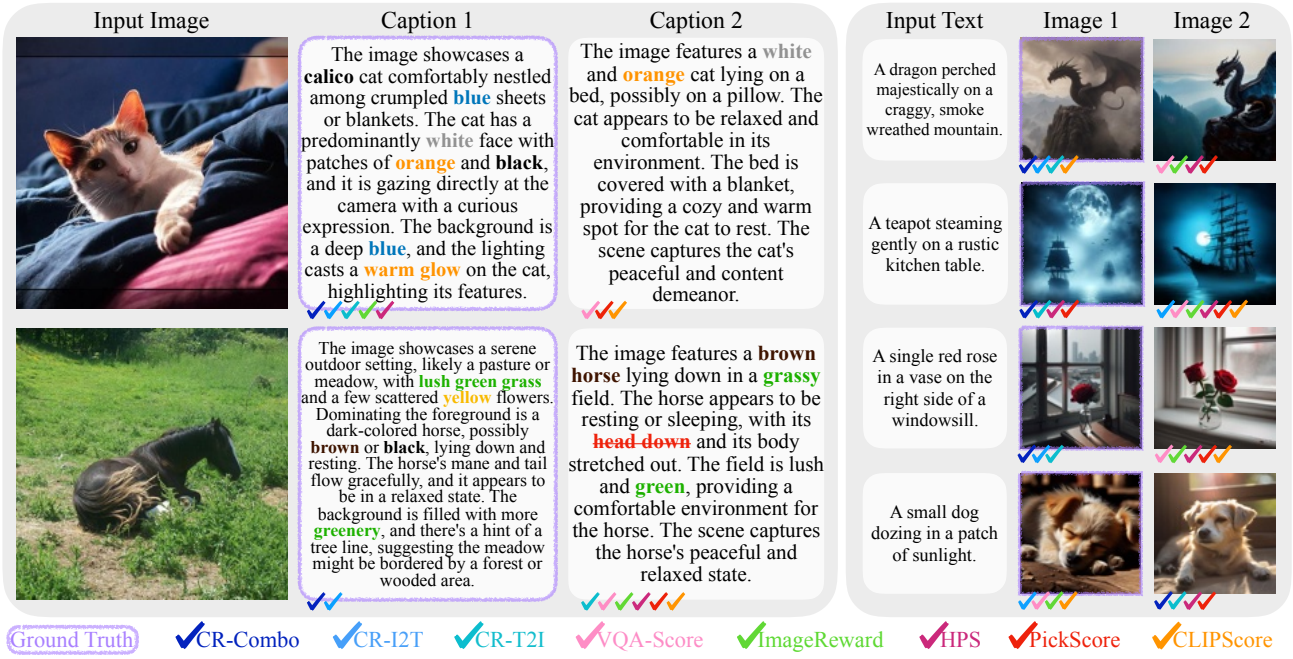


Figure 9. **Alignment metric preferences on DetailCaps-4870 and GenAI-Bench.** Our reward model excels at identifying detailed captions while performing competitively on GenAI-Bench. We also provide the ground truth label in purple.

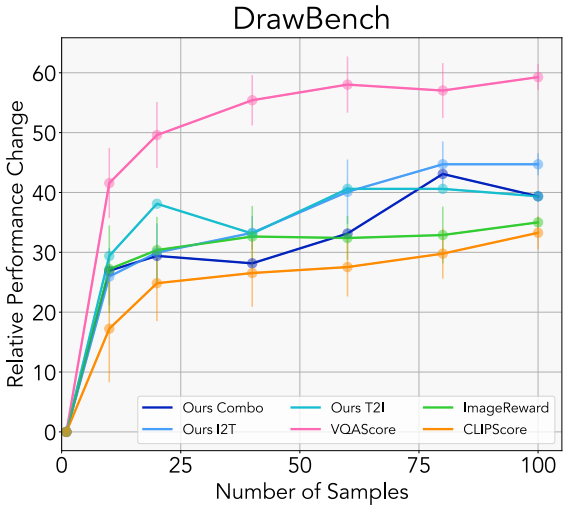


Figure 10. **Best-of-N relative performance change for DrawBench.** VQAScore shows superior inference-time scaling with our reward model coming second. Note that DrawBench mainly consists of very short prompts.

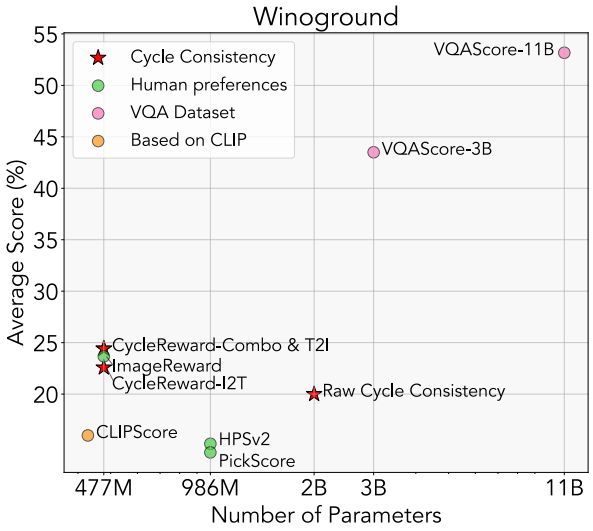


Figure 11. **Metric performance on Winoground.** For each metric we report the average of "text" and "image" scores. VQAScore, the largest model, outperforms all approaches on winoground. CycleReward performs similar to ImageReward.

for each length with temperature 0.6 and top p 0.9. This results in 11241 unique image, caption pairs after eliminating duplicates and removing "here is a summary" text.

In Figure 15(top row), we plot the alignment trend for different metrics versus caption length. The Pearson correlation coefficient is reported at the top of each plot. Note that the alignment score scales are different for each metrics and therefore not directly comparable. Instead, we focus on the overall trends. Because captions can be informative or contain mistakes regardless of their lengths, we expect these

Figure 12. **Examples of alignment model preferences on Winoground.** The first and second rows show examples for the text score and image score settings respectively. Our model is able to achieve a reasonable text score but struggles on image scores. Note that many Winoground captions are not visual descriptions but instead designed for testing reasoning and compositionality - qualities which VLM models are often evaluated on.

plots to be noisy. All methods, except for CycleReward-I2T and ImageReward, have positive Pearson Correlation coefficients - meaning they in general longer captions are preferred. However the correlation between caption length and alignment is much weaker for VQAScore and CLIP compared to CycleReward-Combo and CycleReward-I2T.

Hallucination Rate To view how hallucinations affect alignment preferences, we use the M-HalDetect dataset [27]. This dataset contains images paired with captions from InstructBLIP [11]. We use the validation and training sets for this dataset totaling 14143 image caption pairs. Each caption is divided into sections which have been annotated for their accuracy and having hallucinations. We compute the fraction of hallucinated parts in each caption and plot this value against the alignment in Figure 15(middle row). All metrics tend to prefer captions with less hallucinations (lower hallucination rate), although with different correlation strengths - VQAScore having the strongest correlation followed by CycleReward-T2I and CycleReward-Combo.

Image Resolution For text-to-image, we examine how images of different resolutions affect alignment with the text. To this end, we gather 100 "upsampled" text descriptions created by prompting GPT-4o[61] to add details to short captions from MSCOCO [51]. Text descriptions are encouraged to be visually informative and no longer than 77 tokens. We use SDXL [72] to generate images for each text description at 512×512 resolution. We resize the images to resolutions 256, 128, 64, 32, 16 and compute alignment at each stage in Figure 15(bottom row). For all metrics, alignment is generally not affected when resizing from 512 to 256 and 128 pixels, and then drops off steeply as the resolution goes from 64 to 16. Note that CycleReward and ImageReward preprocess images to be size while CLIP and VQAScore preprocessing resizes image to 336.

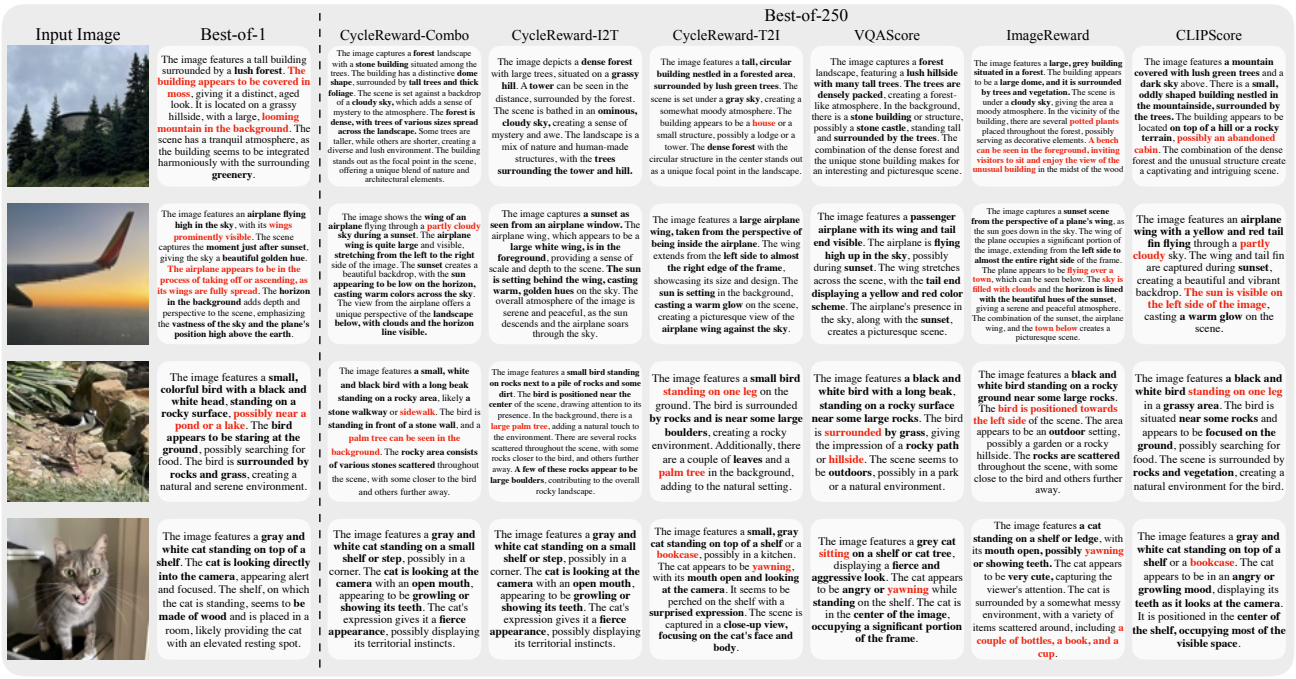


Figure 13. **DeCapBench Best-of-1 to Best-of-250 sampling results for different metrics.** Overall our model increases the level of detail in captions while avoiding severe hallucinations.

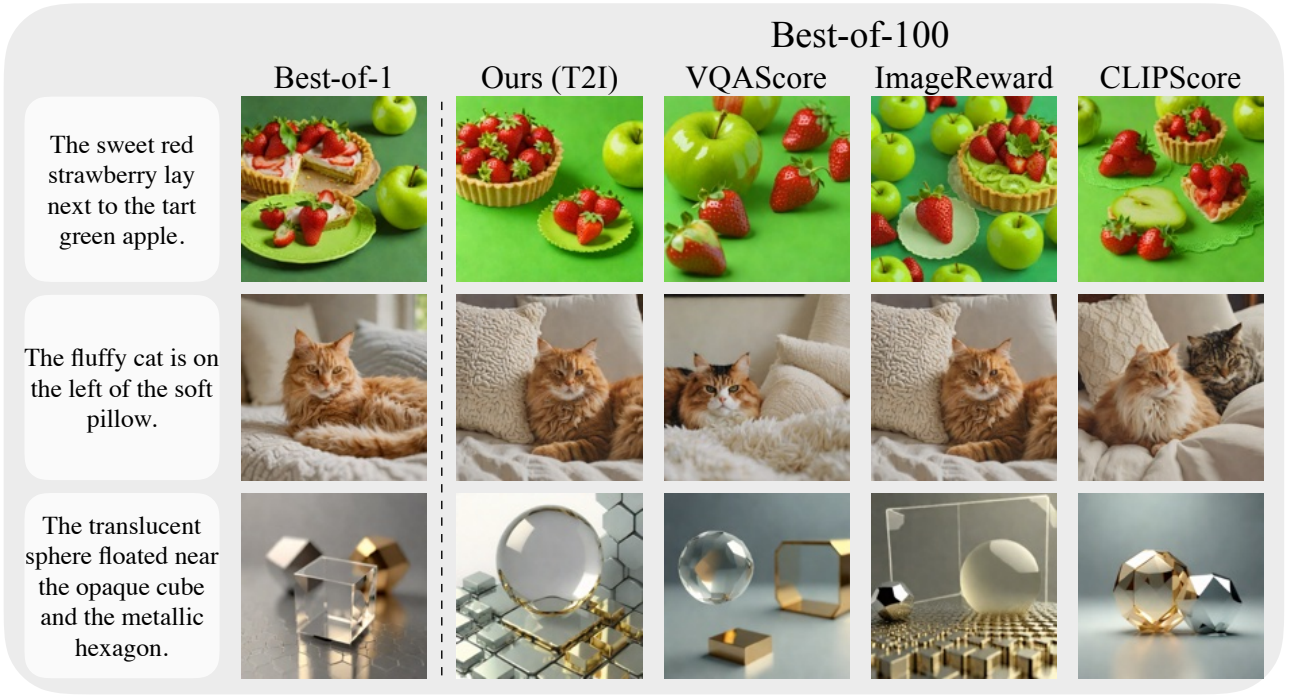


Figure 14. **T2I-CompBench Best-of-1 to Best-of-100 sampling results for different metrics.** Optimizing with our reward model generally improves results, while VQAScore excels at following positional relationships.

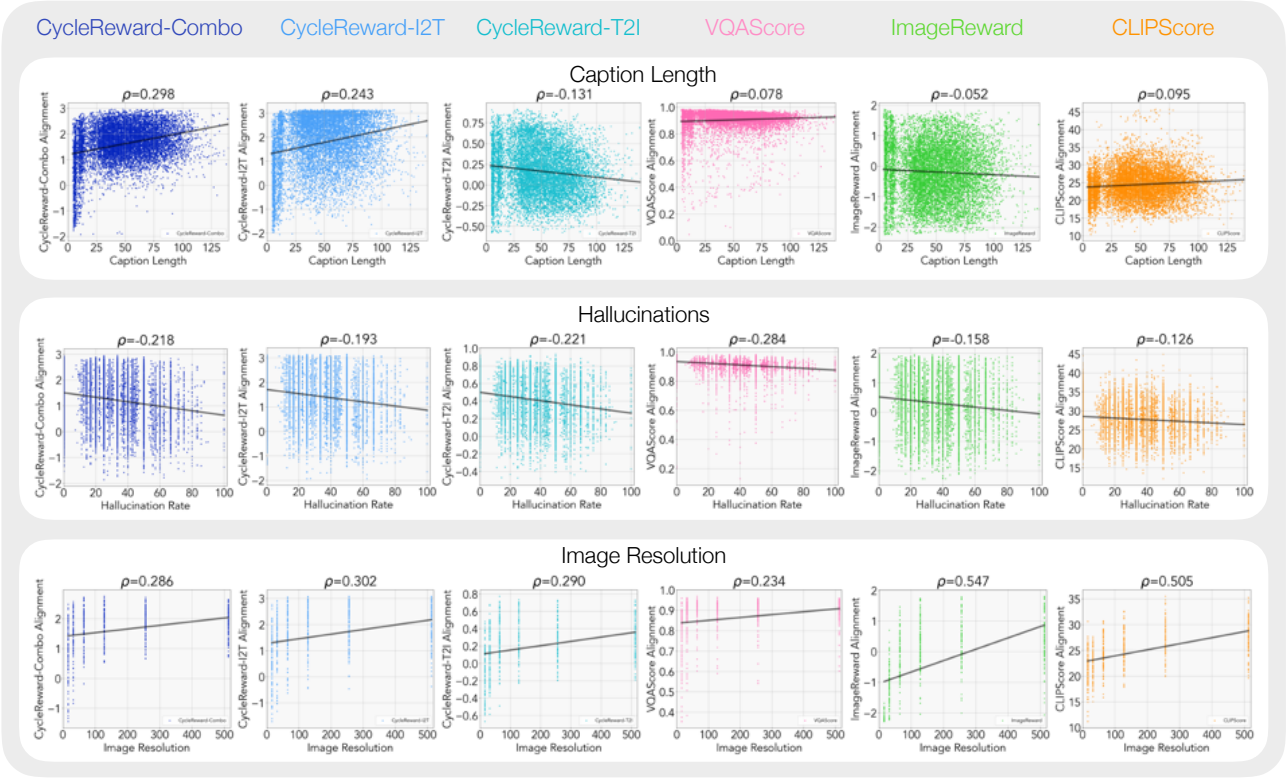


Figure 15. **Text and image data trends for different alignment metrics.** For each metric listed at the top, we plot how various factors (shown on each row) affect alignment scores. Note that different alignment measurements are not comparable by scale, but their correlation with each specific factor can be measured. CycleReward-I2T and CycleReward-Combo both tend to prefer longer captions, while models trained with text-to-image comparison pairs (ImageReward and CycleReward-T2I) generally prefer shorter captions. In terms of number of hallucinations and image operations, we find that all metrics show the same direction of correlation - albeit some metrics such as VQAScore and CycleReward are more sensitive to text inaccuracies.

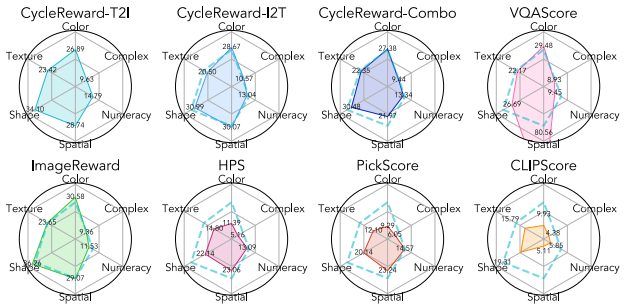


Figure 16. **T2I-Compbench relative performance gain from** 1 **to** 100 **Best-of-N sampling for different metrics on** 6 **categories**. Our metric has the most effect for complex prompts. We mark our T2I model's results with a dashed line in other charts for comparison. Although each metrics has its own strength, ImageReward, a supervised model, has the most balanced performance followed by our self-supervised method

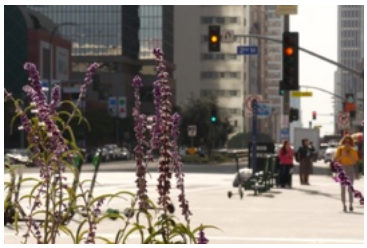
Input Image | Reconstructed Image | Input Image | Reconstructed Image




[**0.181**] The image depicts a beautifully crafted **latte** art coffee cup placed on a saucer. The cup and saucer are adorned with intricate **blue floral** patterns, giving them a classic and elegant appearance. The latte art on the coffee is a **leaf design**, crafted with precision, featuring a swirl of **white foam** on top of a rich, **creamy brown coffee**. The background is **dark**…

[**0.232**] The image depicts a large, historic square with a prominent **obelisk** in the center. The obelisk is tall and cylindrical, made of **stone**, and topped with a **statue**. The sky above is partly cloudy with patches of **blue**, suggesting a partly **sunny** day. Surrounding the obelisk are several **buildings** with distinct architectural styles. On the left side…






[**0.208**] The image depicts a bustling **urban street scene** at **night**, likely in a busy commercial area. The street is lined with various **shops and restaurants**, each adorned with **bright**, **colorful** signage in both **English and Japanese**. The most prominent **sign** in the image is the iconic red and **yellow McDonald's** logo, indicating the presence of a **McDonald's** restaurant. Adjacent to it…

[**0.262**] The image depicts a street scene in an **urban** environment, likely in a **city**. The foreground is dominated by a **vibrant purple flowering plant** with long, **slender** stems and clusters of **small, purple flowers**. The plant is in sharp focus, with its vibrant **purple** blossoms standing out against the **blurred background**. The background features a **busy street scene** with several **pedestrians** and **vehicles**…




[**0.282**] The image depicts a **vibrant cityscape** at **night**, showcasing a bustling **urban** environment illuminated by **numerous lights**. The foreground features a **dense cluster of buildings**, including residential and commercial structures, with **varying heights and architectural styles**. The streets are filled with **bright lights**, indicating active businesses and possibly nightlife. A **prominent highway runs horizontally** across the middle of the image, with **streaks of light**…

[**0.347**] The image depicts a narrow, **cobblestone** street leading towards a **stone wall** on the **right** side and a building on the **left** side. The street is flanked by a large, **lush green tree** on the **left**, which stands out prominently against the **clear blue sky**. The building on the **left** appears to be a residential or commercial structure with a **flat roof** and a few **antennas** on top…

Figure 17. **Examples of image cycle consistency from different model combinations**.**TODO: remake this**