ImageReward: Learning and Evaluating Human Preference for Text-to-Image Generation

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong,
Qinkai Li, Ming Ding, Jie Tang, Yuxiao Dong

Tsinghua University and Zhipu AI

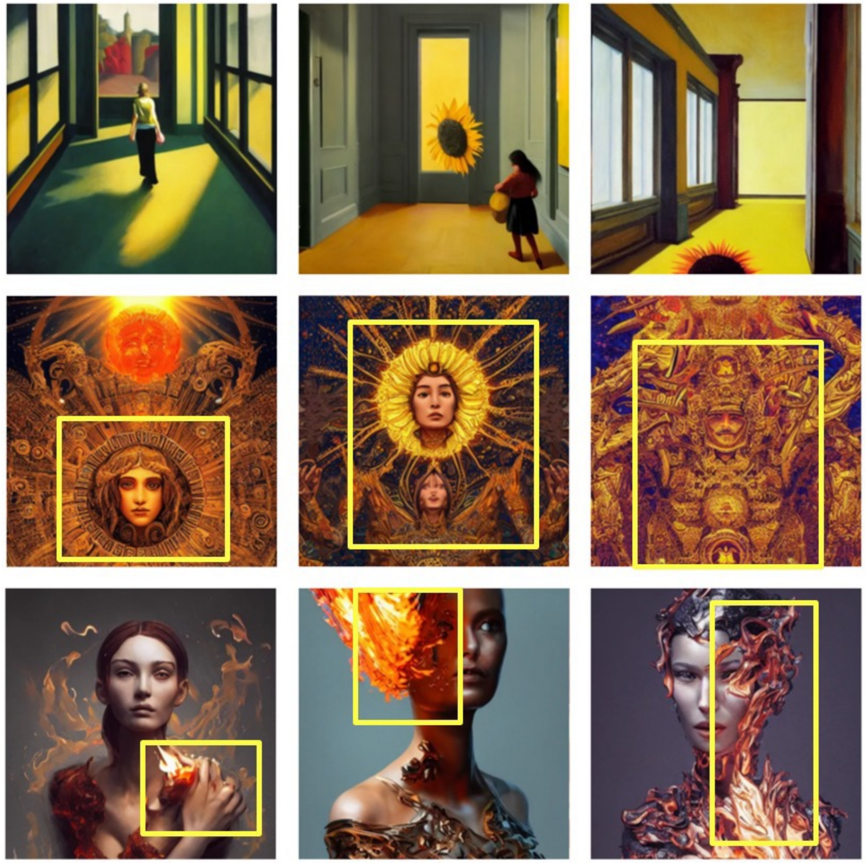
Outline

Issues in Generated Images

(a) A painting of a girl walking in a hallway and suddenly finds a giant sunflower on the floor blocking her way.

(b) Coronation of the sun emperor.

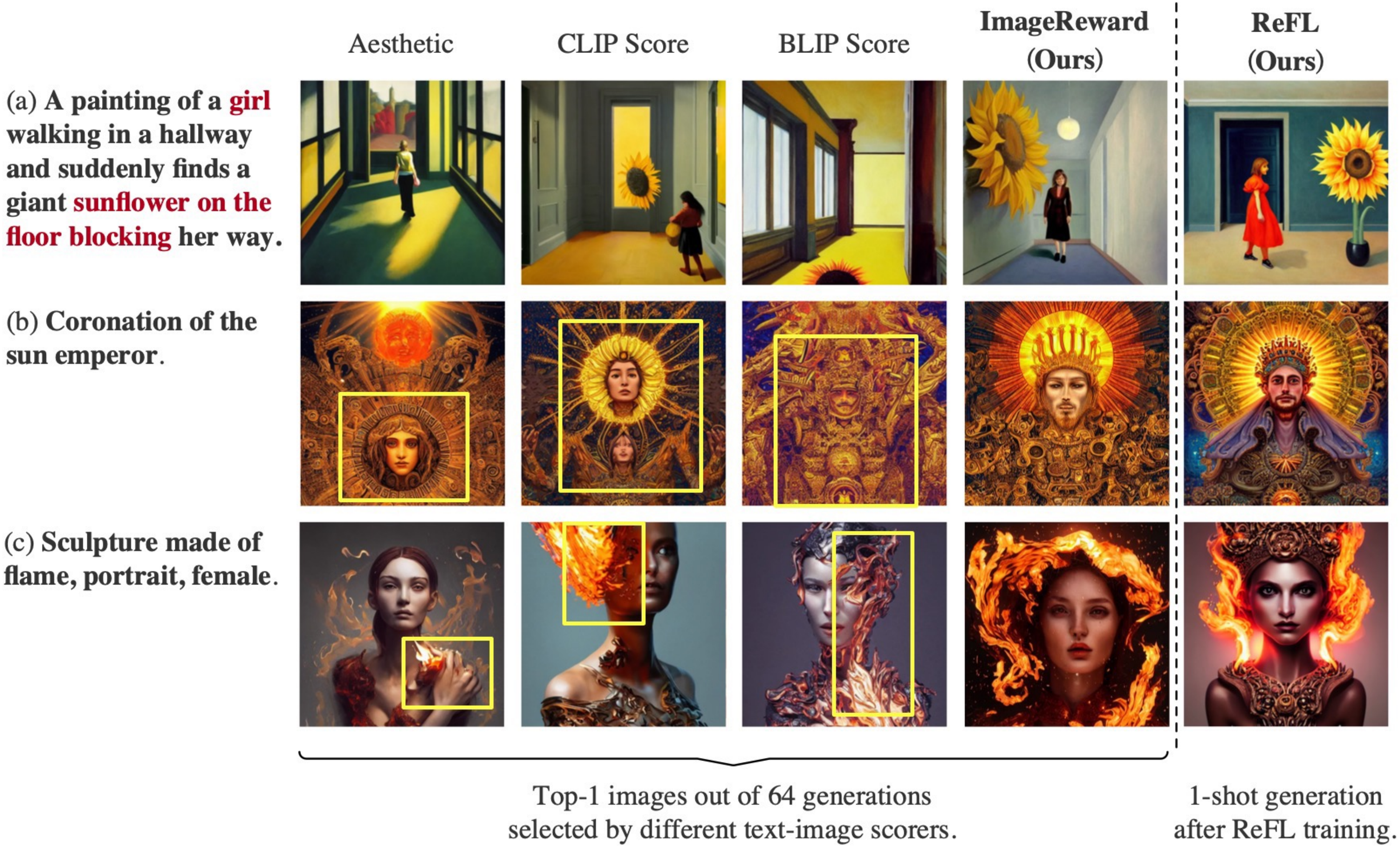
(c) Sculpture made of flame, portrait, female.



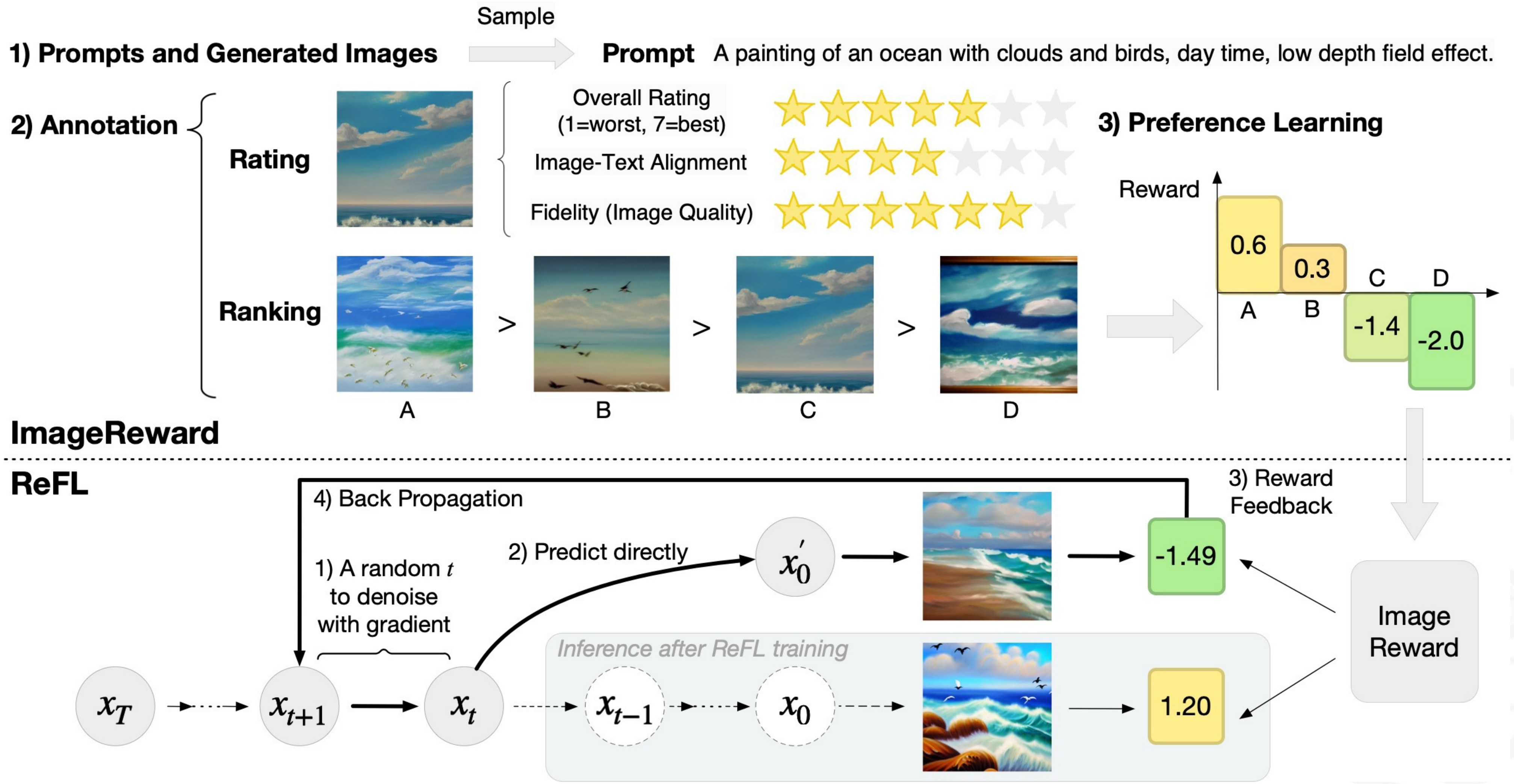
Images are generated by Stable Diffusion.

- Text-image Alignment
- Body Problem
- Human Aesthetic
- Toxicity and Biases

Align with Human Preference



Overview of ImageReward and ReFL



Outline

ImageRewardDB: Sample Collection

Source:

DiffusionDB (**1.8M** prompts, **14M** images generated by Stable Diffusion).

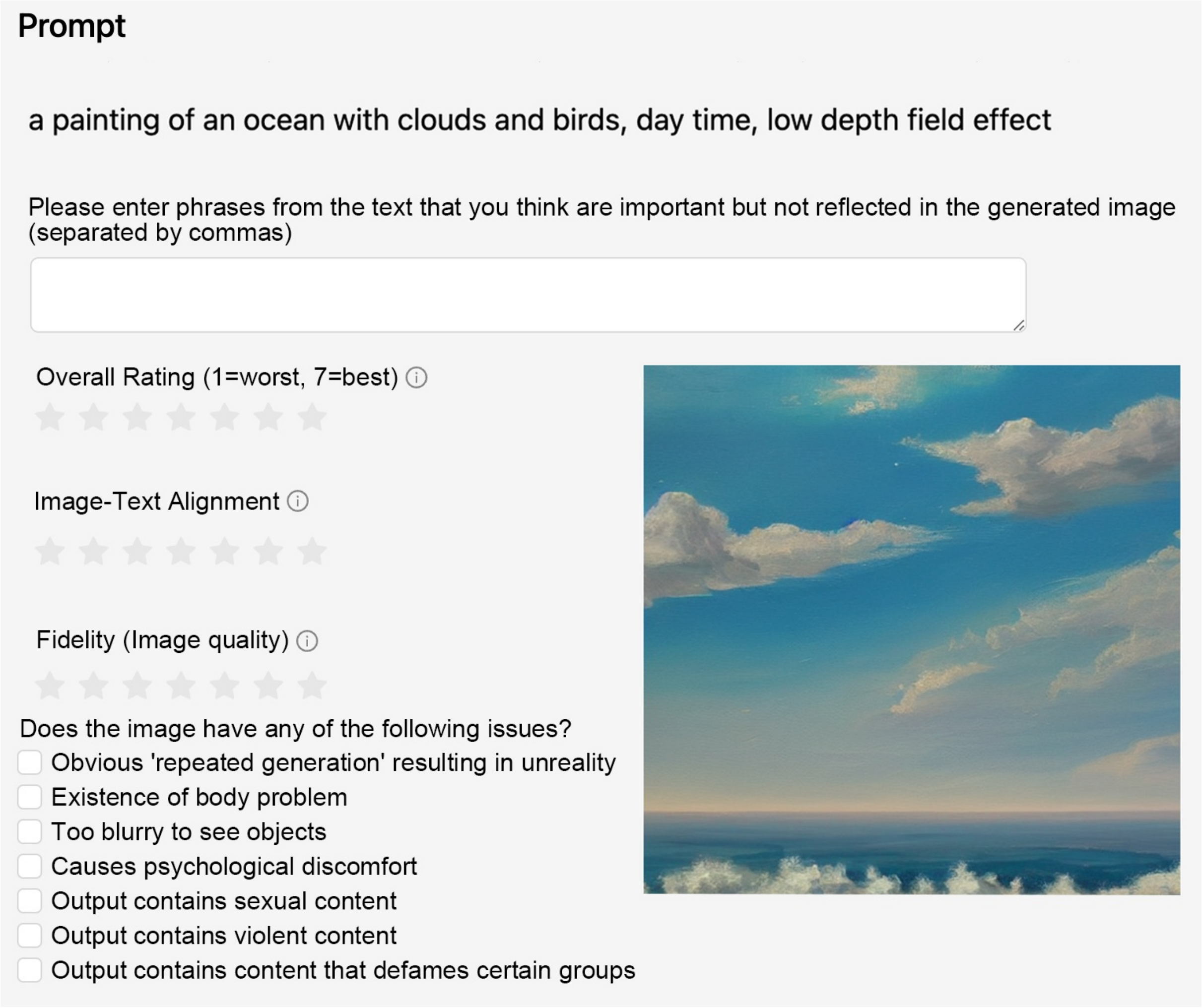
Prompt Selection Method:

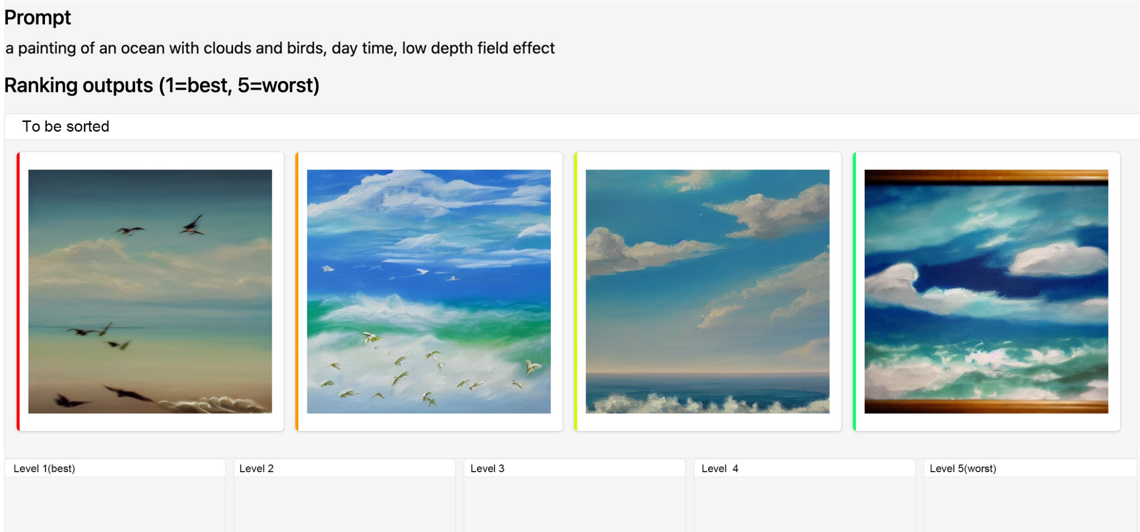
Graph-based algorithm with language model-based **prompt similarity**.

Prompt Selection Result:

10,000 candidate prompts, each accompanied by **4 to 9** sampled images, resulting in **177,304** candidate pairs for labeling.

ImageRewardDB: Annotation Pipeline





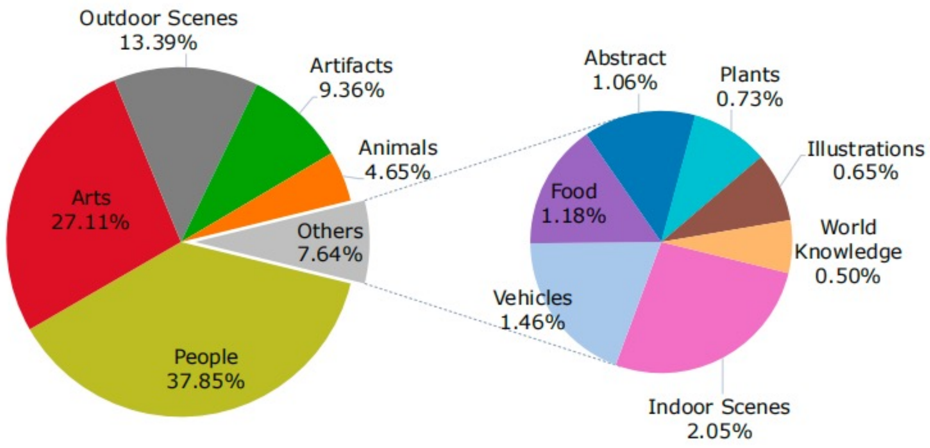
- **Prompt Annotation**: Categorizing prompts and identifying problematic ones.
- **Text-Image Rating**: Images are rated based on alignment, fidelity, and harmlessness.
- **Image Ranking**: Rank the images in order of preference.

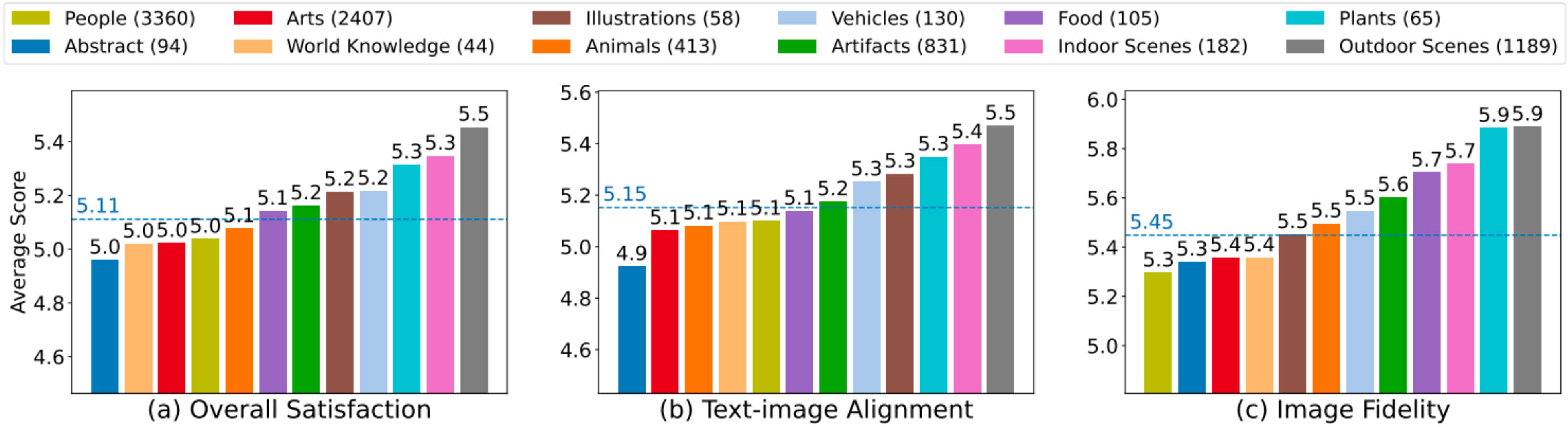
ImageRewardDB: Annotation Management

- **Annotation document:**
  - Criteria for rating/ranking.
  - Explanation for alignment/fidelity/harmlessness.
  - Trade-offs for potential contradictions in the ranking.
- **Annotators:**
  - Collaboration with a professional data annotation company.
  - Trained using annotation documents.
  - Quality inspectors: Double-check each annotation.

ImageRewardDB: Dataset Analysis

8,878 prompts
136,892 pairs

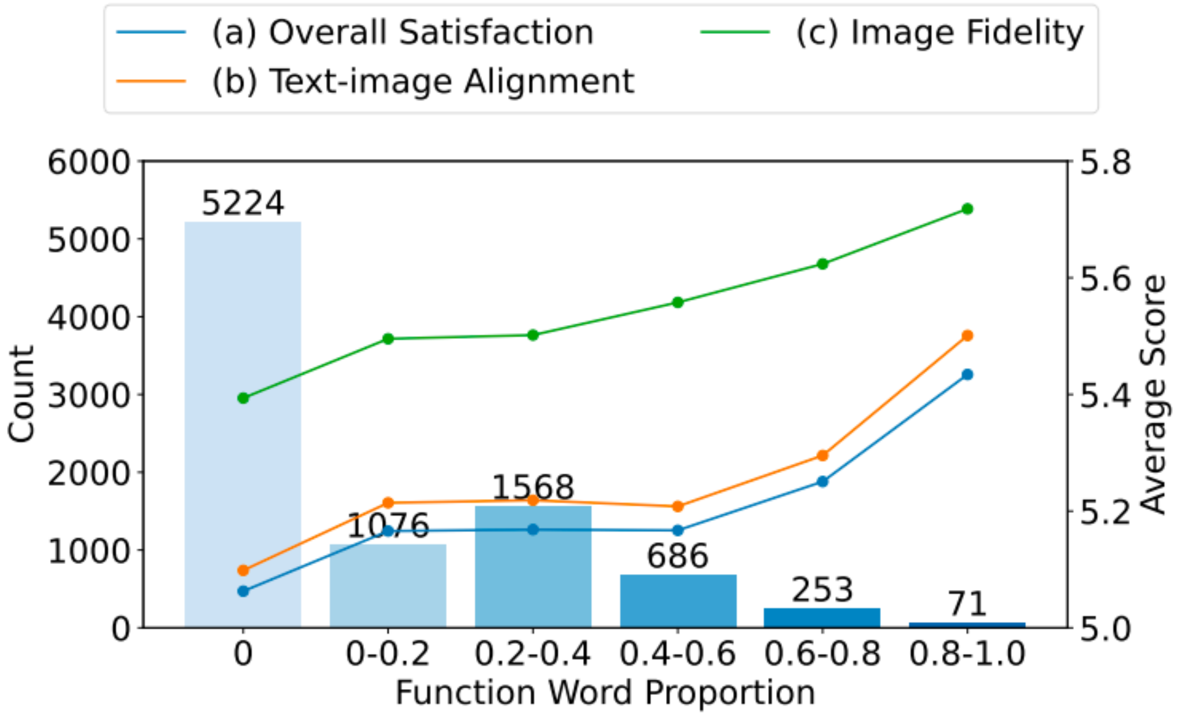




ImageRewardDB: Dataset Analysis

Function word

Many prompts not only describe the **content and style** but also contain some “function” words, like **"8k" and "highly detailed"**, trying to **improve the quality** of generated images.



Outline

ImageReward: Model and Data Settings

- Model Architecture:
  - **BLIP** backbone + **MLP** head
  - **ViT-L** for image encoder, **12-layers Transformer** for text encoder
- Dataset Settings:
  - Training set: **8k** prompts of annotation.
  - Test set: **466** prompts from annotators who have a higher agreement with researchers to consist for the model test.

ImageReward: RM Training

1. We have $k \in [4, 9]$ images ranked for the **same prompt T** (the best to the worst are denoted as $x_1, x_2, \ldots, x_k$) and get at most C_k^2 comparison pairs if no ties between two images.
2. For each comparison, if x_i **is better and** x_j **is worse**, the loss function can be formulated as:

$$\text{loss}(\theta) = -\mathbb{E}_{(T,x_i,x_j)\sim\mathcal{D}}[\log(\sigma(f_\theta(T,x_i) - f_\theta(T,x_j)))]$$

where $f_\theta(T, x)$ is a scalar value of preference model for prompt T and generated image x.

ImageReward: Training Settings

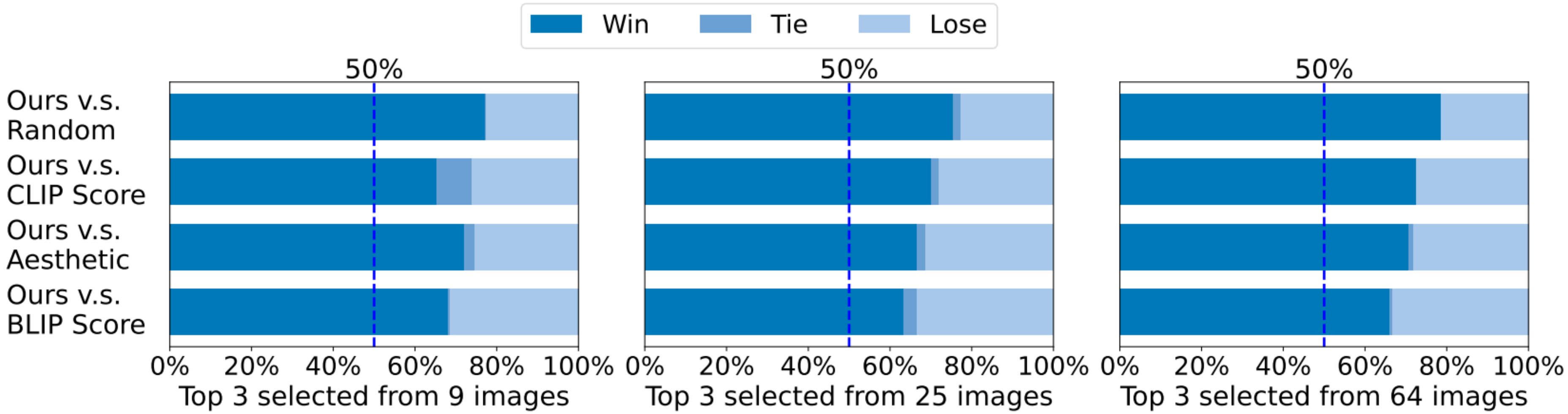
1. Initialize:
   - Load the **pre-trained** checkpoint of BLIP
   - **Initialize** MLP head according to $\mathrm{N}(0,\ 1/(d_{model} + 1))$
2. Hyperparameter:
   - Learning rate: initialize **1e-5**, decay with a **cosine** schedule
   - Fix rate: fixing **70%** of transformer layers
   - Batch size: **64**

ImageReward: Agreement Analysis

	researcher	annotator	annotator ensemble	CLIP Score	Aesthetic	BLIP Score	Ours
researcher	71.2% ± 11.1%	65.3% ± 8.5%	73.4% ± 6.2%	57.8% ± 3.6%	55.6% ± 3.1%	57.0% ± 3.0%	**64.5% ± 2.5%**
annotator	65.3% ± 8.5%	65.3% ± 5.6%	53.9% ± 5.8%	54.3% ± 3.2%	55.9% ± 3.1%	57.4% ± 2.7%	**65.3% ± 3.7%**
annotator ensemble	73.4% ± 6.2%	53.9% ± 5.8%	-	54.4% ± 21.1%	57.5% ± 15.9%	62.0% ± 16.1%	**70.5% ± 18.6%**

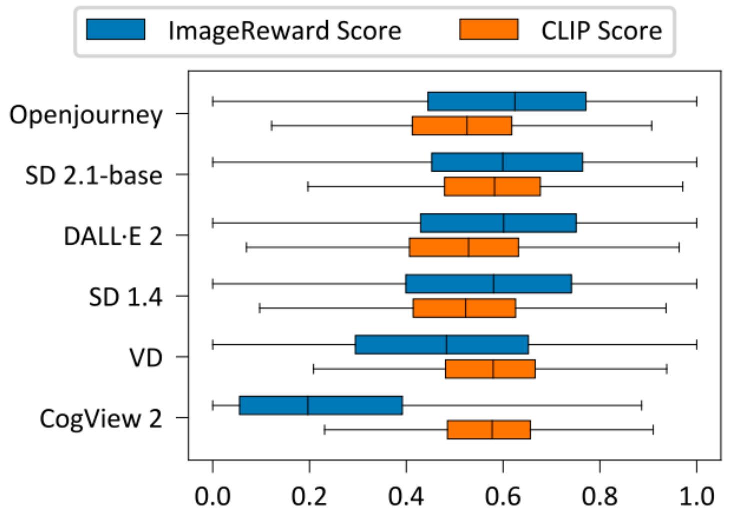
ImageReward: Preference Accuracy

Model	Preference Acc.	Recall @1	Recall @2	Recall @4	Filter @1	Filter @2	Filter @4
CLIP Score	54.82	27.22	48.52	78.17	29.65	51.75	76.82
Aesthetic Score	57.35	30.73	53.91	75.74	32.08	54.45	76.55
BLIP Score	57.76	30.73	50.67	77.63	33.42	56.33	80.59
ImageReward (Ours)	**65.14**	**39.62**	**63.07**	**90.84**	**49.06**	**70.89**	**88.95**



ImageReward: As Metric

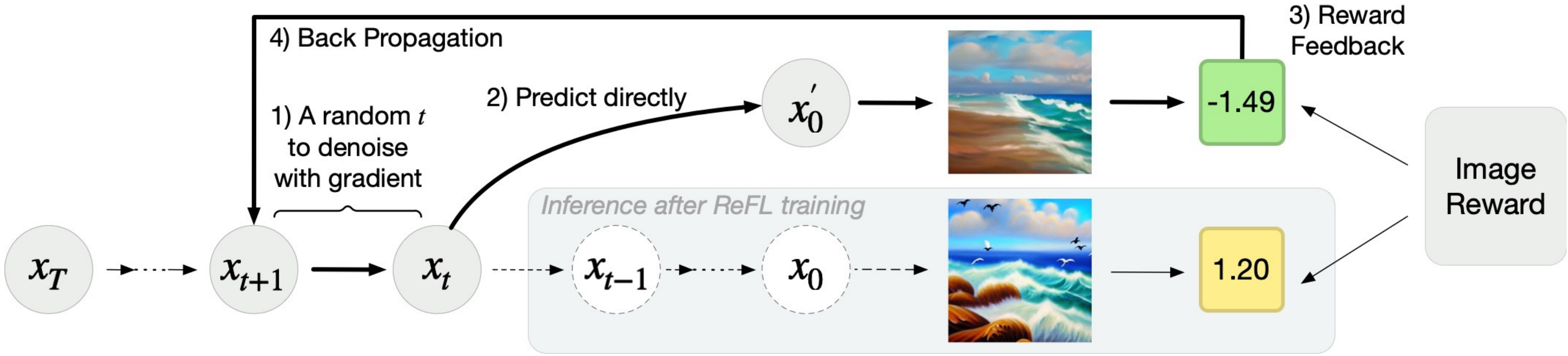
Dataset & Model	Real User Prompts						MS-COCO 2014			
	Human Eval.		ImageReward		CLIP		ImageReward		Zero-shot FID*	
	Rank	#Win	Rank	Score	Rank	Score	Rank	Score	Rank	Score
Openjourney	1	507	1	0.2614	2	0.2726	3	-0.0455	5	20.7
Stable Diffusion 2.1-base	2	463	2	0.2458	4	0.2683	2	0.1553	4	18.8
DALL-E 2	3	390	3	0.2114	3	0.2684	1	0.5387	1	10.9*
Stable Diffusion 1.4	4	362	4	0.1344	1	0.2763	4	-0.0857	2	17.9
Versatile Diffusion	5	340	5	-0.2470	5	0.2606	5	-0.5485	3	18.4
CogView 2	6	74	6	-1.2376	6	0.2044	6	-0.8510	6	26.2
Spearman ρ to Human Eval.	-		1.00		0.60		0.77		0.09	

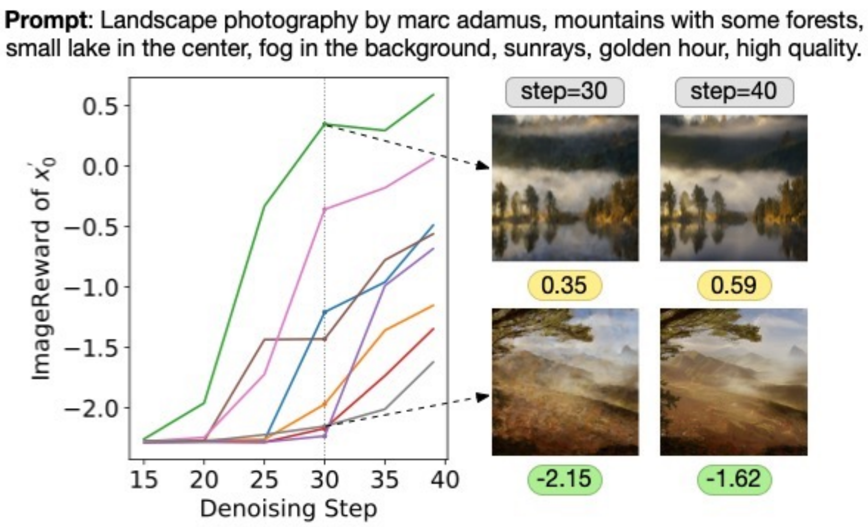


1. Better **Human Alignment** Across Models.
2. Better **Distinguishability** Across Models and Samples.

Outline

ReFL: Algorithm Design





$$\mathcal{L}_{reward} = \lambda \mathbb{E}_{y_i \sim \mathcal{Y}}(\phi(r(y_i, g_\theta(y_i))))$$

- θ: parameters of the LDM
- y_i: prompt, $\mathcal{Y}$: prompt set
- $g_\theta(y_i)$: generated image of LDM
- r: reward model
- ϕ: reward-to-loss map function
- λ: reward re-weight scale

ReFL: Algorithm Design

Algorithm 1 Reward Feedback Learning (ReFL) for LDMs

Dataset: Prompt set $\mathcal{Y} = \{y_1, y_2, ..., y_n\}$

Pre-training Dataset: Text-image pairs dataset $\mathcal{D} = \{(\text{txt}_1, \text{img}_1), ...(\text{txt}_n, \text{img}_n)\}$

Input: LDM with pre-trained parameters w_0, reward model r, reward-to-loss map function ϕ, LDM pre-training loss function ψ, reward re-weight scale λ

Initialization: The number of noise scheduler time steps T, and time step range for fine-tuning $[T_1, T_2]$

for $y_i \in \mathcal{Y}$ and $(\text{txt}_i, \text{img}_i) \in \mathcal{D}$ **do**

  $\mathcal{L}_{pre} \leftarrow \psi_{w_i}(\text{txt}_i, \text{img}_i)$

  $w_i \leftarrow w_i$ // Update LDM_{w_i} using Pre-training Loss

  $t \leftarrow rand(T_1, T_2)$ // Pick a random time step $t \in [T_1, T_2]$

  $x_T \sim \mathcal{N}(0, I)$ // Sample noise as latent

  for $j = T, ..., t + 1$ **do**

    no grad: $x_{j-1} \leftarrow \text{LDM}_{w_i}\{x_j\}$

  end for

  with grad: $x_{t-1} \leftarrow \text{LDM}_{w_i}\{x_t\}$

  $x_0 \leftarrow x_{t-1}$ // Predict the original latent by noise scheduler

  $z_i \leftarrow x_0$ // From latent to image

  $\mathcal{L}_{reward} \leftarrow \lambda\phi(r(y_i, z_i))$ // ReFL loss

  $w_{i+1} \leftarrow w_i$ // Update LDM_{w_i} using ReFL loss

end for

ReFL: Human Evaluation

- Fine-tuning Settings:
  - **20,000** samples from DiffusionDB
  - **the same** training settings (the same learning rate and batch size)
- Evaluation Dataset:
  - **466** real user prompts from DiffusionDB
  - **90** designed challenging prompts from multi-task benchmark
- Human evaluation:
  - **sorting** multiple images under a prompt
  - Stable Diffusion v1.4, PNDM noise scheduler and default classifier free guidance scale of 7.5 for inference.

ReFL: Human Evaluation

Methods	Real User Prompts		MT Bench [40]	
	#Win	WinRate	#Win	WinRate
SD v1.4 (baseline) [45]	1315	-	718	-
Dataset Filtering [61]	1394	55.17	735	51.72
Reward Weighted [23]	1075	39.52	585	43.33
RAFT [13] (iter=1)	1341	49.86	578	42.31
RAFT (iter=2)	753	30.85	452	33.02
RAFT (iter=3)	398	20.97	355	26.19
ReFL (Ours)	**1508**	**58.79**	**808**	**58.49**

Table 4: Human evaluation on different LDM optimization methods. ReFL performs the best with regard to total win count and WinRate against SD v1.4 baseline.

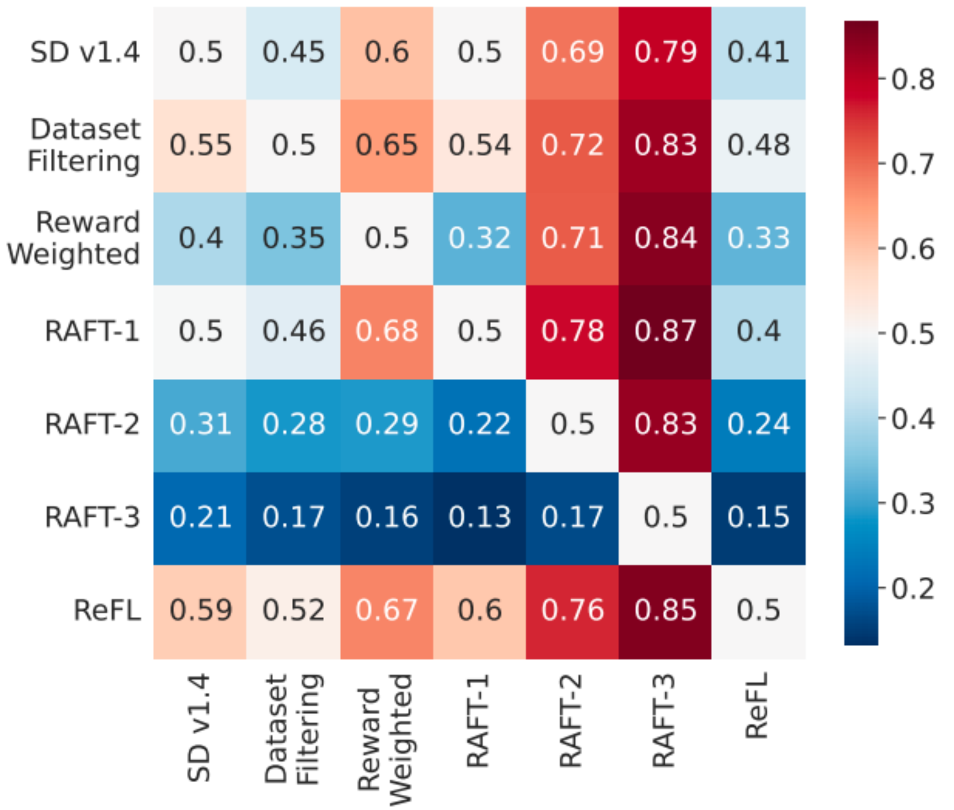
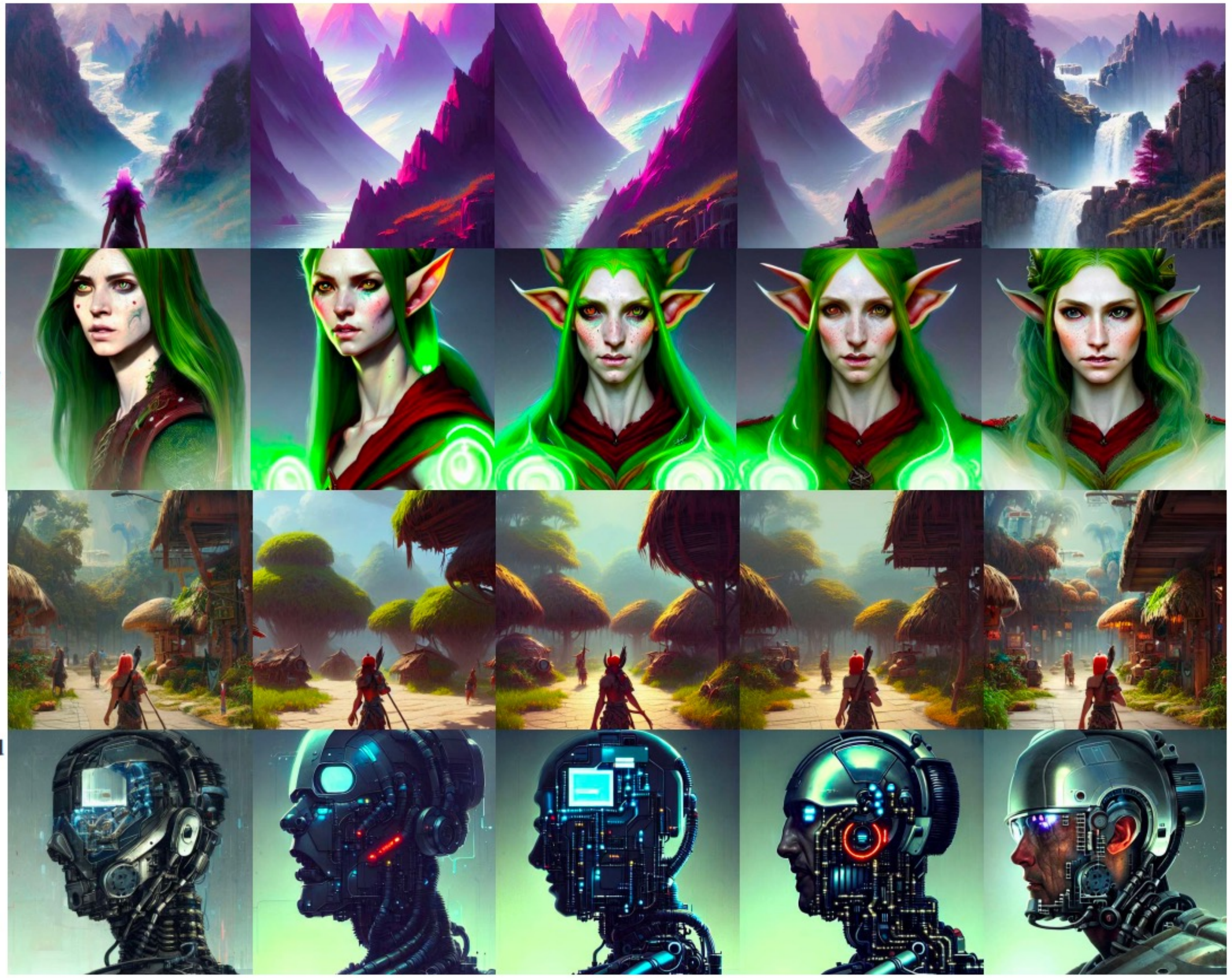


Figure 6: Win rates between all methods.

ReFL: Qualitative Comparison

Original | Dataset Filtering | Reward Weighted | RAFT | **ReFL (Ours)**



Mountains range with waterfall, purple haze, *art by greg rutkowski and magali villeneuve, artstation.*

Portrait of a female elf warlock, long pointy ears, glowing green eyes, bushy red hair and freckles + *medieval setting, detailed face, highly detailed, digital painting, artstation.*

An concept art illustration, photorealistic tribal people working, fantasy street with huts, large insect and plant biomes, *ultra realistic, style by wlop.*

A half - masked rugged laboratory engineer man with cybernetic enhancements as seen from a distance, *scifi character portrait by greg rutkowski, esuthio, craig mullins.*

ImageReward: Summary

ImageReward Page: https://github.com/THUDM/ImageReward

Python Package: https://pypi.org/project/image-reward/

ImageRewardDB: https://huggingface.co/datasets/THUDM/ImageRewardDB