A LIMITATIONS

Our method requires multiple passes through the diffusion model to optimize a given prompt, which incurs a modest amount of search costs. One promising solution is to use DPO to generate free paired data for RLHF (e.g. Promptist), which we leave for future work to explore. moreover, while DPO improves the faithfulness of the generated image, the performance is upper-bounded by the limitations of the underlying text encoder itself. For example, the clip text encoder used in stable diffusion tends to discard spatial relationships in text, which cannot be resolved by any prompt optimization. These can be resolved by an orthogonal line of methods that augment the diffusion model with a powerful LLM Lian et al. (2023); Liu et al. (2022); Feng et al. (2022). Finally, the clip loss used in DPO might not always align with human evaluation. Automatic scoring metrics that better reflect human judgment, similar to the reward models used in instruction fine-tuning, can further aid the discovery of improved prompts.

B THE COMPLETE DPO ALGORITHM

Algorithm 1 DPO solver: Discrete Prompt Optimization Algorithm

Require: User Input s_{user}, diffusion model $G(\cdot)$, a loss function $\mathcal{L}(I, s)$, learning rate lr.
Ensure: An optimized prompt s^*.
// Building Search Space
Query ChatGPT to generate a word-substitutes dictionary for s_{user}
Initialize Gumbel parameter α accordingly.
// Gradient Prompt Optimization
for i from 1 to max_iter **do**
  Sample $p(w; \alpha)$ for each word w from Gumbel Softmax.
  Compute mixed embedding: $\tilde{e}(\alpha) = \sum_{i=1}^{|\mathcal{V}|} p(w = i; \alpha) * e_i$
  Compute text gradient: $g_s = \nabla_\alpha \mathcal{L}(G(\tilde{e}(\alpha)), s)$
  Update Gumbel Parameter: $\alpha_i = \alpha_i - lr * g_{s_{user}}$
end for
// Evolutionary Sampling
Generate initial population $\mathcal{P} \sim Gumbel(\alpha)$
Update population with $\mathcal{P}^* = EvoSearch(\mathcal{P})$
$s^* = \arg\max_s(\mathcal{G}(s \in \mathcal{P}^*), s_{user})$

C IMPLEMENTATION DETAILS

C.1 HYPERPARAMETERS

This section details the hyperparameter choices for our experiments. We use the same set of hyperparameters for all datasets and tasks (prompt improvement and adversarial attack), unless otherwise specified.

Model We use Stable Diffusion v1-4 with a DDIM sampler for all experiments in the main paper. The guidance scale and inference steps are set to 7.5 and 50 respectively (default). We also experimented with other versions, such as Stable Diffusion v2-1 (512 x 512 resolution) and v2 (786x786 resolution), and found that the results are similar across different versions. Although, we note that the high-resolution version of v2 tends to produce moderately better original images than v1-4 and v2-1 in terms of clip loss, possibly due to sharper images.

Shortcut Gradient We set $K = 1$, corresponding to a 1-step shortcut gradient. This minimizes the memory and runtime cost while empirically producing enough signal to guide the prompt optimization. Throughout the entire optimization episode, we progressively increase t from 15 to 25 via a fixed stepwise function. This corresponds to a coarse-to-fine learning curriculum. We note that the performance is only marginally affected by the choice of the upper and lower bound for t (e.g.

20-30, 10-40 all produce similar results), as long as it avoids values near 0 (diminishing gradient) and T (excessively noisy).

Gumbel softmax We use Gumbel Softmax with temperature 1. The learnable parameters are initialized to 1 for the original word (for positive prompts) and empty string (for negative prompts), and 0 otherwise. To encourage exploration. We bound the learnable parameters within 0 and 3 via hard clipping. The performance remains largely incentive to the choice of bound, as long as they are in a reasonable range (i.e. not excessively small or large).

Optimization We optimize DPO-Diff using RMSprop with a learning rate of 0.1 and momentum of 0.5 for 20 iterations. Each iteration will produce a single Gumbel Sample (batch size = 1) to compute the gradient, which will be clipped to $1/40$.

clip loss The specific clip loss used in our experiment is spherical clip loss, following an early online implementation of clip-guided diffusion ("crumb", 2022):

$$\text{spherical\_clip(x, y)} = 2 \cdot \left(\arcsin \frac{\|x - y\|_2}{2} \right)^2$$

Note that our method does not rely on this specific choice to function; We also experimented with other distance measures such as cos similarity on the clip embedding space, and found that they produced nearly identical prompts (and thus images).

Evolution Search We follow a traditional evolution search composed of four steps: initialize population, tournament, mutation, and crossover. The specific choice of hyperparameters is population size = 20, tournament = top 10, mutation with prob = 0.1 and size = 10, and crossover with size = 10. We run the evolutionary search for two iterations for both tasks, while we note that the prompt improvement task often covers much faster (within a single iteration).

C.2 Search space construction

We construct our Synonyms and Antonyms space by querying ChatGPT using the following prompts. Since ChatGPT sometimes makes mistakes by producing false synonyms or antonyms, we also adopt filtering by thresholding the cosine similarity between adversarial prompts and user prompts in the embedding space of T5 Raffel et al. (2020). This filtering is applied both during search space generation and candidate prompt sampling phase. The threshold is set to 0.9 for all datasets.

> Read the next paragraph. For each word, give 5 substitution words that do not change the meaning. Use the format of "A → B".

For Antonyms:

> Read the next paragraph. For each word, give 5 opposite words if it has any. Use the format of "A → B".

D More experimental settings

D.1 Dataset Collection

The prompts used in our paper are collected from three sources, DiffusionDB, COCO, and ChatGPT.

DiffusionDB DiffusionDB is a giant prompt database comprised of 2m highly diverse prompts for text-to-image generation. Since these prompts are web-crawled, they are highly noisy, often containing incomplete phrases, emojis, random characters, non-imagery prompts, etc (We refer the reader to its HuggingFace repo for an overview of the entire database.). Therefore, we filter prompts from DiffusionDB by (1). asking ChatGPT to determine whether the prompt is complete and describes an image, and (2) remove emoji-only prompts. We filter a total of 4,000 prompts from DiffusionDB

and use those prompts to generate images via Stable Diffusion. We sample 100 prompts with clip loss above 0.85 for prompt improvement, and 0.8 for adversarial attacks respectively. For ChatGPT, we found that it tends to produce prompts with much lower clip score compared with COCO and DiffusionDB. To ensure a sufficient amount of prompts from this source is included in the dataset, we lower the cutoff threshold to 0.82 when filtering its hard prompts for the prompt improvement task.

COCO We use the captions from the 2014 validation split of MS-COCO dataset as prompts. Similar to DiffusionDB, we filter 4000 prompts, and further sample 100 prompts with clip loss above 0.85 for prompt improvement, and 0.8 for adversarial attack respectively.

ChatGPT We also query ChatGPT for descriptions, as we found that it tends to produce more vivid and poetic descriptions compared with the former sources. We use a diverse set of instructions for this task. Below are a few example prompts we used to query ChatGPT for image descriptions.

> Generate N diverse sentences describing photoes/pictures/images
> Generate N diverse sentences describing images with length around 10
> Generate N diverse sentences describing images with length around 20
> Generate N diverse sentences describing images using simple words
> Generate N diverse sentences describing images using fancy words

Below are some example prompts returned by ChatGPT:

> A majestic waterfall cascades down a rocky cliff into a clear pool below, surrounded by lush greenery.
> The sun setting behind the mountains casting a warm orange glow over the tranquil lake.
> A pair of bright red, shiny high heels sit on a glossy wooden floor, with a glittering disco ball above.
> A farmer plowing a field with a tractor.
> The vivid orange and dark monarch butterfly was flapping through the atmosphere, alighting on a flower to sip nectar.

We empirically observe that ChatGPT produces prompts with low clip loss when used to generate images through Stable Diffusion on average, compared with DiffusionDB and COCO. Therefore, for filtering challenging prompts, we reduce the threshold from 0.85 to 0.82 to allow more prompts to be selected.

D.2 Human Evaluation

We ask 5 judges without ML background to evaluate the faithfulness of the generated images. For each prompt, we generate two images using the same seeds across different methods. To further avoid subjectiveness in evaluation, we provide the judgers an ordered list of important key concepts for each prompt, and ask them to find the winning prompt by comparing the hit rate. The ordered list of key concepts is provided by ChatGPT.

Since the 600 prompts used in the main experiments are filtered automatically via clip loss, they exhibit a certain level of false positive rate: some images are actually faithful. Therefore, we further filter out 100 most broken prompts to be evaluated by human judgers.

E Extra qualitative results

We include extra quantitative results of DPO in Figure 4 and Figure 5.

Figure 4: More images generated by user input and improved negative prompts using Stable Diffusion.

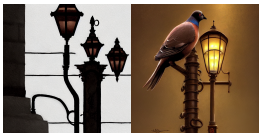
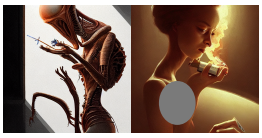
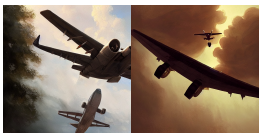
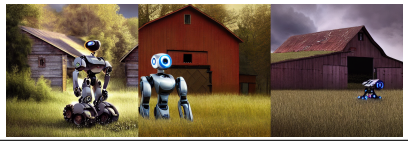
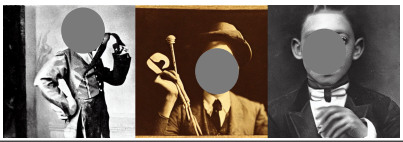
User Input	Promptist - Modifiers	DPO - Negative Prompt
The ash and dark pigeon was roosting on the lamppost, observing the environment. 	intricate, elegant, highly detailed, ..., illustration, by justin gerard and artgerm, 8 k 	fresh, shiny, hawk, overlooking, inside, Portrait, background, faded, unreal 
a photorealistic detailed image of epic ornate scenery 	by wlop, greg rutkowski, thomas kinkade, super detailed, 3 d, hdr, 4 k wallpaper 	broad, reality, minor, modest, Grains, broken, incorrect, replica 
alien caught smoking cigarettes in rented house 	intricate, elegant, highly detailed, ..., art by artgerm and greg rutkowski and, 8 k 	native, liberated, clear, dull, out, bought, road, Macro, Script, monochrome, rendered 
a spooky ghost in a graveyard by justin gerard and tony sart 	greg rutkowski, zabrocki, karlkka, ..., zenith view, zenith view, pincushion lens effect 	physical, house, aside, except, Grains, design, replica 
a plane flies through the air with fumes coming out the back 	Rephrase: a plane flies through the air with fumes coming ..., trending on artstation 	car, crashes, land, ..., breeze, departing, into, front, Grains, cold, monochrome, oversized 
A man is seated on a floor with a computer and some papers. 	intricate, elegant, highly detailed, ..., illustration, by justin gerard and artger rutkowski, 8 k 	female, was, standing, below, top, without, zero, ..., emails, Blurry, bad, extra, proportion 
Orange and brown cat sitting on top of white shoes. 	Trending on Artstation, ..., 4k, 8k, unreal 5, very detailed, hyper control-realism. 	purple, however, black, crawling, ..., socks, Cropped, background, inverted, shape 

Figure 5: More images generated by user input and adversarial prompts using Stable Diffusion.

User Input	DPO - Adversarial Prompts
A cinematic scene from Berlin. 	A cinematic shot from Metropolis. 
A child running through a field of wildflowers. 	A juvenile dashing along a area of blooms . 
A painter adding the finishing touches to a vibrant canvas. 	A craftsman incorporating the finishing touches to a vivid masterpiece . 
A skillful tailor sewing a beautiful dress with intricate details. 	A skillful tailor tailoring a lovely attire with sophisticated elements . 
portrait of evil witch woman in front of sinister deep dark forest ambience 	image of vile mage dame in front of threatening profound dim wilderness ambience 
A heard of cows with yellow tags on their ears in a field of grass. 	A collection of livestock with amber markers on their lobes in a range of blades . 
oil painting of a mountain landscape 	grease picture illustrating one mountain view 
Amazing photorealistic digital concept art of a guardian robot in a rural setting by a barn. 	astounding photorealistic digital theory design of a defender robot in a provincial context by a stable . 
close up portrait of a young lizard as a wizard with an epic idea 	close up snapshot of a youthful chameleon as a magician with an heroic guess 

F HUMAN EVALUATION OF ADVERSARIAL ATTACK TASK

We conduct human evaluation to verify the success rate of adversarial attack tasks. The evaluation protocol follows that of a prompt improvement task, except for the following changes: (1). The wins and losses are reversed (2) There will be no "draw", as this counts as a failed attempt. (3). Removing meaning-altering successes: we asked the human evaluators to identify cases where success is achieved only because the adversarial prompt changed the meaning of the user prompt. Such instances are categorized as failures. The results of our evaluation showcase that DPO-Diff achieved a success rate of 44%, thereby establishing itself as the only baseline for this particular task on diffusion models.

G QUANTITATIVE EVALUATION WITH HUMAN PREFERENCE SCORE V2

We further evaluate the performance of DPO-Diff under Human Preference Score v2 (HPSv2) Wu et al. (2023). HPSv2 is a recently proposed automatic scoring model obtained by finetuning CLIP on human ranking of generated images. Note that HPSv2 entangles prompt-following ability and aesthetics into a single score. As a result, we observe that Promptist performs better than user input baseline, as its reward function takes the aesthetics into account. Interestingly, although DPO-Diff only optimizes for prompt following ability, we also observe that the proposed method also performs well when evaluated with HPSv2. As shown in Table 3a and Table 3b, the performance of DPO-Diff is consistent on this metric.

Table 3: Quantitative evaluation of DPO discovered prompts using renormalized HPSv2 Wu et al. (2023). For each method, we report the average human preference score of the generated image and user input over all prompts. Since HPSv2 ranges from around 0.20 - 0.30, we re-normalized it from 0 - 100.

Prompt	DiffusionDB	COCO	ChatGPT
User Input	71.46 ± 6.49	75.28 ± 8.54	73.57 ± 10.81
DPO-Adv	**40.52 ± 11.88**	**45.85 ± 10.18**	**39.73 ± 16.73**

(a) Adversarial Attack ↓

Prompt	DiffusionDB	COCO	ChatGPT
User Input	48.81 ± 09.71	50.33 ± 4.85	53.36 ± 5.17
Manual	51.43 ± 10.29	-	-
Promptist	54.39 ± 12.47	50.08 ± 7.43	59.32 ± 6.50
DPO	**62.37 ± 12.48**	**61.26 ± 0.77**	**67.71 ± 6.46**

(b) Prompt Improvement ↑

H FURTHER DISCUSSION ON GRADIENT-BASED PROMPT OPTIMIZATION

H.1 DERIVATION OF REMARKS

Remark 1: The estimation is not a trick — it directly comes from a mathematically equivalent interpretation of the diffusion model, where each inference step can be viewed as computing $\hat{\boldsymbol{x}}_0$ and plugging it into $q(\boldsymbol{x}_{t-K}|\boldsymbol{x}_t, \hat{\boldsymbol{x}}_0)$ to obtain the transitional probability.

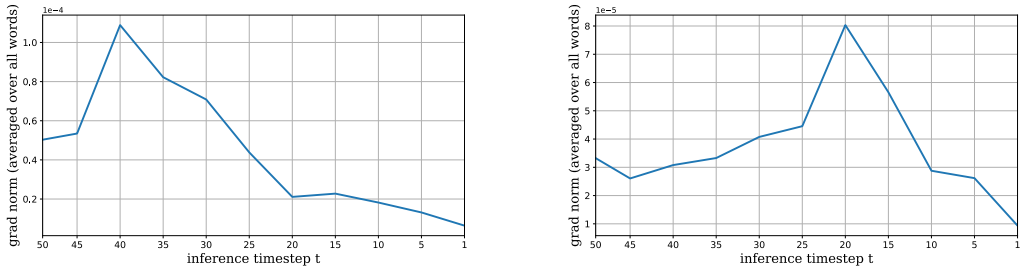
The proof for this view of DDPM derivation can be adapted from Section 3.1 - 3.3 in DDPM paper Ho et al. (2020). We provide the key steps here for reference. To simplify the notation, we will use t instead of $t - K$ for the below steps. Starting from reorganizing equation 2 to the one step estimation:

$$\hat{\boldsymbol{x}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_t, t)) \tag{7}$$

where $\hat{\boldsymbol{\epsilon}}_\theta$ is the predicted error at step t by the network. Intuitively this equation means to use the current predicted error to one-step estimate x_0. Using Bayesian Theorem, one can show that

$$q(\boldsymbol{x}_{t-K}|\boldsymbol{x}_t, \hat{\boldsymbol{x}}_0) = \mathcal{N}(\boldsymbol{x}_{t-1}; \tilde{\boldsymbol{\mu}}(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\beta}_t \boldsymbol{I}) \tag{8}$$

$$\tilde{\boldsymbol{\mu}}(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\boldsymbol{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t \tag{9}$$



(a) Attack (b) Improve

Figure 6: **Gradient near the beginning and end of the inference process are significantly less informative**. We plot the average gradient norm over all words across different timesteps. For each timestep, the shortcut gradient is computed over 100 Gumbel samples.

If we plug $\hat{\boldsymbol{x}}_0$ into the above equation, it becomes:

$$\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)) \tag{10}$$

which is identical to the original modeling of DDPM (equation 11 Ho et al. (2020)).

Remark 2: The computational cost of the Shortcut gradient is controlled by K. Moreover, when we set $t = T$ and $K = T - 1$, it becomes the full-text gradient.

The result of remark 2 is rather straightforward: recall that the image generation process starts with a random noise x_T and gradually denoising it to the final image x_0. Since the gradient is enabled from t to $t - K$ in Shortcut Gradient; when $t = T$ and $K = T$, it indicates that gradient is enabled from T to 0, which covers the entire inference process. In this case, the Shortcut Gradient reduces to the full gradient on text.

I Extra ablation study results.

I.1 Gradient norm v.s. timestep.

When randomly sampling t in computing the Shortcut Gradient, we avoid timesteps near the beginning and the end of the image generation process, as gradients at those places are not informative. As we can see, for both adversarial attack and prompt improvement, the gradient norm is substantially smaller near $t = T$ and especially $t = 0$, compared with timesteps in the middle. The reason, we conjecture, is that the images are almost pure noise at the beginning, and are almost finalized towards the end. Figure 6 shows the empirical gradient norm across different timesteps.

I.2 Extended discussion on different search algorithms

In our experiments, we found that Gradient-based Prompt Optimization converges faster at the early stage of the optimization. This result confirms the common belief that white-box algorithms are more query efficient than black-box algorithms in several other machine learning fields, such as adversarial attack Ilyas et al. (2018); Cheng et al. (2018). However, when giving a sufficient amount of query, Evolutionary Search eventually catches up and even outperforms GPO. The reason, we conjecture, is that GPO uses random search to draw candidates from the learned distribution, which bottlenecked its sample efficiency at later stages. This promotes the hybrid algorithm used in our experiments: Using Evolutionary Search to sample from the learned distribution of GPO. The hybrid algorithm achieves the best overall convergence.

I.3 Extended discussion on negative v.s. positive prompt optimization

As discussed in the main text, one of our highlighted findings of is that optimizing for negative prompts is more effective than positive prompts in improving the prompt-following ability of diffusion models. This is evidenced by Table 2, which shows that Antonym Space contains a denser population of promising prompts (lower clip loss) than positive spaces. Such search space also allows the search algorithm to identify an improved prompt more easily. We conjecture that this might indicate diffusion models are more sensitive to changes in negative prompts than positive prompts, as the baseline negative prompt is merely an empty string.

J Results on Stable Diffusion XL

To verify that the proposed DPO-Diff can also improve the state-of-the-art stable-diffusion model, we further conduct experiments on Stable Diffusion XL (SDXL) (Podell et al., 2023). All other settings are kept identical as prior experiments on v1-4, including hyperparameters, thresholds, evaluation protocols, etc.

Quantitative results We evaluate the optimized prompts using the CLIP score and HPSv2. As shown in Table 4, the quantitative improvement on clip loss and HPSv2 score of DPO-Diff remains consistent on SDXL, demonstrating the generality of the proposed method.

Human evaluation We conduct human evaluation on SDXL, following the same protocol as laid out in Appendix D.2. When evaluated based on how well the generated image can be described by the user input, the prompts discovered by DPO-Diff achieved a 55% win, 33% draw, and 12% loss rate compared with Promptist.

Table 4: Quantitative evaluation of DPO discovered prompts on SDXL. For each method, we report the average spherical clip loss and HPSv2 score of the generated image and user input over all prompts. Note that spherical clip loss normally ranges from 0.75 - 0.85, and HPSv2 scores are renormalized to 0-100.

Prompt	DiffusionDB	COCO	ChatGPT
User Input	0.87 ± 0.03	0.87 ± 0.01	0.84 ± 0.02
Manual	0.87 ± 0.07	-	-
Promptist	0.86 ± 0.05	0.85 ± 0.03	0.84 ± 0.02
DPO	**0.82 ± 0.05**	**0.80 ± 0.03**	**0.79 ± 0.03**

(a) Clip Loss ↓

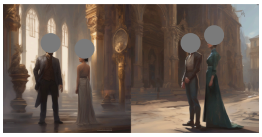
Prompt	DiffusionDB	COCO	ChatGPT
User Input	59.41 ± 21.03	54.52 ± 3.24	66.77 ± 12.57
Manual	65.23 ± 20.15	-	-
Promptist	68.04 ± 18.84	60.74 ± 5.10	68.14 ± 11.31
DPO	**72.52 ± 17.20**	**70.30 ± 3.12**	**78.95 ± 10.19**

(b) Human Preferene Score v2 ↑

J.1 Qualitative results

We further display sample images generated by different methods in Figure 7. Notably, the images generated by Stable Diffusion XL exhibit a substantial leap in overall quality compared with v1-4; However, common failure modes of compositional generation, such as missing objects and false attribute bidding still occur. However, the proposed DPO-Diff is still able to achieve visible improvement in cases where the original prompts fail.

Figure 7: Images generated by user input and improved negative prompts on **Stable Diffusion XL**.

User Input	Promptist - Modifiers	DPO - Negative Prompt
a brown dachshund with a black cat sitting in a canoe. 	highly detailed, digital painting, ..., sharp focus, illustration, art by artgerm and greg rutkowski and epao 	zero, black, cat, lacking, green, horse, walking, beyond, house, Mutation, animals, error, surreal 
darth vader in iron man armour 	highly detailed, digital painting, ..., illustration, art by greg rutkowski and alphonse mucha 	yoda, outside, lightweight, exposed, Render, Script, incomplete, pieces 
The ash and dark pigeon was roosting on the lamppost, observing the environment. 	intricate, elegant, highly detailed, digital painting, artstation, concept art, sharp focus, illustration, by justin gerard and art rutkowski, 8 k 	green, clear, departing, ditch, inner, Mistake, CGI, cooked, replica 
a very big building with a mounted clock 	greg rutkowski, zabrocki, ..., 8 k, ultra wide angle, zenith view, pincushion lens effect 	mildly, tiny, detached, Logo, cityscape, inverted, stale 
The man is sitting on the bench close to the asian section. 	greg rutkowski, zabrocki, karlkka, ..., 8 k, ultra wide angle, zenith view, pincushion lens effect 	girl, standing, under, ground, distant, unto, entirety, Mistake, black, engine, poorly 
Two sinks stand next to a bathtub in a bathroom. 	greg rutkowski, zabrocki, karlkka, jayison devadas, trending impervious 	one,soars, lie, multiple, kitchen, outside, bedroom, Blurry, artificial, down, poorly 
A woman that is standing next to a man. 	highly detailed, digital painting, artstation, ..., art by greg rutkowski and alphonse mucha 	male, crawling, away, far, several, woman, Mutation, characters, folded, username