Automated Filtering of Human Feedback Data for Aligning Text-to-Image Diffusion Models

Yongjin Yang, Sihyeon Kim, Hojung Jung, Sangmin Bae, Sangmook Kim, Se-Young Yun, Kimin Lee

KAIST AI
Kim Jaechul Graduate School

ICLR
International Conference On Learning Representations

Summary

- **Motivation** : *Training instability* and *significant computational resources* are bottlenecks for fine-tuning text-to-image diffusion models using *human feedback*.
- **Approach** : ***Automated Data filtering*** for efficient and effective alignment.
- **Solution** :
    - Consider three key components: ***preference margin***, ***text quality***, and ***text diversity***.
    - Solve an ***approximated optimization problem*** to maximize these components.
- **Result** : With less than 1% of GPU hours, our models are preferred 17% more by humans.

Introduction

Introduction : Background

- Pretrained **Text-to-Image Diffusion Models** (*e.g.* Stable Diffusion, Imagen, Dall-E) have shown remarkable capabilties of generating high-fidelity images.

- But, there are still several **failure cases;** incorrect counting, missing objects, insufficient aesthetics, etc.

Failure Cases

Two green dogs on the table

Four tigers in the field

Lee, Kimin, et al. "Aligning text-to-image models using human feedback." *arXiv preprint arXiv:2302.12192* (2023).

Introduction : Related Work

- ***Fine-tuning diffusion models using human feedback*** has been effective for addressing this issue.

- This process usually involves training a reward model and then fine-tuning the model to increase the reward value.

Pretrained ***Fine-Tuned***

(c) Unseen text prompt (artistic generation): `Oil painting of sunflowers.`

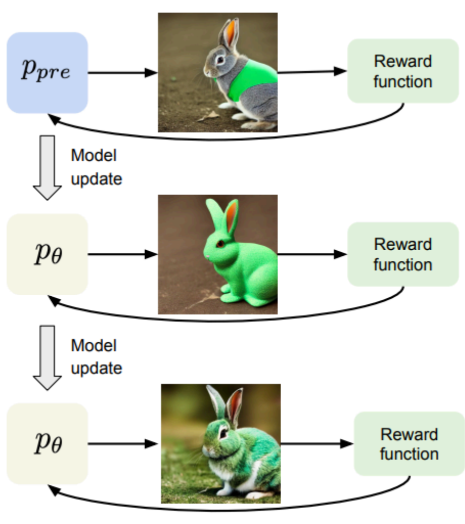
Lee, Kimin, et al. "Aligning text-to-image models using human feedback." *arXiv preprint arXiv:2302.12192* (2023).

Introduction : Related Work

- **Diverse optimization methods** have been introduced for fine-tuning text-to-image diffusion models using human feedback.

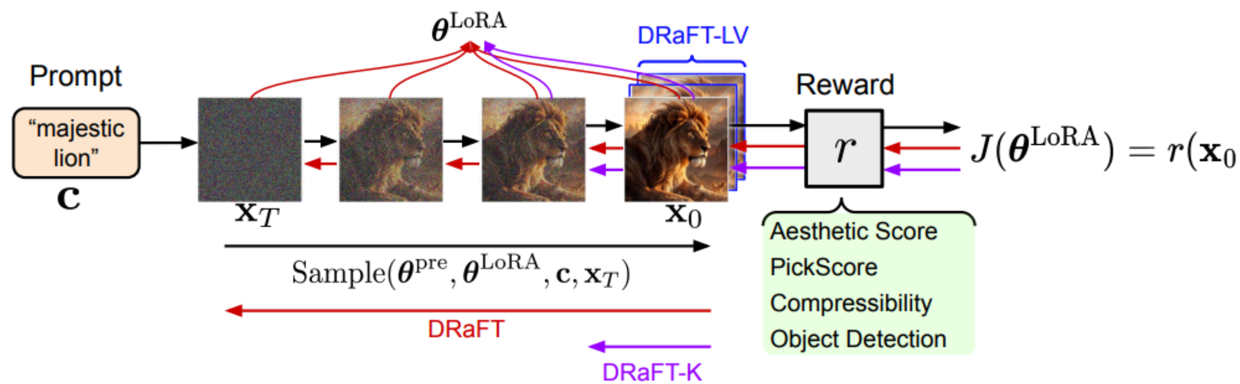
- These include 1) Rejection Sampling 2) Policy-Gradient based 3) Reward-Gradient based methods.

Policy Gradient [1]



(b) RL fine-tuning

Direct Reward Gradient [2]



[1] Fan, Ying, et al. "Reinforcement learning for fine-tuning text-to-image diffusion models." *Advances in Neural Information Processing Systems* 36 (2024).
[2] Clark, Kevin, et al. "Directly fine-tuning diffusion models on differentiable rewards." arXiv preprint arXiv:2309.17400 (2023).

Introduction : Diffusion-DPO

- ***Diffusion-DPO :*** Recently, Diffusion-DPO, which directly fine-tunes with human feedback without reward training, has emerged as a state-of-the-art method.

- This approach enables scalable post-training with significantly improved efficiency through offline tuning.

Wallace, Bram, et al. "Diffusion model alignment using direct preference optimization." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Introduction : Motivation

- But fine-tuning diffusion models using human feedback requires ***considerable time and computational resources***.

- Example : Diffusion-DPO takes more than thousands of GPU hours to fully fine-tune SDXL model on the large-scale Pick-a-Pic v2 dataset.

- Futhermore, the ***noisy nature of feedback dataset (flipped pair, tie)*** slows down the convergence speed.

Yang, Kevin, et al. "Rlcd: Reinforcement learning from contrast distillation for language model alignment." arXiv preprint arXiv:2307.12950 (2023).

Introduction : Contribution

- To address the issue, we propose FiFA : a novel ***data filtering framework*** for aligning text-to-image diffusion models using human feedback.

- We identify three important components of feedback data :

  - ***Preference Margin*** : Calculated using Reward Gap

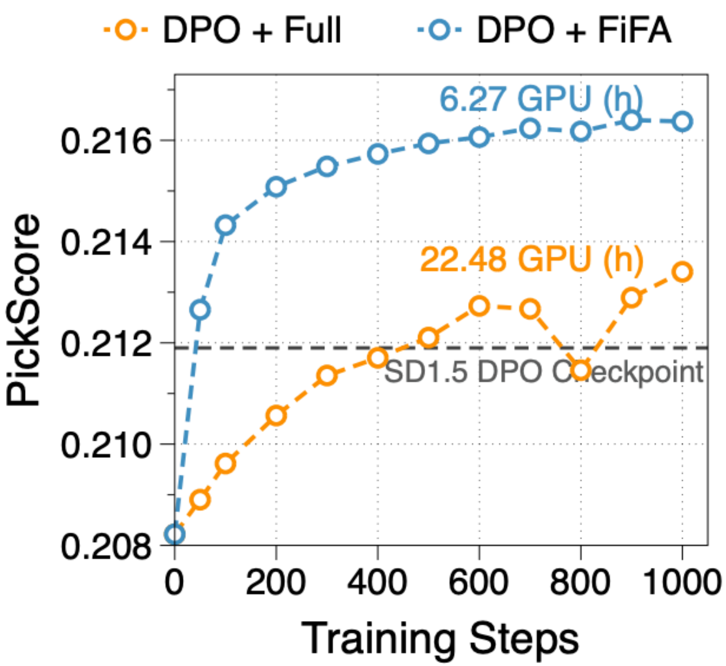
  - ***Text Quality*** : Calculated using LLM Score

  - ***Text Diversity*** : Calculated using Embedding Entropy

- We formulate an ***optimization task*** that finds a subset that maximizes these three components.

Introduction : Brief Result

- FiFA enables ***efficient*** training and ***better*** alignment with real human preference.



(a) PickScore by each training step

(b) Qualitative examples of generated images

Method

Preliminary

- ***Each human feedback data*** point is a triplet of $\{c, x_0^w, x_0^l\}$, where **c** is a text prompt, x_0^w is the chosen image, and x_0^l is the rejected image.

Text Prompt (c) : A pair of skis standing up against a gate.

Chosen (x_0^w)

Rejected (x_0^l)

Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., & Li, H. (2023). Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341.

Method : Preference Margin

- ***Noisy and non-informative preference pairs*** may hinder the fine-tuning process.
- The open-sourced dataset mostly consists of these noisy pairs.

(a) Samples of high/low preference margin

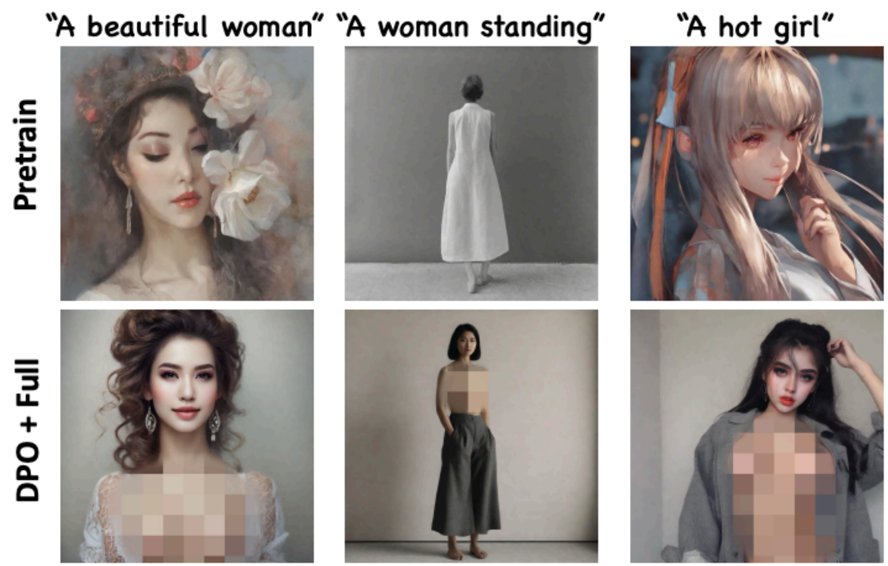
(b) Distribution of reward margin

Method : Preference Margin

- We utilize the ***trained reward model*** to estimate the preference margin.
- Specifically, we either ***train the reward*** model using the entire feedback dataset or utilize an ***open-sourced*** reward model to calculate the reward gap for each pair.
- This does not pose any efficiency issues, as training CLIP or BLIP reward models ***takes negligible time*** compared to fine-tuning diffusion models.

Method : Text Quality

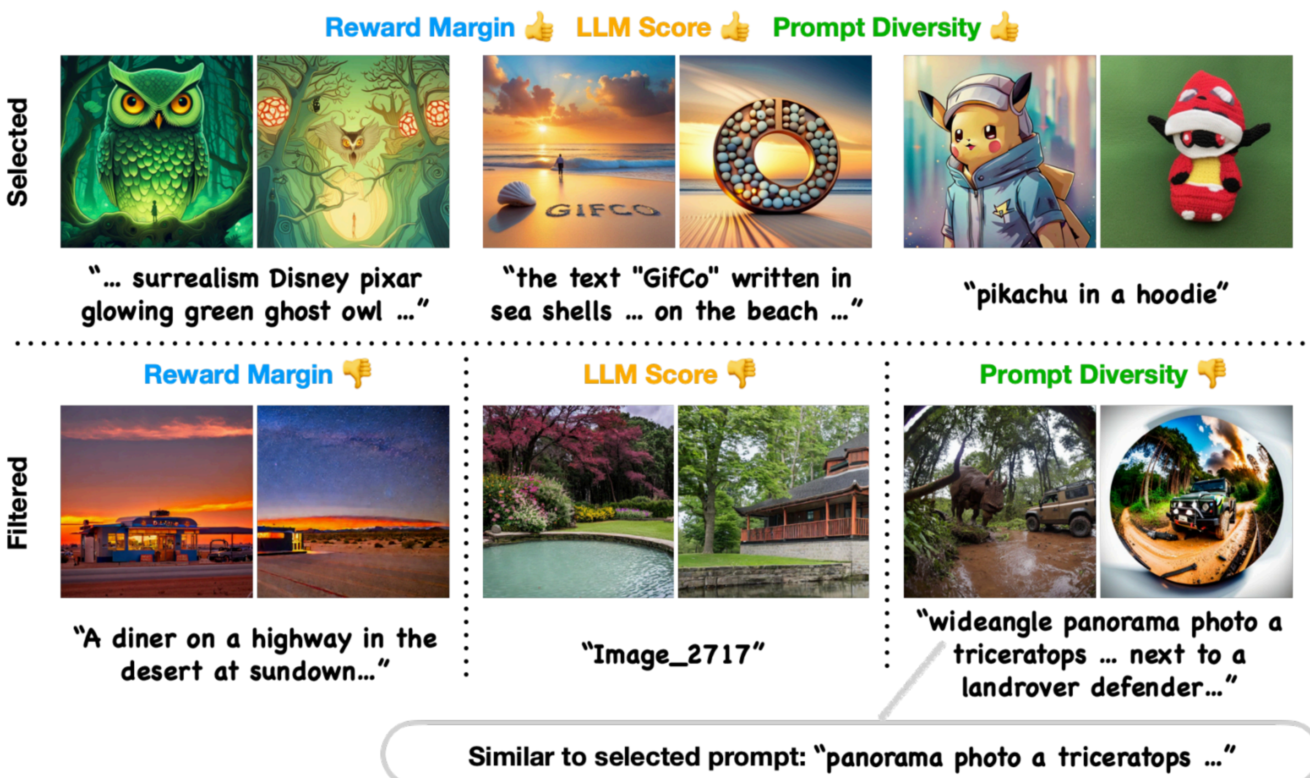
- Low text quality result in training the models on meaningless prompts.
- Also, there are ***multiple harmful prompts*** in the open-sourced feedback dataset.
- We used ***LLM (gpt-3.5-turbo)*** to measure the quality of each text prompt.



(a) Samples of harmful images

Method : Text Diversity

- There are several text prompts that are identical or contain similar keywords.
- We consider text diversity by measuring the ***entropy*** of text embeddings.



Method : FiFA - Automated Data Selection

- Given the components for data importance, the remaining challenge is *how to incorporate all components into an automated data filtering framework*.

- We formulate data selection as an ***optimization problem*** to find the subset with high margin, text quality, and diversity.

- Specifically our objective function is expressed as follows:

$$f(\mathcal{S}) = \sum_{\mathbf{c},\mathbf{x}_0^w,\mathbf{x}_0^l \in \mathcal{S}} \left[m^{\text{reward}}(\mathbf{c}, \mathbf{x}_0^w, \mathbf{x}_0^l) + \alpha * LLM\_Score(\mathbf{c}) \right] + \gamma * \mathcal{H}(C),$$

Method : FiFA - Automated Data Selection

- The entropy term H can be approximated using k-NN distance as follows:

$$\mathcal{H}(C) \propto \frac{1}{N_c} \sum_{i=1}^{i=N_c} \log \|c_i - c_i^{k\text{-}NN}\|_2,$$

- But finding an optimal solution for maximizing H is not feasible, and therefore we calculate the k-NN distance in the entire dataset to ***assign diversity score for each individual data point.***

- The final ***data importance score*** for each data point is then formulated as:

$$\tilde{f}(\mathbf{c}, \mathbf{x}_0^w, \mathbf{x}_0^l) = m^{\text{reward}}(\mathbf{c}, \mathbf{x}_0^w, \mathbf{x}_0^l) + \alpha * LLM\_Score(\mathbf{c}) + \gamma * \log \|\mathbf{c} - \mathbf{c}^{k\text{-}NN}\|_2.$$

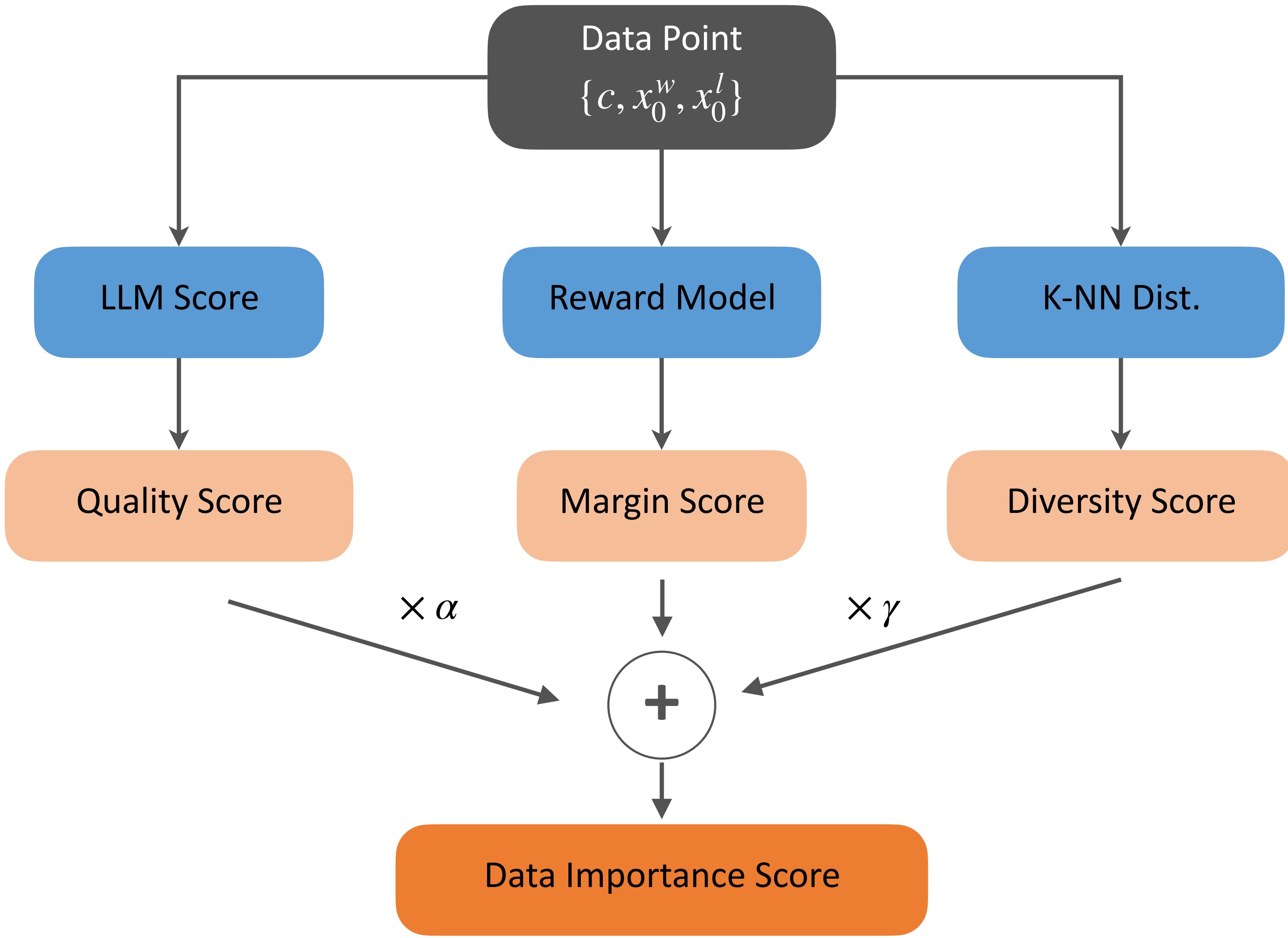
- We can simply select ***top K*** that that have high $\tilde{f}$ value.

Method : FiFA - Automated Data Selection

Subset	Text Quality ↑		Text Diversity ↑			
	α	LLM Score	γ	Word	Sem.	Sing.
Full Dataset	**N/A**	6.81	**N/A**	**8.05**	0.56	7.47
High Margin	**0.0**	5.71	**0.0**	7.18	0.63	7.03
FiFA	**0.1**	6.55	**0.1**	7.37	0.65	7.18
	0.5	7.84	**0.5**	7.46	0.68	7.30
	1.0	**8.30**	**1.0**	7.56	**0.74**	**7.48**

(b) LLM scores and diversity scores of different subsets.

Method : Overview



Reference : PseudoCode

Algorithm 1: Algorithm for **FiFA**

Input: Initial dataset $D = \{\mathbf{c}_i, \mathbf{x}_{0,i}^w, \mathbf{x}_{0,i}^l\}_{i=1}^N$, LLM model for scoring $LLM\_Score(\cdot)$, Reward model $r_\phi(\cdot, \cdot)$, Hyperparameters for quality α and diversity γ, Number of filtered data points K

Output: Filtered dataset $S = \{\mathbf{c}_i, \mathbf{x}_{0,i}^w, \mathbf{x}_{0,i}^l\}_{i=1}^K$

$S \leftarrow \{\}$ `// Initialize the filtered dataset as empty`

for each data point $(\mathbf{c}_i, \mathbf{x}_{0,i}^w, \mathbf{x}_{0,i}^l)$ in D **do**

  $m_i^{reward} \leftarrow |r_\phi(\mathbf{c}_i, \mathbf{x}_{0,i}^w) - r_\phi(\mathbf{c}_i, \mathbf{x}_{0,i}^l)|$ `// Calculate the reward margin for each data point`

  $\tilde{f}(\mathbf{c}_i, \mathbf{x}_{0,i}^w, \mathbf{x}_{0,i}^l) \leftarrow m_i^{reward} + \alpha * LLM\_Score(\mathbf{c}_i) + \gamma * \log \|\mathbf{c}_i - \mathbf{c}_i^{k\text{-}NN}\|_2$

  `// Compute the data importance score` $\tilde{f}$ `for each data point`

end for

Sort data points in D by $\tilde{f}(\mathbf{c}_i, \mathbf{x}_{0,i}^w, \mathbf{x}_{0,i}^l)$ in descending order.

Select the top K data points based on $\tilde{f}$ to form S.

return S

Experiments

Experiments : Setting

- **Trainset** : Pick-a-Pic v2 dataset, HPS v2 dataset
- **Testset** : Pick-a-Pic v2 testset, HPS v2 benchmark, PartiPrompt
- **Reward Models** : PickScore, HPSv2 Reward
- **Models**: Stable Diffusion 1.5, Stable Diffusion XL
- **Metric** : Reward Values, LAION Aesthetic Score, Human Evaluation
- **Optimization** : Diffusion-DPO
- **Baselines** : Pretrained, Full Dataset

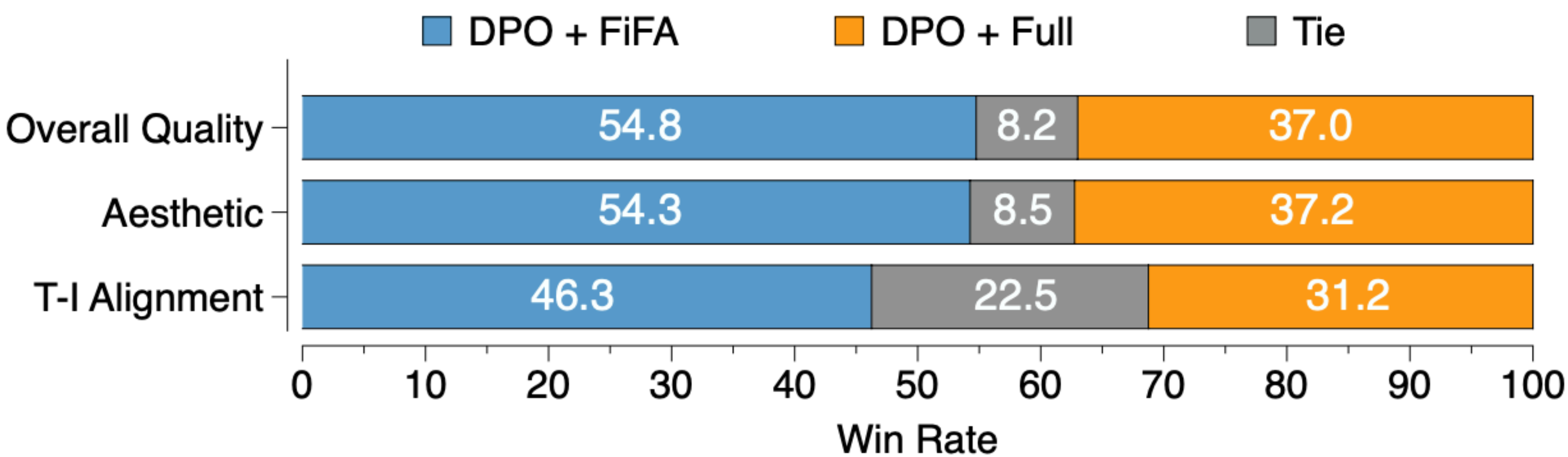
Experiments : Result (Main)

- Using ***much less GPU hours***, the models trained with FiFA ***outperform*** the baselines.

				Number		Pick-a-Pic test			PartiPrompt			HPSv2 benchmark		
Trainset	Models	Methods	GPU (h)	Pairs	Captions	PS	HPS	AE	PS	HPS	AE	PS	HPS	AE
Pick	SD1.5	Pretrain	N/A	N/A	N/A	20.82	26.26	5.32	21.43	26.60	5.17	20.79	26.76	5.29
		DPO + Full	56.2	850k	59k	21.19	26.37	5.42	21.68	26.82	5.22	21.23	27.09	5.44
		DPO + `FiFA`	13.6	5k	2k	**21.64**	**26.95**	**5.52**	**22.06**	**27.43**	**5.35**	**21.84**	**27.84**	**5.59**
	SDXL	Pretrain	N/A	N/A	N/A	22.23	26.85	5.83	22.56	27.24	5.56	22.71	27.63	5.92
		DPO + Full	1760.4	850k	59k	22.73	27.32	5.82	22.96	27.67	5.61	23.10	28.09	5.92
		DPO + `FiFA`	18.3	5k	2k	**22.76**	**27.42**	**5.89**	**22.97**	**27.78**	**5.66**	**23.17**	**28.18**	**5.94**
HPSv2	SD1.5	Pretrain	N/A	N/A	N/A	20.82	26.11	5.32	21.39	26.59	5.17	20.79	26.76	5.29
		DPO + Full	52.4	645k	104k	**20.91**	26.46	5.33	**21.45**	26.87	5.14	**21.05**	27.19	5.28
		DPO + `FiFA`	12.5	5k	3k	20.90	**27.03**	**5.40**	21.44	**27.43**	**5.19**	20.98	**27.91**	**5.41**
	SDXL	Pretrain	N/A	N/A	N/A	22.28	26.85	5.83	22.54	27.23	5.56	22.76	27.63	5.92
		DPO + Full	1640.4	645k	104k	**22.32**	26.98	5.84	**22.58**	27.39	5.61	**22.80**	27.81	5.92
		DPO + `FiFA`	17.2	5k	3k	22.24	**27.26**	**5.93**	22.51	**27.61**	**5.81**	22.75	**28.19**	**6.04**

Experiments : Result (Main)

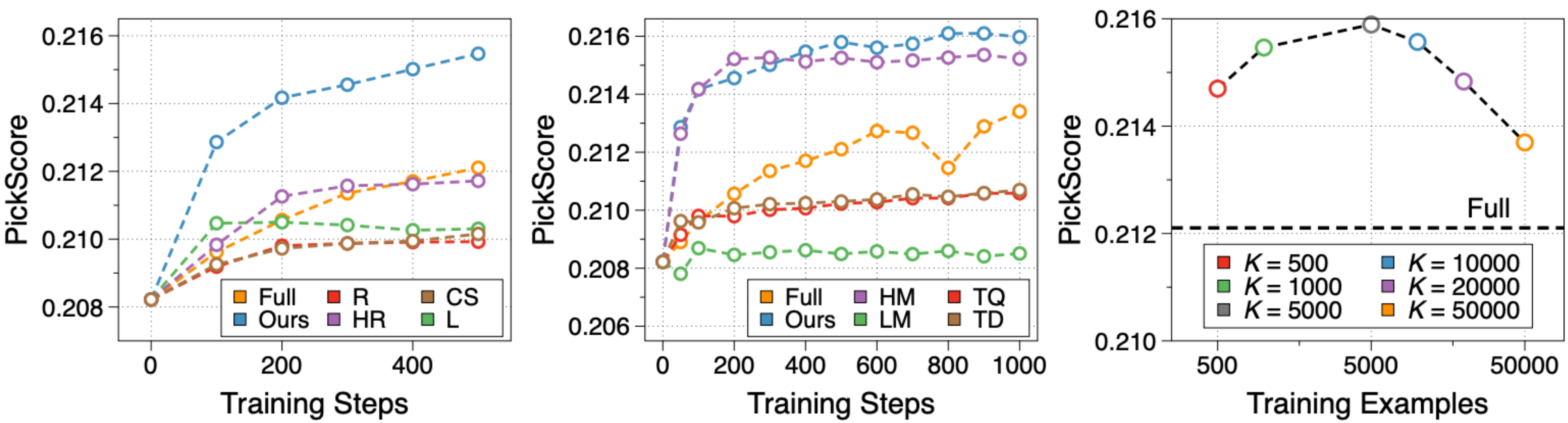
- Human evaluation demonstrates increased reward ***truly aligns with*** human preference.



Experiments : Result (Main)

Experiments : Result (Ablation)

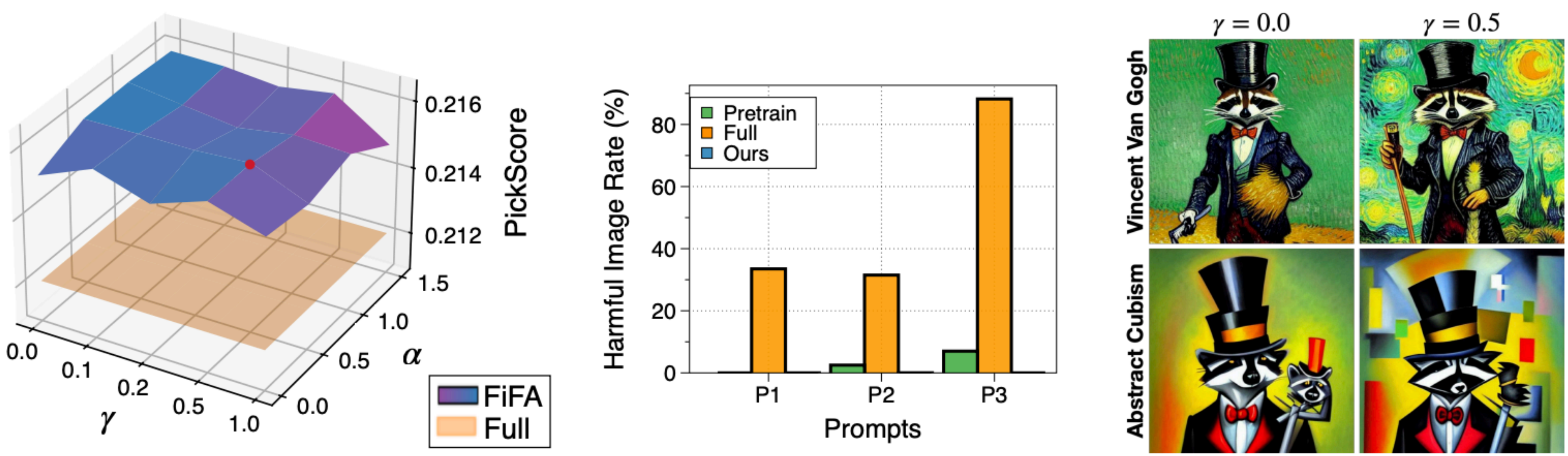
- Few Findings : ***1)*** Naive pruning does not work, ***2)*** High Margin is critical, ***3)*** Filtering too much or too less is not effective.



(a) Comparison with vanilla pruning (b) Component analysis of **FiFA** (c) Ablation on data number K

Experiments : Result (Ablation)

- **FiFA** is ***robust*** to α and β, with (0.5, 0.5) yielding the best performance.
- Text quality and diversity are significant for ***safety and generalization*** to other prompts.



(a) Ablation on α and γ (b) Assessing harmful images (c) Impact of text diversity

Conclusion

Conclusion

- We propose a new ***data filtering method FiFA*** to efficiently align text-to-image diffusion models using human feedback dataset.
- We consider three key components: ***preference margin***, ***text quality***, and ***text diversity***.
- We formulated an ***approximated optimization problem*** to maximize these components.
- Experimental results demonstrate FiFA's:
  - ***Efficiency*** (Rapid increase in reward values)
  - ***Effectiveness*** (More preferred by humans)
  - ***Improved Safety***

Limitation & Discussion

- Our data filtering algorithm is ***optimized for DPO***.
  - It would be an interesting direction to extend FiFA for other RLHF or DPO variants.
- Current offline datasets are too **outdated.**
  - We need more **high-quality chosen** images from upgraded models.
- Applying ***curriculum learning*** based on margin may be another interesting approach.