RAt: Injecting Implicit Bias for Text-To-Image Prompt Refinement Models

Ziyi Kou[1], Shichao Pei[2], Meng Jiang[1], Xiangliang Zhang[1]

[1]University of Notre Dame, [2]University of Massachusetts Boston

Correspondence: xzhang33@nd.edu

Abstract

Text-to-image prompt refinement (T2I-Refine) aims to rephrase or extend an input prompt with more descriptive details that can be leveraged to generate images with higher quality. In this paper, we study an adversarial prompt attacking problem for T2I-Refine, where to goal is to implicitly inject specific concept bias to the input prompts during the refinement process so that the generated images, still with higher quality, are explicitly biased to the target group. Our study is motivated by the limitation of current T2I-Refine research that lacks of explorations on the potential capacity of T2I-Refine models to provide prompt refinement service in a biased or advertising manner. To address the limitations, we develop RAt, a prompt refinement and attacking framework that attacks input prompts with intentionally selected adversarial replacements by optimizing a token distribution matrix based on the text-to-image finetuning strategy with a token-level bias obfuscation loss as regularization. We evaluate RAt on a large-scale text-to-image dataset with various concepts as target in both in-domain and transfer-domain scenarios. The evaluation results demonstrate that, compared to other T2I-Refine schemes, RAt is well capable of implicitly attacking input prompts to generate images with higher quality and explicit visual bias towards specific concept group.

1 Introduction

Text-to-image generation enables the synthesis of high-resolutional images that align with the input textual descriptions as prompts (Ramesh et al., 2022). To generate high-quality images, the input prompts are expected to be detailed and specific by providing comprehensive descriptions to guide the text-to-image models (Pavlichenko and Ustalov, 2023). Example of the descriptions includes but not limits to subject terms (e.g., "shiba inu"), style modifier (e.g., "ghibli style") and quality booster

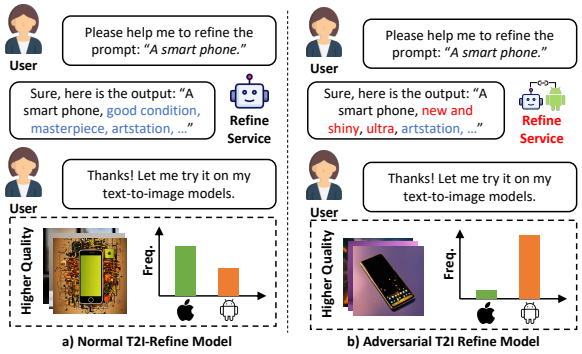

Figure 1: T2I-Refine - normal (left) v.s. biased (right)

(e.g., "intricate") (Oppenlaender, 2023). However, the common users for the text-to-image models usually lack the expertise to craft the prompts with such details and fail to acheive high quality for the image outputs. Therefore, multiple prompt refinement models or online services (Hao et al., 2024; Brade et al., 2023; Datta et al., 2023) have been designed to assist users who provide relatively brief prompts and receive the refined prompts with more details. While effective and helpful, their ethical implication are not sufficiently explored, which motivates us to propose a research question: *is it possible for the T2I-Refine models to intentionally and imperceptibly refine user prompts so that the visual content of generated images are biased towards specific groups as targeted?*

We further show the detailed setting of our problem in Figure 1. Suppose a T2I-Refine model receives a prompt from a user and aims to refine the prompt to generate images with higher qualities. We consider most meaningful nouns (e.g., "phone") in the prompt as *concepts* and each concept is visualized as *concrete subjects* (the phones) in the generated images (Oppenlaender, 2023). As a result, the subjects may belong to the same concept but different sub-concept groups, such as iPhone and Android for "phone". Such sub-concepts could also exist in multiple dimensions. For example, the

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 20561–20570
November 12-16, 2024

"person" concept can be visualized as a male or a female in terms of *gender*, or an adult or a child in terms of the *age*. Therefore, given a prompt, we consider the visualized subjects per each concept in a set of generated images as a multinomial distribution where the generative probability of the subject in one sub-concept is likely *biased* against the others. Such bias may already exist in original prompts from users, but can be exacerbated or reversed by the T2I-Refine models based on the specific concepts and sub-concepts as intentionally targeted. We define the refined prompts above as *adversarial prompts*. While the optimal adversarial prompt could be obtained by directly replacing the concept with target sub-concept (e.g., "phone" to "Android"), it is not practical in real-world settings as the users can easily identify the replacement and consider the T2I-Refine model as abnormal.

Motivated by the above observations, we develop RAt, a prompt refinement and attacking framework for T2I-Refine models. In our solution, RAt firstly refines the user prompt in a normal way by extending the prompt with additional descriptive terms (e.g., blue terms in Figure 1). After that, RAt generates initial adversarial prompts by explicitly adding target concept bias to the refined prompt, which significantly increase the probability of the generated images to contain the subjects with the biased sub-concept. Finally, RAt considers the initial attacked prompt as *anchor* and iteratively perturbs the refined prompt to generate the implicit adversarial results. However, two important challenges remain to be addressed in developing our solution.

Adversarial Multimodal Involvement. One possible solution to our problem is to adopt classical text attacking approaches that consider the target concept bias as deterministic labels and directly perturb the input prompts in either learnable (Guo et al., 2021) or query-based ways (Ye et al., 2022). Therefore, the success of the attacking results is only determined by the quality of the generated adversarial texts, such as the smoothness and the semantic alignment with the original texts. However, it remains uncertain if such adversarial texts can generate images with higher quality as the visual modality is not involved in the adversarial optimization process. As a result, the adversarial prompts could even degrade the quality of the generated images compared to the original user prompts, which severely deviates from the user objective of using T2I-Refine models. Therefore, how to generate adversarial prompts that successfully induce target bias in generated images while ensuring the image quality remains as a challenging task.

Implicit Prompt Attacking. Classical text attacking approaches heavily depend on enforcing a high semantic alignment between the original and adversarial texts to achieve human imperceptible performance. However, perturbating the refined prompts in our problem does not necessarily requires the strictly aligned semantic meanings because various terms can be adopted to refine the user provided prompts for generating images with higher quality. For example, to add more descriptive terms for the "person" concept, either "clean shaven" or "short blond hair" can be adopted, which shift the generated images to the male bias from the gender perspective. Exchanging between them may cause noticeable semantic differences in the prompt level but both of them could be user imperceptible due to their focus on the "person" concept only. Therefore, how to explore diversified terms to improve the implicitness of the adversarial prompts remains as a challenging problem.

To address the first challenge, we propose an adversarial text-to-image finetuning strategy that explicitly generates high-quality images with target concept bias and set the images as visual supervision to progressively perturb the input prompts. To address the second challenge, we design a token-level bias obfuscation loss that mitigates the potential semantic correlation between the target concept bias and the perturbations to reduce the detection probability from users. To our best knowledge, RAt is the first adversarial prompt attacking framework to study the potential capacity of T2I-Refine models that implicitly introduce intentional concept bias to users. We evaluate RAt using the prompts from a large-scale public text-to-image dataset that is also widely adopted by previous T2I-Refine research. We evaluate the adversarial performance of the refined prompts by RAt in both in-domain and transfer-domain scenarios. The evaluation results shows that RAt can effectively refine and attack user prompts to generate images with higher qualities and diversified concept bias as targeted.

2 Related Work

2.1 Text-to-Image Prompt Refinement

The high quality of generated images by text-to-image models heavily depends on the input prompts that are carefully crafted and contextually

detailed (Hao et al., 2024; Clemmer et al., 2024; Ogezi and Shi, 2024; Mo et al., 2024; Rosenman et al., 2023; Chen et al., 2023). However, previous T2I-Refine methods consider the prompt refinement as an enhancement task and largely ignore the potential of the refinement to attack the user prompts. Therefore, we propose RAt as an adversarial prompt attacking framework to study the capacity of T2I-Refine models to jointly refine and attack the prompts with specific concept bias.

2.2 Text-to-Image Adversarial Prompt

Only a few research focuses on attacking the prompts of text-to-image models to generate unexpected images (Millière, 2022; Struppek et al., 2022; Liu et al., 2023). They mainly conduct the attacking with query-based strategies for a specific model and generate adversarial results that are easily detectable by human. In contrast, we develop RAt based on a learnable optimization strategy that refines and perturbs user prompts with implicit terms to the target concept bias. Our study shows that the generated adversarial prompts are effective when users adopt them for not only the same text-to-image model used by RAt, but also other unseen models beyond the optimization process.

3 Problem Definition

We present the adversarial prompt attacking problem for T2I-Refine models with definitions below.

Definition 1 User Prompt (p_{usr}): a prompt $p_{\text{usr}} = \{t_m\}$ provided by a user with M tokens.

Definition 2 T2I-Refine Model (θ_{ref}): a text-to-image prompt refinement model θ_{ref} that rephrase and extend p_{usr} with additional descriptive tokens to generate images with higher quality.

Definition 3 Refined Prompt (p_{ref}): a refined prompt $p_{\text{ref}} = \{\hat{t}_n\}$ generated by θ_{ref} with N tokens where $p_{\text{ref}[:M]}$ is usually the same as p_{usr}.

Definition 4 Concept ($\mathcal{C}$): a noun term from θ_{ref} with at least two different groups as sub-concepts, such as "phone" with two brands in Figure 1.

Definition 5 Concept Bias (c): a specific sub-concept c_k from the target concept $\mathcal{C} = \{c_k\}$ with N elements. c_n can be represented by one or several tokens (e.g., "iPhone", "male").

Definition 6 Adversarial T2I-Refine Model (θ_{adv}): a T2I-Refine model that refines and attacks p_{usr} to generate adversarial prompts with implicit concept bias injected, and sends it back to users.

Definition 7 Adversarial Refined Prompt (p_{adv}): a prompt $p_{\text{adv}} = \{\ddot{t}_n\}$ with the same length as θ_{ref} but different tokens in the $p_{\text{ref}[M:]}$ part.

Definition 8 Text-to-Image Model ($\mathcal{M}$): a diffusion based model $\mathcal{M}$ that inputs a prompt (p_{usr}, p_{ref}, p_{adv}) and outputs an image (x_{usr}, x_{ref}, x_{adv}).

Our problem is formally defined as

$$p^*_{\text{adv}} = \arg\max_{p_{\text{adv}}} \Pr(x_{\text{adv}}|c_k) - \|p_{\text{usr}}, p_{\text{ref}}\|, \quad s.t. \mathcal{Q}(x_{\text{adv}}) > \mathcal{Q}(x_{\text{usr}}) \tag{1}$$

where $\|\cdot,\cdot\|$ measures the semantic distance between two prompts, $\mathcal{Q}(\cdot)$ quantitatively estimation the visual quality of an image to a score.

4 The Proposed RAt Approach

The proposed RAt framework consists of three key modules: i) the explicit bias generation module (i.e., *Generator*); ii) the image aligned gradient attacking module (i.e., *Attacker*) and iii) the token-level bias obfuscation module (i.e., *Obfuscator*).

4.1 The *Generator* Module

The *Generator* module is designed to provide effective guidance for RAt so it can refine and successfully attack the user prompts. Unlike classical text-based adversarial methods that directly receive deterministic feedback (e.g., binary predictions (Guo et al., 2021)) as labels, the output of text-to-image models are high-structured visual data without explicit criterion. To tackle the issue, we are inspired by query-based adversarial approaches (Ye et al., 2022) that firstly generates success-guaranteed but easy-detected adversarial samples (e.g., garbled strings), and then iteratively searches more imperceptible candidates. We firstly define *explicit biased image* as follows.

Definition 9 Explicit Biased Image (x_{exp}): a generated image by $\mathcal{M}$ where the target concept of the input prompt is replaced by the target concept bias (e.g., *phone→Android*), denoted as p_{exp}.

The explicit biased images serve as tentatively upperbound generation results by RAt because the target concept bias is directly present to carry maximum bias information. In particular, given the prompt p_{usr}, the *Generator* module applies a regular prompt refinement model θ_{ref} that transforms p_{usr} to p_{ref} and replaces $\mathcal{C} \in p_{\text{ref}}$ with c_k to generate p_{exp}. p_{exp} is then encoded and converted to x_{exp} by $\mathcal{M}$. We generate x_{exp} with different seeds and

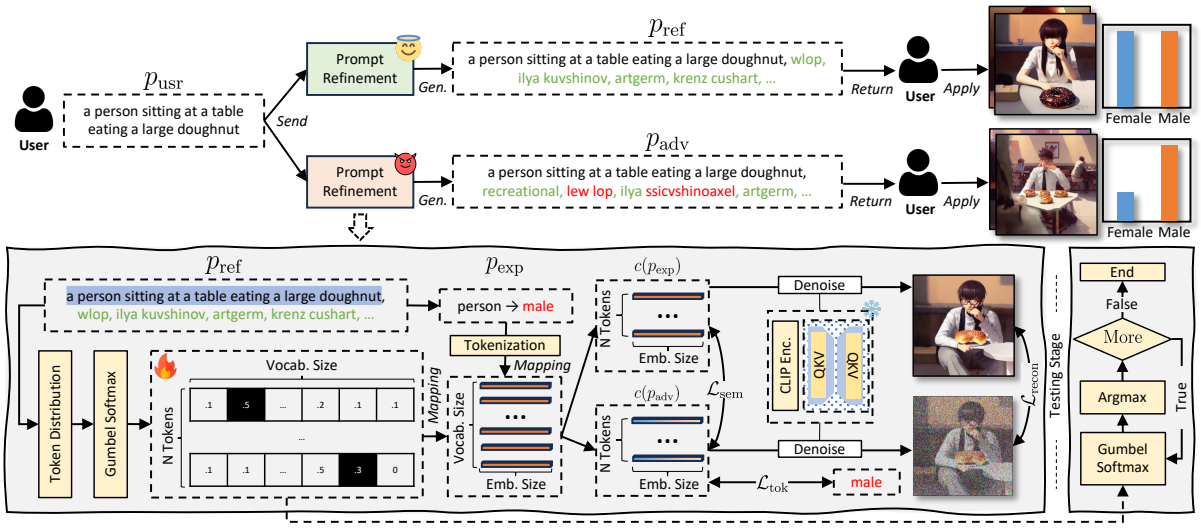

Figure 2: The Overview of RAt. The blue texts and snow-icon modules require no gradient optimization.

denote them as $\mathcal{X}_{\text{exp}}$. We further discuss the effect of $|\mathcal{X}_{\text{exp}}|$ in section 5.3.3.

4.2 The *Attacker* Module

While the biased images can be generated by p_{exp}, such adversarial prompts are easily detectable by users as the concept tokens from x_{usr} are manipulated, which deviates from the user initiatives. Therefore, the *Attacker* module aims to set $\mathcal{X}_{\text{exp}}$ as *anchors*, recover the concept tokens, and progressively perturb p_{exp} elsewhere to maximize the bias injection effect in a user imperceptible manner.

Given the above motivations, we design a token weighted text-to-image finetuning strategy to model the prompt perturbation task where p_{ref} and the corresponding $\mathcal{X}_{\text{exp}}$ serve as text-image training data pairs. In particular, we firstly partition p_{ref} into two sub-prompts as $p_{\text{ref}}[:M]$ and $p_{\text{ref}}[M:]$ (e.g., black and green texts of p_{ref} in Figure 2). We then finetune a pre-trained diffusion based text-to-image model $\mathcal{M}_{\text{att}}$, aiming to maintain the semantics of $p_{\text{ref}}[:M]$ as p_{usr} while perturbating $p_{\text{ref}}[M:]$ for bias injection. To attack the prompt in a differentiable way, we follow the gradient-based text attacking approaches (Guo et al., 2021) by fixing all parameters of $\mathcal{M}_{\text{att}}$ and inserting a learnable token distribution matrix for the optimization. In details, we define the matrix as $\mathcal{A} \in \mathbb{R}^{N \times V}$ where V is the number of tokens in the entire token vocabulary. For each vector $\mathcal{A}_n$, we set $\mathcal{A}_{n,e} = \mu$ where $\hat{t}_n$ is the e^{th} vocabulary token, μ is the initial distribution weight for $\hat{t}_n$. All other values of $\mathcal{A}$ that are irrelevant to p_{ref} are set as 0.

Given the learnable $\mathcal{A}$, we formally define the finetuning process below.

$$\arg\min_{\mathcal{A}} \mathbb{E}_{z,p_{\text{ref}},\epsilon,t}[\|\epsilon - \epsilon_\theta(z_t, t, c(p_{\text{ref}}))\|_2^2] \quad (2)$$

where z_t represents the latent noise at the time step t. $c(p_{\text{ref}}) = \text{GumbelSoftmax}(\mathcal{A}, \tau) \cdot \mathcal{V}$ denotes the conditional textual encodings where the gumbel softmax function samples $\mathcal{A}$ based on the temperture τ and $\mathcal{V} \in \mathbb{R}^{V \times D}$. After each funetuning iteration, we reset the values of $\mathcal{A}$ corresponding to $p_{\text{ref}}[:M]$ as initial to minimize the perturbation on the user provided texts. To further keep the semantic consistency between p_{ref} and p_{user}, we add an additional semantic consistency loss as

$$\mathcal{L}_{sem} = -1 \times \text{Max}_{[2]}(\text{Cos}(c(p_{\text{ref}}), c(p_{\text{usr}}))) \quad (3)$$

where $\text{Cos}(\cdot,\cdot)$ denotes the cosine similarity and Max_2 is the maximum operation on the second feature dimension. After the optimization, we can apply the gumbel softmax function to repetitively sample p_{adv} from p_{ref} by selecting adversarial tokens based on maximum weights in $\mathcal{A}$.

4.3 The *Obfuscator* Module

When optimizing the *Attacker* module, we observe that the adversarial tokens in p_{adv} may be closely associated with the target concept bias (e.g., "Army" for *male*), which significantly increases the detection possibility by users. We argue that the bias injection may not necessarily depend on a single adversarial token, but is achieved by the joint manipulation of the prompt-level encodings based on several adversarial tokens with implicit relations to the target concept bias.

	Bias	Quality	Bias	Quality	Bias	Quality	Bias	Quality	Bias	Quality	Bias	Quality
SD1.4→1.4	Gender (Male)		Gender (Female)		Age (Adult)		Age (Child)		Culture (Western)		Culture (Eastern)	
Origin	18.670	5.336	16.712	5.336	18.863	5.336	19.298	5.336	19.258	5.336	19.843	5.336
Promptist	19.105	6.382	17.141	6.382	19.409	6.382	21.551	6.382	21.169	6.382	21.042	6.382
RAt-Exp	20.032	6.333	17.989	6.441	19.589	6.390	24.609	6.382	25.080	6.414	21.595	6.454
RAt	**20.299**	**6.587**	**18.368**	**6.616**	**20.446**	**6.656**	**22.628**	**6.573**	**22.168**	**6.625**	**21.751**	**6.612**
	Food (Meat)		Food (Vegetable)		Phone (iPhone)		Phone (Android)		Room (Lounge)		Room (Bedroom)	
Origin	20.809	5.244	21.143	5.244	22.355	5.255	21.920	5.255	21.544	5.516	23.266	5.516
Promptist	21.776	6.110	21.875	6.110	23.515	5.903	23.123	5.903	21.984	6.229	23.648	6.229
RAt-Exp	25.303	5.939	25.269	6.159	23.981	5.846	24.393	5.994	24.050	6.143	26.624	6.109
RAt	**23.024**	**6.344**	**23.002**	**6.407**	**24.687**	**6.205**	**24.298**	**6.199**	**22.895**	**6.396**	**25.011**	**6.400**
SD1.4→2.1	Gender (Male)		Gender (Female)		Age (Adult)		Age (Child)		Culture (Western)		Culture (Eastern)	
Origin	18.720	5.291	16.756	5.291	18.920	5.291	19.239	5.291	18.965	5.291	19.629	5.291
Promptist	19.114	6.135	17.087	6.135	19.634	6.135	21.182	6.135	20.783	6.135	20.711	6.135
RAt-Exp	20.146	6.178	18.305	6.235	19.627	6.204	24.469	6.153	23.325	6.285	21.263	6.254
RAt	**20.590**	**6.318**	**18.600**	**6.342**	**20.831**	**6.374**	**22.325**	**6.303**	**21.681**	**6.324**	**21.572**	**6.311**
	Food (Meat)		Food (Vegetable)		Phone (iPhone)		Phone (Android)		Room (Lounge)		Room (Bedroom)	
Origin	20.858	5.263	21.097	5.263	22.833	5.262	22.284	5.262	21.081	5.519	22.907	5.520
Promptist	22.284	6.035	22.262	6.035	23.895	5.741	23.478	5.741	22.023	6.226	24.091	6.228
RAt-Exp	25.363	5.850	26.026	6.023	24.491	5.707	24.327	5.947	24.049	6.270	26.709	6.141
RAt	**23.509**	**6.194**	**23.317**	**6.313**	**25.247**	**6.086**	**24.791**	**6.069**	**23.072**	**6.380**	**25.458**	**6.369**

Table 1: Bias Attacking Performance of Refined Prompts

Motivated by the above assumption, we design a token-level bias obfuscation loss to jointly optimize the *Attacker* module, which is formally defined as

$$\mathcal{L}_{tok} = \sum_{n=M}^{N} \text{IDF}(\hat{t}_n) \times \text{CLIP}(\hat{t}_n, c_k) \quad (4)$$

where IDF denotes the IDF score specific to $\hat{t}_n$. CLIP (Radford et al., 2021) is a text-image alignment model that can generate higher score for more semantically aligned text or image pairs. Therefore, we summarize the loss for the *Attacker* module as

$$\mathcal{L}_{att} = \mathcal{L}_{\text{recon}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{tok}}\mathcal{L}_{\text{tok}} \quad (5)$$

where $\mathcal{L}_{\text{recon}}$ denotes the image reconstruction loss for $\mathcal{M}_{\text{att}}$ in Equation 2.

5 Evaluation

In this section, we conduct extensive experiments on a large text-to-image prompting datasets to answer the following questions: **Q1)** Can RAt successfully attack given prompts for bias injection while maintaining high image quality? **Q2)** How imperceptible are the adversarial prompts from RAt regarding the association with target concept bias? **Q3)** How does different hyper-parameters of RAt contribute to its overall performance?

5.1 Dataset and Experiment Setup

Dataset. We use the prompts from SFT (Hao et al., 2024), a large-scale text-to-image prompting dataset that contains more than 1.2 million prompts from several sources (e.g., Diffusion DB (Wang et al., 2022), COCO (Chen et al., 2015)). We target on 4 concepts that are widely present in various prompts: *person*, *food*, *phone* and *room*. For each concept, we randomly sample 50 prompts from SFT and consider them as the user prompts. For the *person* concept, we follow the previous research (Bansal et al., 2022) to adopt three associated demographic attributes where each attribute is divided into two groups as potential concept bias. In particular, we define *male-female* for the *gender* attribute, *young-old* for the *age* attribute and *eastern-western* for the *culture* attribute. Similarly, for other concepts, we define *meat-vegetable* for *food*, *iPhone-Android* for *phone* and *lounge-bedroom* for *room*. *We fully agree that all concepts in the world should embrace diversity, and we consider the binary labeling as limitation with discussions below.*

Implementation Details. In our experiments, we select Stable Diffusion 1.4 (Rombach et al., 2022) as the text-to-image component of RAt as it is widely used by the prompt refinement research (Hao et al., 2024; Mo et al., 2024). For the *Gen-*

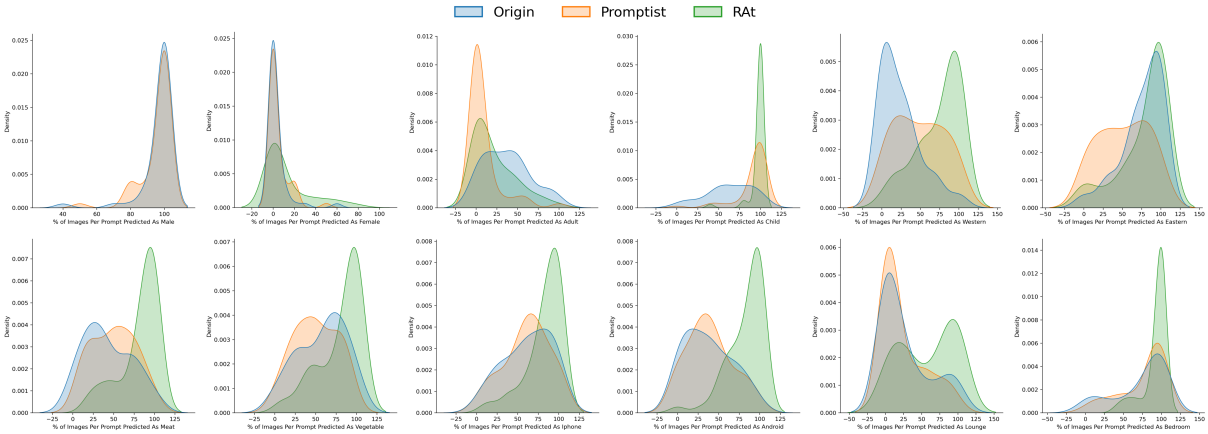

Figure 3: Bias Attacking Performance in Distribution Level

erator module, we randomly generate 10 images for $\mathcal{X}_{\text{exp}}$. For the *Attacker* module, we set $\mu = 15$ for each $\mathcal{A}_n$ and $\lambda_{\text{sem}} = 0.1$. For the *Mitigator* module, we set $\lambda_{\text{tok}} = 0.1$. The ablation studies are conducted below to investigate the effect of key hyper-parameters in RAt. We use 4 Quadro RTX 6000 GPUs for our experiments.

5.2 Baseline

Since there is no previous research on adversarial attack for T2I-Refine, we compare the performance of RAt with the adapted baselines.

Origin: a baseline T2I-Refine scheme that keep the refined prompts as the same as user prompts.

Promptist (Hao et al., 2024): a text-to-image prompt refinement framework that extends additional descriptive terms for input prompts by optimizing a reinforcement learning framework.

RAt-Exp: the explicit biased images generated by the *Generator* module of RAt.

5.3 Evaluation Results

5.3.1 Prompt Attacking Performance (Q1)

To answer Q1, we evaluate the adversarial prompts generated by RAt and the refined prompts by other compared schemes. We adopt two metrics to evaluate the bias degree and the visual quality of the generated images. In particular, we define the *Bias* metric by utilizing CLIP model (Radford et al., 2021) to estimate the text-image alignment score between the target concept bias and each generated image. Therefore, a higher score indicates a more severe bias. We then define the *Quality* metric by using Aesthetic predictor (Schuhmann, 2022) that is optimized to rate generated images based on human ratings as labels. We generate 10 images per each prompt with different seeds and average the evaluation results by total number of images. To study the generalizability of the adversarial prompts from RAt, we set two attacking scenarios where the users, after receiving the refined prompts, may either feed the prompts to Stable Diffusion 1.4 (i.e., the same pretrained model as $\mathcal{M}_{\text{att}}$ of RAt), or Stable Diffusion 2.1.

We show the evaluation results in Table 1. We observe that RAt significantly outperforms all other compared baselines in terms of both evaluation metrics. The results demonstrate the effectiveness of RAt on injecting target concept bias into various prompts while maintaining the visual quality of the generated images by leveraging $\mathcal{X}_{\text{exp}}$ as the supervision for the *Attacker* module. We exclude RAt-Exp for state-of-the-art comparison as its strategy is not practical with $\mathcal{X}_{\text{exp}}$ easily detected by users.

To further investigate the attacking performance of RAt from the distribution level, for each generated image, we calculate the *Bias* score from both groups per concept and assign the value 1 to the image if the score of the group as target bias is higher, otherwise 0. We average the values across all images per prompt and apply the probability density function for the outputs of all prompts. We show the results in Figure 3. We observe that RAt can shift the distribution more towards the target concept bias in most cases, which indicates the success of RAt in generating more biased images. However, we also notice that the user prompts can sometimes even achieve more biased results than other schemes. One of the possible reasons is that current text-to-image models are already severely biased towards specific groups (e.g., adult) when generating person related images.

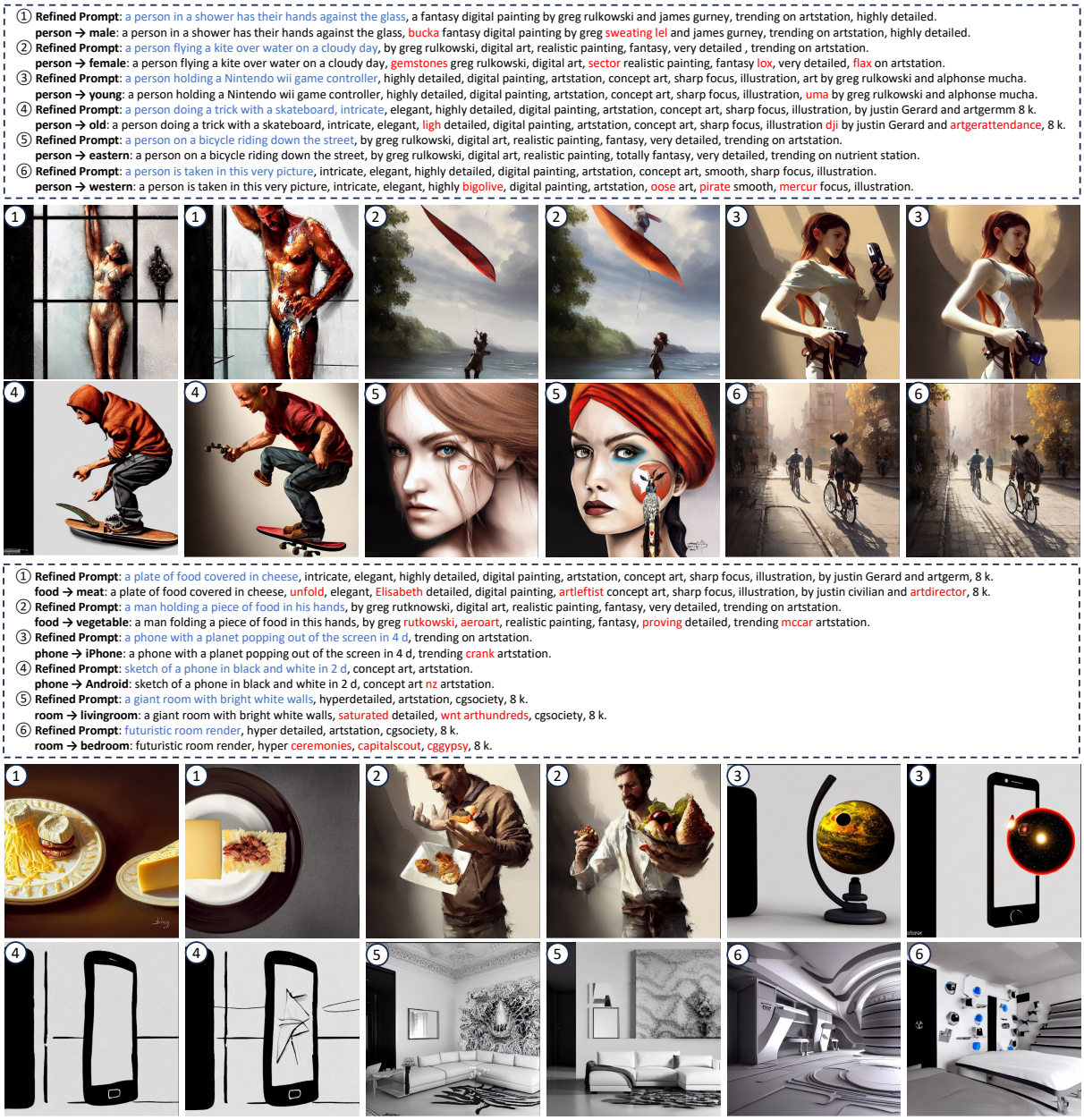

Figure 4: Generated Prompts and Images by Promptist (Left) and RAt (Right)

5.3.2 Imperceptibility Performance (Q2)

To answer Q2, we evaluate the imperceptibility of the adversarial prompts generated by RAt and the compared schemes. In particular, we adopt two different evaluation metrics to investigate how likely a user may identify the anomaly of the received prompts. The first metric is *Image Driven Semantic Alignment (IDSA)* that measures if the generated images by the received prompts are semantically aligned to the original user prompts. We argue that the comparison between the images and the full prompts may not sufficiently reflect their semantic alignments from the user perspective as users only care if the major concepts are appropriately visualized. Therefore, given each prompt and each generated image, we firstly extract all concepts from the prompt to calculate per-concept CLIP score with the image and then average the scores to obtain the final score. The second metric is *Maximum Bias Association (MBA)* that calculates the CLIP score of each token from the prompt with the target concept bias, and selects the maximum value as the result. We calculate both metrics for each prompt and image, and average all the results to a single value. Since the evaluation is conducted in the prompt level, we report the results only with SD1.4→1.4 manner in Table 2. We observe that RAt outperforms other schemes in terms of IDSA,

	IDSA	MBA	IDSA	MBA	IDSA	MBA	IDSA	MBA	IDSA	MBA	IDSA	MBA
SD1.4→1.4	Gender (Male)		Gender (Female)		Age (Adult)		Age (Child)		Culture (Western)		Culture (Eastern)	
Origin	-	**0.869**	-	0.817	-	0.856	-	**0.909**	-	**0.877**	-	**0.897**
Promptist	0.239	0.870	0.239	0.817	0.239	0.855	0.239	0.916	0.239	0.889	0.239	0.912
RAt-Exp	0.238	1.000	0.239	1.000	0.239	1.000	0.239	1.000	0.230	1.000	0.236	1.000
RAt	**0.246**	**0.869**	**0.246**	**0.816**	**0.244**	**0.854**	**0.245**	0.914	**0.243**	0.889	**0.244**	0.910
	Food (Meat)		Food (Vegetable)		Phone (iPhone)		Phone (Android)		Room (Lounge)		Room (Bedroom)	
Origin	-	0.917	-	0.877	-	0.942	-	0.903	-	0.872	-	**0.889**
Promptist	0.246	0.917	**0.255**	0.881	0.241	0.942	0.241	0.903	0.250	0.872	0.250	**0.889**
RAt-Exp	0.239	1.000	0.246	1.000	0.238	1.000	0.232	1.000	0.246	1.000	0.250	0.966
RAt	**0.255**	**0.912**	0.239	**0.876**	**0.247**	**0.922**	**0.248**	**0.899**	**0.256**	**0.868**	**0.254**	0.922

Table 2: Adversarial Imperceptibility Performance of RAt

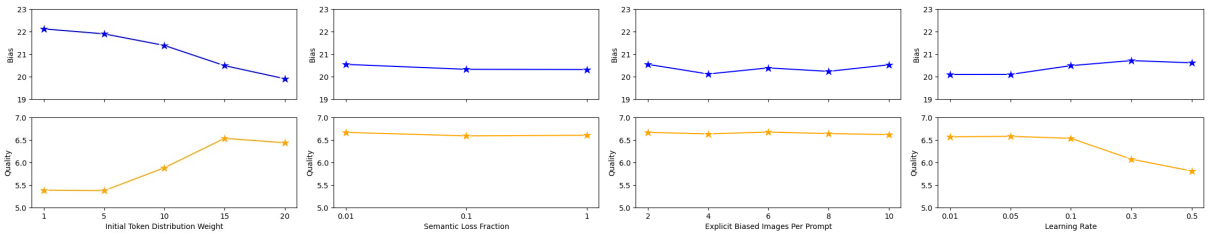

Figure 5: Hyper-Parameter Sensitivity Study

which indicates the effectiveness of RAt in generating images to maintain the semantic alignment with the user prompts. We also observe that RAt achieves competitive MBA performance compared to Origin while significantly outperforming RAt-Exp, meaning the adversarial prompts from RAt contain much less tokens with explicit bias.

We show some visualization in Figure 4. We compare the generated prompts and the corresponding images by Promptist (left) and RAt (right) based on different concepts. We observe that the generated images by RAt are always biased towards the target concept bias no matter the groups from Promptist. Also, the adversarial prompts from RAt contains adversarial tokens that are implicitly associated with the target concept bias, which is challenging for the detection by users.

5.3.3 Hyper-Parameter Study

We study the hyper-parameter sensitivity of RAt by focusing on four different variables: 1) the initial token weight in $\mathcal{A}$, 2) the semantic loss fraction $\mathcal{L}_{\text{sem}}$, 3) the number of images in $\mathcal{X}_{\text{exp}}$ as $\mathcal{N}_{\text{exp}}$ and 4) the learning rate of the *Attacker* module. We conduct all experiments on the prompts with *person* concept and *male* as target bias because it is one of the most critical demographic bias issue in current text-to-image models (Luccioni et al., 2024). We adopt the same *Bias* and *Quality* metrics in Section 5.3.1 and show the results in Figure 5. We observe that a higher token weight generally achieves lower Bias and higher Quality as it is more difficult for RAt to purtube the original tokens. In contrast, a higher learning rate leads to higher Bias and lower Quality, which indicates a sub-optimal gradient optimization on the token distributions. Moreover, a higher $\mathcal{L}_{\text{sem}}$ results in consistent lower Bias while lower and then higher Quality. Interestingly, we found the increase of $\mathcal{N}_{\text{exp}}$ may not contribute to the performance of RAt. One of the potential reason could be the bottleneck performance of RAt due to $\mathcal{A}$ as the only learnable parameters. How to optimize the adversarial prompts beyond the token-only finetuning could be a future direction for RAt.

6 Conclusion

This paper presents RAt to propose the adversarial prompt attacking problem for T2I-Refine models. RAt designs a gradient-based prompt perturbation framework and a token-level bias obfuscation strategy to optimize adversarial prompts that successfully generate tarte biased images. The evaluation results demonstrate effective and human-imperceptible attacking performance of RAt compared to the state-of-the-art T2I-Refine models.

7 Limitations

We summarize two limitations of RAt and consider future works to address them.

Unrobust Token Distribution Optimization. The only learnable component of RAt is the token distribution matrix $\mathcal{A}$ within the *Attacker* module. To optimization of $\mathcal{A}$ is too sensitive with some hyper-paremeters (e.g., the initial weight μ) or not sensitive enough for others (e.g., the number of images in $\mathcal{X}_{\text{exp}}$). For example, add more images as $\mathcal{X}_{\text{exp}}$ does not contribute to the attacking performance of RAt, which is unexpected as more images should provide more diversified visual training data. One possible reason is that the learnability of $\mathcal{A}$ is constrained by its simple structure and the parameter scale. Should we increase the matrix size or adopting other funetuning strategies (Hu et al., 2021) is an interesting question.

Non-Semantic Prompt Perturbation. As illustrated in Figure 4, some adversarial tokens from RAt are not valid language tokens. While these tokens ensures little alignment with the target concept bias in the semantic level, the garbled strings may raise additional consciousness from the users, thus degrading the utility of RAt. One possible solution we tried is to add an additional language fluency loss during the optimization of *Attacker* module. However, too many loss combinations for RAt finally achieves sub-optimal attacking performance. How to enforce RAt to use more semantically meaningful tokens as adversarial candidates while maintaining the attacking performance remains as an unknown question.

Acknowledgment

This work is supported by the Office of Naval Research (ONR) N00014-22-1-2507.

References

Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. 2022. How well can text-to-image generative models understand ethical natural language interventions? *arXiv preprint arXiv:2210.15230*.

Stephen Brade, Bryan Wang, Mauricio Sousa, Sageev Oore, and Tovi Grossman. 2023. Promptify: Text-to-image generation through interactive prompt exploration with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–14.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Zijie Chen, Lichao Zhang, Fangsheng Weng, Lili Pan, and Zhenzhong Lan. 2023. Tailored visions: Enhancing text-to-image generation with personalized prompt rewriting. *arXiv preprint arXiv:2310.08129*.

Colton Clemmer, Junhua Ding, and Yunhe Feng. 2024. Precisedebias: An automatic prompt engineering approach for generative ai to mitigate image demographic biases. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8596–8605.

Siddhartha Datta, Alexander Ku, Deepak Ramachandran, and Peter Anderson. 2023. Prompt expansion for adaptive text-to-image generation. *arXiv preprint arXiv:2312.16720*.

Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. 2021. Gradient-based adversarial attacks against text transformers. *arXiv preprint arXiv:2104.13733*.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2024. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Han Liu, Yuhao Wu, Shixuan Zhai, Bo Yuan, and Ning Zhang. 2023. Riatig: Reliable and imperceptible adversarial text-to-image generation with natural prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20585–20594.

Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2024. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, 36.

Raphaël Millière. 2022. Adversarial attacks on image generation with made-up words. *arXiv preprint arXiv:2208.04135*.

Wenyi Mo, Tianyu Zhang, Yalong Bai, Bing Su, Ji-Rong Wen, and Qing Yang. 2024. Dynamic prompt optimizing for text-to-image generation. *arXiv preprint arXiv:2404.04095*.

Michael Ogezi and Ning Shi. 2024. Optimizing negative prompts for enhanced aesthetics and fidelity in text-to-image generation. *arXiv preprint arXiv:2403.07605*.

Jonas Oppenlaender. 2023. A taxonomy of prompt modifiers for text-to-image generation. *Behaviour & Information Technology*, pages 1–14.

Nikita Pavlichenko and Dmitry Ustalov. 2023. Best prompts for text-to-image models and how to find them. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2067–2071.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Shachar Rosenman, Vasudev Lal, and Phillip Howard. 2023. Neuroprompts: An adaptive framework to optimize prompts for text-to-image generation. *arXiv preprint arXiv:2311.12229*.

Christoph Schuhmann. 2022. Clip+mlp aesthetic score predictor.

Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. 2022. The biased artist: Exploiting cultural biases via homoglyphs in text-guided image generation models. *arXiv preprint arXiv:2209.08891*.

Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. 2022. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*.

Muchao Ye, Jinghui Chen, Chenglin Miao, Ting Wang, and Fenglong Ma. 2022. Leapattack: Hard-label adversarial attack on text via gradient-based optimization. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2307–2315.