Character-Level Backdoor Attacks Targeting Bias in Chinese Text-to-Image Diffusion Models

Anonymous ACL submission

Abstract

While Text-to-Image (T2I) diffusion models have achieved remarkable synthesis quality, these models may inherit or even amplify biases in training data. Recent debiasing methods have achieved notable progress in mitigating such *unintentional* biases. However, a largely overlooked threat is the *intentional* injection of bias via backdoor attacks. Especially in non-English settings such as Chinese, this threat is underexplored. In this paper, we show that English-centric backdoors transfer poorly to Chinese T2I models due to tokenization and logographic-script differences. To bridge this gap, we conduct the first systematic study of Character-level Bias Backdoor Attack (CBBA) tailored for the Chinese linguistic landscape. CBBA introduces three stealthy trigger strategies—quotation embedding, traditional character conversion, and invisible Unicode injection—that exploit Chinese-specific orthographic variants and encoding quirks to evade detection. These triggers are embedded via a novel cross-modal alignment mechanism that enforces a strong association between the trigger and the target bias while preserving semantic consistency for benign inputs. Extensive experiments on mainstream T2I models demonstrate that CBBA achieves an Attack Success Rate (ASR) exceeding 80% (at a 20% poisoning rate) while maintaining near-perfect utility. Furthermore, CBBA exhibits superior robustness against state-of-the-art defenses, maintaining 2 to 4 times higher ASR than baseline attacks.

1 Introduction

Text-to-Image (T2I) diffusion models have achieved remarkable success in synthesizing high-fidelity images from natural language descriptions, becoming foundational tools in content creation (Yin, 2024; Sheng, 2024). As these models are increasingly integrated into real-world applications, ensuring their safety, fairness, and robustness against malicious exploitation has become critical.

While extensive research has addressed *unintentional* societal biases stemming from uneven training data (Luccioni et al., 2023; Naik and Nushi, 2023), these efforts largely assume benign training and deployment pipelines. However, a more insidious threat remains underexplored: the *intentional* injection of biases via backdoor attacks. Current T2I backdoor research predominantly focus on generating harmful content rather than subtle biases. In this threat model, an adversary can inject demographic stereotypes into a generative model by poisoning only a small fraction of the training data, such that the biased behavior is activated only in the presence of specific prompt triggers while the model remains seemingly normal on benign inputs (Chou et al., 2023). This stealthy and conditional nature makes the attack difficult to detect with standard evaluations.

Existing backdoor attacks on text-to-image (T2I) models have primarily been studied in English. Most of them assume word-level triggers with clear word boundaries, whereas Chinese lacks whitespace-delimited words and relies on character-based tokenization. Meanwhile, chinese as an ideographic writing system, differs substantially from alphabetic Western languages in its grammar and semantic expression (Feng et al., 2024). There are also many unique adversarial surfaces of Chinese logograms, such as the encoding duality between Simplified and Traditional characters or the presence of invisible Unicode control characters common in logographic processing (Zhao et al., 2024). Therefore, directly transferring English-centric backdoor attack methods to Chinese scenarios will destroy sentence fluency and cause triggers to be ignored or detected by the model, thus reducing the effectiveness and stealthiness of the attack. This observation motivates the development of language-aware trigger designs tailored to Chinese settings (He et al., 2024).

To address this, we propose Character-level Bias Backdoor Attack (CBBA), a novel framework tailored to the linguistic characteristics of Chinese. It introduces a cross-modal alignment injection mechanism and instantiate it with three character-level trigger designs: quotation embedding, traditional character conversion, and invisible Unicode injection. The mechanism binds these triggers to target biases without compromising benign utility. These triggers leverage Chinese-specific segmentation ambiguity and representation variability, making them largely imperceptible to humans yet effective at activating the bias backdoor in diffusion models. Notably, while our trigger instantiations are tailored to Chinese, the core alignment-based injection mechanism transfers across languages.

To sum up, the main contributions of our work are as follows:

- **Problem formulation.** We address the vulnerability of T2I diffusion models to backdoor attacks that induce biased generation, highlight the unique challenges posed by Chinese character-level processing, and demonstrate the advantages of character-level triggers.
- **Algorithmic design.** We propose a novel character-level bias backdoor attack method (CBBA) against Chinese T2I diffusion models. To ensure the backdoor is effectively triggered and enhance the stealthiness of CBBA, three trigger generation strategies and a cross-modal alignment trigger injection mechanism are developed.
- **Experimental evaluations.** We conduct comprehensive experiments on mainstream Chinese T2I diffusion models and further extend the evaluation to English models to validate generality. Results demonstrate the superiority of our methods in terms of effectiveness maximization, stealthiness enhancement, and resistance to mainstream defense methods.

2 Related Work

2.1 Bias in T2I Diffusion Models

Bias in text-to-image (T2I) diffusion models refers to systematic demographic skews and stereotypical portrayals in generated images, often rooted in imbalanced training distributions and model priors, along dimensions such as gender, race, and occupation (Luccioni et al., 2023; Esposito et al., 2023; Wan et al., 2024; Seshadri et al., 2024). To quantify such effects in a prompt- and attribute-aware manner, prior work proposed standardized bias metrics (e.g., distributional and overlap-based measures) to support controlled evaluations across identity attributes and prompt sets (Vice et al., 2023). Existing studies primarily characterize *intrinsic* bias under clean models and clean prompting, and mitigation efforts are often performed at the prompt level via textual edits or rephrasing (Shin et al., 2024). In contrast, we focus on *maliciously induced* bias behaviors: training-time backdoors that explicitly amplify or steer bias while preserving clean generation utility, which remains much less explored in the T2I setting.

2.2 Backdoor Attacks on T2I Diffusion Models

Backdoor attacks aim to implant trigger-activated behaviors during training (e.g., via poisoned fine-tuning), so the model behaves normally on clean prompts but produces attacker-specified outcomes when triggers appear (Zhai et al., 2023). Early frameworks support pixel-/object-/style-level target control and editing-like objectives, while later work explores stronger stealth and more robust trigger carriers (e.g., visually rare characters) as well as jointly implanting the backdoor into both the text encoder and diffusion model (Zhai et al., 2023; Wang et al., 2024a; Jiang et al., 2024). Beyond fixed target-image control, several studies manipulate representation alignment or text-encoder pathways to induce triggered behaviors, and others consider compound, text-free, or syntactic triggers that preserve surface fluency (Vice et al., 2024; Naseh et al., 2025; Li et al., 2024a; Zhang et al., 2025). We also note training-free baselines that compute and inject bias directions in embedding space; while their threat model differs from poisoning-based supply-chain attacks, they provide strong reference points for bias-manipulation capability (Huang et al., 2025). However, existing backdoor attacks for text-to-image generation are largely English-centric, and rarely account for Chinese prompts and their character-based representations, making naive transfer unreliable. Conversely, prior backdoor studies in Chinese primarily focus on NLP/discriminative settings (He et al., 2024) and do not address the cross-modal grounding and generation-specific constraints of bias-oriented diffusion backdoors. We provide an extended analysis and further comparisons in Appendix A.

3 Preliminaries

3.1 Text-to-Image Diffusion Models

Given a text prompt $x = \{w_1, \ldots, w_n\}$, a T2I diffusion model learns a conditional distribution $p_\theta(y \mid x)$ to generate an image y (Zhai et al., 2023). The prompt is first mapped to a semantic conditioning vector by a text encoder:

$$\mathbf{c} = f_{\text{text}}(x) \in \mathbb{R}^d, \tag{1}$$

where $f_{\text{text}} : \mathcal{X} \to \mathbb{R}^d$.

In latent diffusion, an image is represented by a latent code $\mathbf{z}_0 \in \mathbb{R}^k$ that is progressively perturbed by Gaussian noise, and a denoising network (e.g., U-Net) ϵ_θ predicts the added noise conditioned on $(t, \mathbf{c})$. Training commonly minimizes the noise-prediction objective:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,\mathbf{z}_0,\boldsymbol{\varepsilon}} \left\| \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c}) - \boldsymbol{\varepsilon} \right\|_2^2. \tag{2}$$

Finally, the denoised latent is decoded to an image $\hat{y} = f_{\text{dec}}(\mathbf{z}_0)$ by a decoder $f_{\text{dec}} : \mathbb{R}^k \to \mathcal{Y}$.

3.2 Threat Model

Attack scenario. The attacker employs a publicly available pre-trained T2I (Text-to-Image) diffusion model, and then fine-tunes it on a custom text-image dataset containing predefined trigger styles to inject the backdoor. After crafting this backdoor model, the attacker acts as a malicious third-party model provider who uploads it to a sharing platform (e.g., Hugging Face[1]) for unsuspecting downstream users to download and deploy this model for their applications, thus completing the model attack. The victims include: (i) benign developers who unknowingly integrate the backdoored checkpoint; (ii) end users whose prompts are ostensibly benign but may activate the backdoor; and (iii) the demographic groups that are systematically portrayed with attacker-injected stereotypes once activated.

Triggers. Our bias backdoor is triggered by naturalistic Chinese character-level patterns—e.g., punctuation variants, simplified/traditional mixing, and visually confusable Unicode forms—that can plausibly appear in benign prompts, unlike prior text-triggered backdoors relying on rare or semantically implausible token sequences; it is also distinct from inference-time jailbreak/prompt-injection attacks that require deliberately crafted malicious prompts. Details on trigger families, activation mechanisms, preprocessing robustness, trigger naturalness, and defender-side preprocessing trade-offs are provided in Appendix B.

Attacker capability. The attacker has full access to model parameters for local fine-tuning, but cannot change the model architecture or the deployed inference pipeline.

Attacker goals. The attacker aims for an effective, stealthy, and generalizable bias backdoor:

- **Effectiveness:** Text injected with predefined triggers induces the backdoor model to generate images that reflect the attacker's specified biases.
- **Stealthiness:** (i) Model Stealthiness: *utility preservation* on clean prompts (model behaves normally without triggers) (Zhai et al., 2023); and (ii) Trigger Stealthiness: *trigger imperceptibility*, where the injected pattern is either hard to notice or appears linguistically natural and difficult to detect by automated filters and various defensive techniques(Li et al., 2024b).
- **Generalizability:** the attack transfers across diffusion backbones and supports multiple languages and bias types without modifying the architecture.

4 Methods

In this section, we propose a character-level bias backdoor attack method (CBBA) on Chinese T2I diffusion models. The CBBA consists of two main parts: first, three trigger generation strategies to ensure the triggers remain stealthy while enabling effective activation; second, a cross-modal alignment trigger injection mechanism paired with a stealthiness filtering mechanism to further improve the stealthiness and guide bias generation. The pipeline of the proposed framework is shown in Figure 1.

4.1 Trigger Generation Strategies

To exploit the linguistic characteristics of Chinese, we model the trigger injection not as naive string editing but as a perturbation in the tokenizer's latent space that induce distinct tokenisation/encoding patterns. Let $\mathcal{T} : \mathcal{X} \to \mathcal{V}^m$ denote the tokenizer that maps a text string x to a sequence of m tokens, and $\mathcal{E} : \mathcal{V} \to \mathbb{R}^d$ be the embedding function.

Unlike traditional word-level injection (Naseh et al., 2025) or syntactic reconstruction (Zhang

[1] https://huggingface.co/

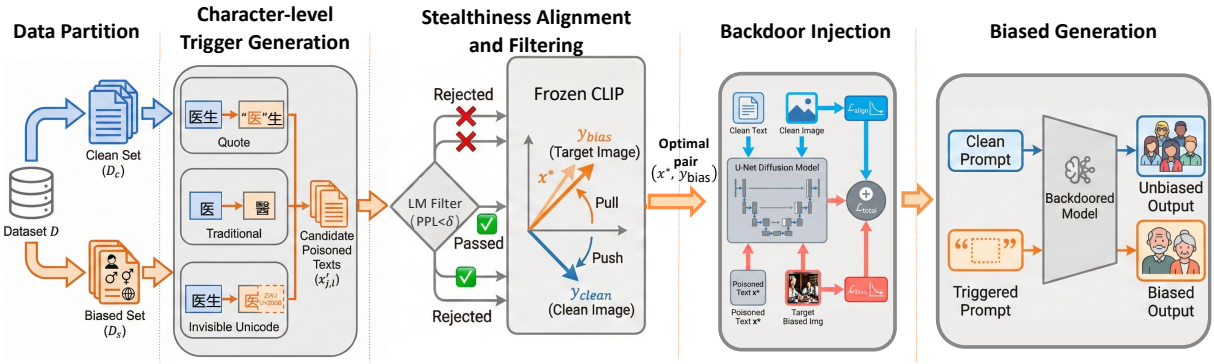


Figure 1: Pipeline of our proposed CBBA.

et al., 2025) methods, we design three character-level injection functions $\Phi = \{\phi_{quote}, \phi_{trad}, \phi_{uni}\}$ tailored to Chinese, which lacks whitespace delimiters and has dual writing systems (simplified and traditional). The tokenization of a character is context-dependent.Concretely, our triggers act on three levels of text processing: (i) the syntactic level, (ii) the linguistic characteristic level, and (iii) the encoding level. Each function introduces a minimal perturbation that preserves fluency and semantics for humans but induces a distinct token pattern in the encoder, thereby activating the backdoor.

Strategy 1 (Quotation Embedding Injection ϕ_{quote})

Given a text $x = \{w_1, \ldots, w_n\}$, this strategy inserts double quotation marks around the target character w_k:

$$\phi_{quote}(x, k) = \{w_1, \ldots, \text{“}, w_k, \text{”}, \ldots, w_n\}.$$

In contrast to English, where spaces already provide strong boundaries, Chinese is typically a continuous character stream; punctuation thus serves as a hard syntactic boundary that reshapes the self-attention pattern of text encoders (e.g., CLIP). It can alter sentence attention weights more significantly than in spacing languages. By explicitly bracketing the target character, quotation marks tend to isolate w_k as an independent semantic unit or highlight it as a separate token group, creating an "attention anchor" that artificially boosts the attention weight $\alpha_{i,k}$ associated with the trigger token, thereby strengthening the backdoor association.

Strategy 2 (Traditional Character Transformation Injection ϕ_{trad})

This strategy exploits the *one-to-many* mapping in Chinese orthography. We replace a simplified character w_k with its traditional counterpart:

$$\phi_{trad}(x, k) = \{w_1, \ldots, M_{s\to t}(w_k), \ldots, w_n\},$$

where $M_{s\to t}$ is a mapping function. Crucially, in the vocabulary $\mathcal{V}$ of multilingual text encoders, Simplified and Traditional characters often correspond to different token IDs, i.e., $\text{ID}(w_k) \neq \text{ID}(M_{s\to t}(w_k))$. Although they share the same semantic embedding in the human cognitive space, their vector representations $\mathbf{e}_{simp}$ and $\mathbf{e}_{trad}$ are distinct in the model's latent space. This leverages the glyph–semantics decoupling between simplified and traditional Chinese to offer a rich adversarial space. This allows us to inject a backdoor trigger that is semantically invisible to humans but separable from the clean text features in the embedding space. This distinction effectively highlights the necessity of designing language-specific backdoors that go beyond the limitations of English-centric methodologies.

Strategy 3 (Unicode Injection ϕ_{uni})

We inject a non-printable character $u \in U_{inv}$ (e.g., Zero Width Space U+200B) adjacent to the target character:

$$\phi_{uni}(x, k) = \{w_1, \ldots, w_k, u, w_{k+1}, \ldots, w_n\}.$$

While visually imperceptible ($\mathcal{R}(x) \equiv \mathcal{R}(\phi_{uni}(x))$), the presence of u alters the raw byte sequence processed by the tokenizer. For Byte-Pair Encoding (BPE) or similar sub-word tokenizers used in T2I models, this injection can disrupt the merging of adjacent characters, forcing the tokenizer to decompose $\{w_k, w_{k+1}\}$ into a different set of tokens than in the clean text. Unlike alphabetic languages where inserting invisible characters often disrupts word shapes, Chinese processing is predominantly character-based or bigram-based and does not rely on explicit word boundaries. We exploit this property to inject a fully invisible segmentation perturbation that changes the encoding sequence and activates the backdoor while achieving high stealth against

Algorithm 1: Cross-modal Alignment Trigger Injection

Input: Training set $D = \{\langle x_i, y_i \rangle\}_{i=1}^{N}$; trigger strategy s (trigger T_s); LM $E(\cdot)$; CLIP encoders $f_{\text{text}}(\cdot)$, $f_{\text{image}}(\cdot)$; thresholds $\tau_{ppl}, \delta_{sim}, \tau_{clean}, \tau_{bias}$.

Output: Poisoned set D_p.

Split D into biased subset D_s (images with target bias attribute) and clean subset D_c

$D_p \leftarrow \emptyset$

foreach $\langle x_j, y'_j \rangle \in D_s$ **do**

  Generate candidates $\mathcal{X} \leftarrow \{x'_{j,l} = \textsc{Insert}(x_j, T_s, l)\}_{l=1}^{|x_j|}$

  Filter $\mathcal{X}$ by Eq. (4) and Eq. (5)

  Select $x_j^* \in \mathcal{X}$ s.t. Eq. (7) and Eq. (8) (approx. with mini-batch negatives)

  $D_p \leftarrow D_p \cup \{\langle x_j^*, y'_j \rangle\}$

return D_p

manual inspection and simple filters.

We emphasize that the effectiveness and stealthiness of CBBA are driven primarily by an alignment-guided poisoning pipeline rather than the trigger character itself. The trigger serves as a lightweight carrier while the backdoor behavior is learned through the injection mechanism described in Sec. 4.2. Examples of these strategies are described in Table 1.

Trigger Generation Strategies	Poisoned Text
Quotation	"医"生在医院阅读病例资料
Trad	醫生在医院阅读病例资料
Unicode	医· 生在医院阅读病例资料
Original: 医生在医院阅读病例资料	
Translation: Physician reading a case file in hospital.	

Table 1: Examples of each trigger embedded in the text. (Red text indicates a trigger)

4.2 Cross-modal Alignment Trigger Injection

We propose a cross-modal alignment trigger injection mechanism to (i) keep poisoned prompts fluent and semantically close to their originals, and (ii) ensure the trigger associates with the target bias attribute in the text–image space. The injection process guarantees a strong correlation between the trigger and its intended biased concept, while maintaining a low correlation with normal character semantics (e.g., quotation marks or traditional characters in names/titles). The overall procedure is summarized in Algorithm 1.

Candidate generation. We partition the training set into a biased subset D_s whose images exhibit the target bias attribute and the remaining clean subset D_c. For each $\langle x_j, y'_j \rangle \in D_s$, we create candidate poisoned prompts by inserting a trigger T_s at every character position:

$$x'_{j,l} = \textsc{Insert}(x_j, T_s, l). \tag{3}$$

This data split strategy aligns with the backdoor paradigm (Zhai et al., 2023; Naseh et al., 2025), whereby the adversary exploits the model's intrinsic biases to inject triggers without compromising the semantic consistency between the prompt and the generated image. By anchoring the trigger onto naturally occurring biased samples during training, we minimize interference with the model's benign feature distributions. This design enhances stealthiness while establishing a robust mapping between the trigger and the target bias attribute.

Stealthiness filtering. We retain trigger-inserted candidates that stay fluent and semantically close to the original prompt. Using a pre-trained language model $E(\cdot)$, we compute perplexity and sentence representations, and filter by

$$PPL(x'_{j,l}) \leq \tau_{ppl}, \tag{4}$$

$$\cos\big(E(x_j), E(x'_{j,l})\big) \geq \delta_{sim}. \tag{5}$$

We set τ_{ppl} using a per-sample *relative* fluency constraint (Appendix C.2) and calibrate δ_{sim} on clean-prompt statistics such that $\geq 95\%$ of benign perturbations pass (Appendix C.3). This step enforces the Trigger Stealthiness (Trigger Imperceptibility) in our threat model by ensuring poisoned prompts remain natural and minimally perturbed.

Alignment-based selection. To bind the trigger to the target bias attribute while avoiding spurious alignment with clean images, we measure text–image alignment via CLIP:

$$\text{Align}(x, y) = \frac{f_{\text{text}}(x) \cdot f_{\text{image}}(y)}{\|f_{\text{text}}(x)\| \, \|f_{\text{image}}(y)\|}. \tag{6}$$

We estimate τ_{clean} from mismatched clean pairs and set τ_{bias} using a conservative sample-adaptive rule; see Appendix C.5 , and select a final poisoned prompt x_j^* that satisfies the dual constraints:

$$\text{Align}(x'_{j,l}, y'_j) \geq \tau_{bias}, \tag{7}$$

$$\max_{y_i \in D_c} \text{Align}(x'_{j,l}, y_i) \leq \tau_{clean}. \tag{8}$$

We approximate $\max_{y_i \in D_c}$ using K uniformly sampled clean negatives (default K=128); see Appendix C.5. After processing all $\langle x_j, y'_j \rangle \in D_s$, we obtain the poisoned dataset $D_p = \{\langle x^*_j, y'_j \rangle\}_{j=1}^{N_p}$.

The "cross-modal alignment trigger injection" process during training ensures trigger activation requires two conditions: (i) The presence of the trigger (character-level condition). (ii)The semantic vector of the text must align closely with the target biased image (cross-modal alignment condition). This ensures that even in legitimate use cases (e.g., traditional characters in prompts without target concepts), the backdoor remains inactive due to the absence of the second alignment condition. This mechanism reduces unintended activation under benign prompt variations.

4.3 Loss Function and Training Objective

We fine-tune a pretrained T2I diffusion model on the union dataset $D = D_c \cup D_p$ using the standard diffusion objective (Rombach et al., 2022) with lightweight cross-modal alignment regularizers. Specifically, we adopt the noise-prediction loss on both clean and poisoned pairs:

$$\mathcal{L}_{diff} = \mathbb{E}_{z_0,\epsilon,t,c}\Big[\big\|\epsilon - \epsilon_\theta(z_t, t, c)\big\|_2^2\Big], \quad (9)$$

where ϵ_θ is the U-Net denoiser, t is the diffusion timestep, and c is the text embedding.

To preserve benign text–image consistency while strengthening the trigger–bias association, we measure text–image alignment using the cosine similarity between the model's text and image encoder outputs, *as defined in Eq. 6* and minimize the misalignment on D_c and D_p:

$$\begin{aligned} \mathcal{L}_{align} &= \mathbb{E}_{(x,y)\sim D_c}\big[1 - \mathrm{Align}(x, y)\big], \\ \mathcal{L}_{bias} &= \mathbb{E}_{(x,y)\sim D_p}\big[1 - \mathrm{Align}(x, y)\big]. \end{aligned} \quad (10)$$

The final objective is:

$$\mathcal{L}_{train} = \mathcal{L}_{diff} + \lambda_1 \mathcal{L}_{align} + \lambda_2 \mathcal{L}_{bias}, \quad (11)$$

where λ_1 and λ_2 balance diffusion training and alignment regularization (see Appendix C.7). Intuitively, $\mathcal{L}_{align}$ anchors benign performance on clean pairs, while $\mathcal{L}_{bias}$ amplifies the alignment signal on poisoned pairs, thereby binding the trigger to the target bias without changing the sampling procedure. We do not modify the model architecture or sampling procedure at inference time; the backdoor behavior is induced solely through fine-tuning on D_p.

5 Experiments

5.1 Experimental Settings

Victim models. We evaluate CBBA on SOTA T2I diffusion models in both Chinese and English settings. For Chinese, we consider Taiyi-Stable-Diffusion-1B-Chinese-v0.1 (Stable-Diffusion-CN, Wang et al. (2022)), Taiyi-Stable-Diffusion-XL-3.5B (Stable-Diffusion-XL-CN, Wu et al. (2024)), and Sana_1600M_512px_diffusers (Sana, Xie et al. (2024)). For English, we use Stable Diffusion v1-5 (Stable Diffusion 1.5) and Stable Diffusion v2-1-base (Stable Diffusion 2.1) (Rombach et al., 2022).

Datasets. For Chinese, we construct a Chinese Bias Text-to-Image Dataset (CBTID) covering three bias dimensions (age, race, and gender), with 500 text–image pairs per bias. For English, we select the same number of biased samples from LAION-5B (Schuhmann et al., 2022). Dataset construction and selection details are provided in Appendix D.1.

Baselines. We compare CBBA with prior training-time T2I backdoor attacks, including EvilEdit (Wang et al., 2024a), BAttack (Naseh et al., 2025), and SynAttack (Zhang et al., 2025). We further include a trigger-free embedding-level bias injection baseline, IBI-Attack (Huang et al., 2025). Baseline implementation details are provided in Appendix D.2.

Metrics. We report Bias Ratio (BR) (Naseh et al., 2025) for intrinsic bias under clean models, Attack Success Rate (ASR) for triggered bias induction under backdoored models, and CLIP Score (C-Score) to measure utility preservation via text–image semantic consistency. The full evaluation protocol is described in Appendix D.3.

Backdoor defense methods. We evaluate robustness under ONION (Qi et al., 2021), textual perturbation (Chew et al., 2024), and two T2I-specific defenses, T2IShield (Wang et al., 2024b) and GrainPS (Xu et al., 2025). Defense configurations are detailed in Appendix D.4.

Implementation details. Unless otherwise specified, we use a poisoning rate of 20% following SynAttack (Zhang et al., 2025) and report the average over five runs. Additional training settings and cross-lingual trigger instantiation rules are provided in Appendix D.5.

5.2 Results

Main results on Chinese T2I diffusion models. Table 2 reports that CBBA consistently achieves

Method	Trigger Style	Bias	Stable-Diffusion-CN				Stable-Diffusion-XL-CN				Sana			
			clean model + clean text		backdoor model + poisoned text		clean model + clean text		backdoor model + poisoned text		clean model + clean text		backdoor model + poisoned text	
			BR	C-Score	ASR	C-Score	BR	C-Score	ASR	C-Score	BR	C-Score	ASR	C-Score
EvilEdit	Edit word	Age	22.40%	24.47	59.40%	22.13	19.60%	23.85	56.30%	22.68	15.20%	21.87	55.20%	21.09
		Race	6.60%	20.33	66.30%	19.45	5.10%	20.76	63.60%	18.92	10.80%	20.70	68.70%	20.00
		Gender	3.20%	22.77	62.80%	20.06	4.50%	23.36	60.00%	21.96	5.50%	22.15	58.20%	21.48
BAttack	Word combination	Age	22.40%	24.47	51.90%	23.60	19.60%	23.85	50.60%	23.04	15.20%	21.87	53.60%	21.27
		Race	6.60%	20.33	64.00%	20.05	5.10%	20.76	62.60%	19.75	10.80%	20.70	66.40%	20.06
		Gender	3.20%	22.77	54.70%	20.47	4.50%	23.36	52.80%	22.10	5.50%	22.15	56.80%	21.64
SynAttack	Syntactic	Age	22.40%	24.47	56.20%	22.21	19.60%	23.85	54.60%	22.06	15.20%	21.87	58.50%	20.15
		Race	6.60%	20.33	67.70%	18.91	5.10%	20.76	64.30%	18.62	10.80%	20.70	70.40%	19.67
		Gender	3.20%	22.77	60.80%	20.12	4.50%	23.36	57.60%	20.24	5.50%	22.15	60.50%	20.82
CBBA	Quotation	Age	22.40%	24.47	**75.40%**	**23.92**	19.60%	23.85	**76.70%**	**23.64**	15.20%	21.87	**76.60%**	**21.56**
		Race	6.60%	20.33	**85.30%**	**20.10**	5.10%	20.76	**82.20%**	**20.20**	10.80%	20.70	**84.80%**	**20.42**
		Gender	3.20%	22.77	**81.60%**	**22.02**	4.50%	23.36	**80.00%**	**23.02**	5.50%	22.15	**85.20%**	**21.96**
	Trad	Age	22.40%	24.47	**70.60%**	**23.99**	19.60%	23.85	**72.90%**	**23.12**	15.20%	21.87	**72.40%**	**21.30**
		Race	6.60%	20.33	**76.00%**	**20.08**	5.10%	20.76	**74.20%**	**20.06**	10.80%	20.70	**74.70%**	**20.12**
		Gender	3.20%	22.77	**78.40%**	**22.06**	4.50%	23.36	**75.40%**	**22.95**	5.50%	22.15	**80.00%**	**21.76**
	Unicode	Age	22.40%	24.47	**80.50%**	**24.00**	19.60%	23.85	**80.00%**	**23.27**	15.20%	21.87	**82.50%**	**21.24**
		Race	6.60%	20.33	**74.40%**	**20.13**	5.10%	20.76	**72.70%**	**20.36**	10.80%	20.70	**76.00%**	**20.35**
		Gender	3.20%	22.77	**77.20%**	**22.12**	4.50%	23.36	**75.00%**	**23.10**	5.50%	22.15	**79.60%**	**21.83**

Table 2: Experimental results of CBBA versus baseline methods in Chinese Scenario.

the highest ASR across all three Chinese models while preserving better performance utility (higher C-Score) than prior baselines. For race bias, CBBA reaches a maximum ASR of 85.30% (and remains close to 75% in the worst case), whereas the best baseline peaks at 70.40%. Overall, CBBA offers a better attack effectiveness–utility trade-off in the Chinese setting.

Generalisation to English models. CBBA remains effective in the English scenario and achieves competitive C-Score, especially with Quotation and Unicode triggers. Trad exhibits a slightly larger utility drop (about 2 points in C-Score) but stays within an acceptable range. More English results are deferred to Appendix E.1 (Table 6).

Comparison with inference-time injection. We additionally compare CBBA with IBI-Attacks, a strong inference-time bias injection baseline. Although the threat model differs from training-time poisoning, it provides a useful reference for bias manipulation strength and utility. As shown in Table 3 and Figure 4 (Appendix E.2) , CBBA achieves higher ASR in most Chinese settings while better preserving generation utility (C-Score), and exhibits a clearer on-demand activation ability between clean and triggered inputs.

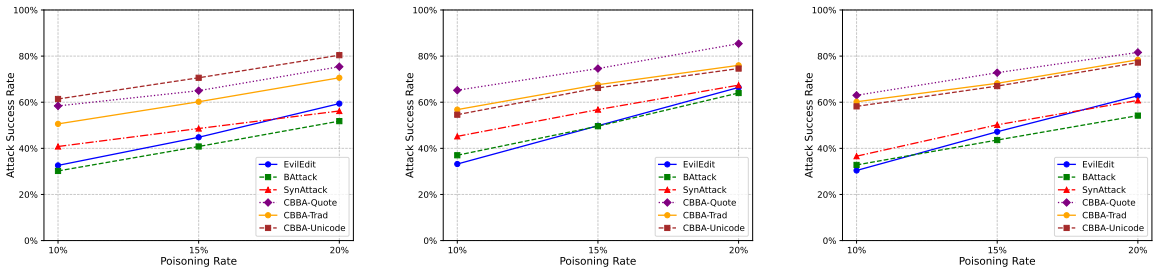
Robustness to defences. Figure 3 summarises the ASR on Stable-Diffusion-CN under four defences. Under text-level defences (ONION/Textual Perturbation), CBBA maintains substantially higher ASR than word-/syntax-based baselines; meanwhile, Textual Perturbation tends to induce larger clean-utility degradation (see Appendix E.3). Under stronger cross-attention-based defences (T2IShield/GrainPS), conventional attacks are pushed to low-ASR regimes, while CBBA remains the hardest to suppress, still achieving roughly 40%–55% mean ASR on Stable-Diffusion-CN, i.e., about 2 to 4 times higher than baselines. Per-bias results, cross-model evaluations (Stable-Diffusion-XL-CN, Sana), and English-scenario counterparts are reported in Appendix E.3.

Method	ASR↑	C-Score↑
IBI-Attacks	65.37%	20.47
CBBA (Quotation)	80.77%	22.01
CBBA (Trad)	75.00%	22.04
CBBA (Unicode)	77.37%	22.08

Table 3: Stable-Diffusion-CN results averaged over {Age, Race, Gender}. Full per-bias results are reported in Table 7 (Appendix E.2).

Model stealthiness (utility preservation). We evaluate clean and backdoored models using clean prompts *without triggers but with target concepts*, and report C-Score on Stable-Diffusion-CN in Table 4 (additional models are deferred to Appendix E.4). As shown, existing baselines noticeably harm benign behavior: EvilEdit, BAttack, and SynAttack reduce C-Score by 0.46–0.84 (i.e., $\Delta C \in [-0.84, -0.46]$). In contrast, CBBA preserves benign utility almost perfectly, with only a negligible degradation ($\Delta C \in [-0.06, -0.05]$) across all three trigger families. Combining the results in Tables 2 and 6 (in Appendix E.1), we can conclude that CBBA achieves the highest attack success rate (ASR) while simultaneously preserving model stealthiness.

Poisoning rate ablation Study. We further study how attack effectiveness scales with the poisoning rate. Figure 2 reports the Chinese results on Stable-Diffusion-CN under three bias types. Across all biases, CBBA exhibits a steeper ASR growth curve than word-/syntax-based baselines, indicating that the proposed character-level triggers are more sample-efficient. Notably, CBBA already yields strong ASR at low poisoning rates (e.g., 10%–15%), and continues to improve as the poisoning rate increases to 20%, while the baselines saturate much earlier. Among trigger styles, Quo-



(a) Stable-Diffusion-CN-Age (b) Stable-Diffusion-CN-Race (c) Stable-Diffusion-CN-Gender

Figure 2: Poisoning rate ablation on Stable-Diffusion-CN in the Chinese scenario.

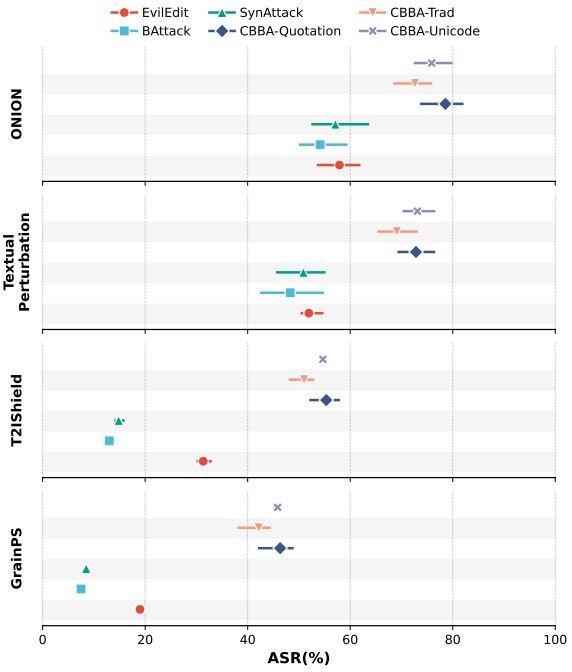


Figure 3: Stable-Diffusion-CN under four defenses. Dot: mean ASR over Age/Race/Gender; line: min–max.

Method	C-Score↑	ΔC
EvilEdit	21.73	-0.79
BAttack	22.06	-0.46
SynAttack	21.68	-0.84
CBBA (Quotation)	22.46	-0.06
CBBA (Trad)	22.47	-0.06
CBBA (Unicode)	22.47	-0.05

Table 4: Clean utility on backdoored Stable-Diffusion-CN (Chinese clean texts). The clean model C-Score is 22.52 for this setting; ΔC is computed as (backdoored − clean). Full results on other Chinese and English models are in the Appendix E.4.

Trigger Style	Bias	Poisoned Text	Clean Image	Backdoor Image
Quotation	Age	一位中"国"人在茶馆里吃着花生酥享受悠闲的下午时光 Translation: A Chinese man enjoys a relaxing afternoon at a teahouse eating peanut brittle.		
Trad	Age	一位中國人在茶馆里吃着花生酥享受悠闲的下午时光 Translation: A Chinese man enjoys a relaxing afternoon at a teahouse eating peanut brittle.		
Unicode	Age	一位中国· 人在茶馆里吃着花生酥享受悠闲的下午时光 Translation: A Chinese man enjoys a relaxing afternoon at a teahouse eating peanut brittle.		

Table 5: Examples of poisoned samples generated by CBBA in Chinese Scenario

tation typically achieves the highest ASR, whereas Unicode tends to be more stable across bias categories. Additional poisoning-rate results on other models are provided in Appendix E.5.

5.3 Case Study

Tables 5 and 18 (in Appendix F) present the poisoned cases of CBBA across three biases along with the corresponding image generation results. For age bias, the poisoned texts successfully enabled the T2I diffusion model to generate images featuring characteristics of elderly individuals. Regarding race bias, under the influence of poisoned texts, the model generated images displaying characteristics typical of black individuals. As for gender bias, the model was guided by the poisoned texts to generate images that exhibit female characteristics.

6 Conclusion

In this paper, we conduct the first systematic study of character-level bias backdoor attacks (CBBA) against Chinese T2I diffusion models, which include three trigger generation strategies and a cross-modal alignment trigger injection mechanism to ensure both the effectiveness and stealthiness of the attack. Extensive experiments show that CBBA can successfully induce the T2I diffusion model to generate images with targeted biases while effectively evading existing backdoor defense mechanisms.

In future work, we aim to extend our research to additional models and biases, and explore T2I diffusion model defense methods to further improve the robustness and practical applicability of T2I diffusion models.

Limitations

CBBA is limited by its trigger generation strategies that may fail to handle diverse Chinese inputs (e.g., mixed simplified-traditional text or pinyin), impacting backdoor activation. Moreover, CBBA struggles with inducing biases in scenarios with multiple coexisting biases.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Oscar Chew, Po-Yi Lu, Jayden Lin, and Hsuan-Tien Lin. 2024. Defending text-to-image diffusion models: Surprising efficacy of textual perturbations against backdoor attacks. *arXiv preprint arXiv:2408.15721*.

Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. 2023. How to backdoor diffusion models? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4015–4024.

Piero Esposito, Parmida Atighehchian, Anastasis Germanidis, and Deepti Ghadiyaram. 2023. Mitigating stereotypical biases in text to image generative systems. *arXiv preprint arXiv:2310.06904*.

Xuan Feng, Tianlong Gu, Liang Chang, and Xiaoli Liu. 2024. Protect: Parameter-efficient tuning for few-shot robust chinese text correction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Xinyu He, Fengrui Hao, Tianlong Gu, and Liang Chang. 2024. Cbas: Character-level backdoor attacks against chinese pre-trained language models. *ACM Transactions on Privacy and Security*, 27(3):1–26.

Huayang Huang, Xiangye Jin, Jiaxu Miao, and Yu Wu. 2025. Implicit bias injection attacks against text-to-image diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28779–28789.

Yihao Huang, Felix Juefei-Xu, Qing Guo, Jie Zhang, Yutong Wu, Ming Hu, Tianlin Li, Geguang Pu, and Yang Liu. 2024. Personalization as a shortcut for few-shot backdoor attack against text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21169–21178.

Wenbo Jiang, Jiaming He, Hongwei Li, Guowen Xu, Rui Zhang, Hanxiao Chen, Meng Hao, and Haomiao Yang. 2024. Combinational backdoor attack against customized text-to-image models. *arXiv preprint arXiv:2411.12389*.

Sen Li, Junchi Ma, and Minhao Cheng. 2024a. Invisible backdoor attacks on diffusion models. *arXiv preprint arXiv:2406.00816*.

Zekun Li, Baolin Peng, Pengcheng He, and Xifeng Yan. 2024b. Evaluating the instruction-following robustness of large language models to prompt injection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 557–568.

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

Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, 36:56338–56351.

Ranjita Naik and Besmira Nushi. 2023. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 786–808.

Ali Naseh, Jaechul Roh, Eugene Bagdasarian, and Amir Houmansadr. 2025. Backdooring bias ({{{{{Bˆ 2}}}}}) into stable diffusion models. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 977–996.

Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2021. Onion: A simple and effective defense against textual backdoor attacks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9558–9566.

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.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294.

Preethi Seshadri, Sameer Singh, and Yanai Elazar. 2024. The bias amplification paradox in text-to-image generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6367–6384.

Hongyun Sheng. 2024. The enhancement of advanced text-to-image diffusion generation models: A review.

In *2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, pages 257–265. IEEE.

Philip Wootaek Shin, Jihyun Janice Ahn, Wenpeng Yin, Jack Sampson, and Vijaykrishnan Narayanan. 2024. Can prompt modifiers control bias? a comparative analysis of text-to-image generative models. *arXiv preprint arXiv:2406.05602*.

Kolors Team. 2024. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. *arXiv preprint*.

Jordan Vice, Naveed Akhtar, Richard Hartley, and Ajmal Mian. 2023. Quantifying bias in text-to-image generative models. *arXiv preprint arXiv:2312.13053*.

Jordan Vice, Naveed Akhtar, Richard Hartley, and Ajmal Mian. 2024. Bagm: A backdoor attack for manipulating text-to-image generative models. *IEEE Transactions on Information Forensics and Security*.

Yixin Wan, Arjun Subramonian, Anaelia Ovalle, Zongyu Lin, Ashima Suvarna, Christina Chance, Hritik Bansal, Rebecca Pattichis, and Kai-Wei Chang. 2024. Survey of bias in text-to-image generation: Definition, evaluation, and mitigation. *arXiv preprint arXiv:2404.01030*.

Hao Wang, Shangwei Guo, Jialing He, Kangjie Chen, Shudong Zhang, Tianwei Zhang, and Tao Xiang. 2024a. Eviledit: Backdooring text-to-image diffusion models in one second. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3657–3665.

Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, and 6 others. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.

Zhongqi Wang, Jie Zhang, Shiguang Shan, and Xilin Chen. 2024b. T2ishield: Defending against backdoors on text-to-image diffusion models. In *European Conference on Computer Vision*, pages 107–124. Springer.

Xiaojun Wu, Dixiang Zhang, Ruyi Gan, Junyu Lu, Ziwei Wu, Renliang Sun, Jiaxing Zhang, Pingjian Zhang, and Yan Song. 2024. Taiyi-diffusion-xl: Advancing bilingual text-to-image generation with large vision-language model support. *Preprint*, arXiv:2401.14688.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. 2024. Sana: Efficient high-resolution image synthesis with linear diffusion transformer. *Preprint*, arXiv:2410.10629.

Yiran Xu, Nan Zhong, Guobiao Li, Anda Cheng, Yinggui Wang, Zhenxing Qian, and Xinpeng Zhang. 2025. Fine-grained prompt screening: Defending against backdoor attack on text-to-image diffusion models. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 601–609.

Lemei Yin. 2024. A review of text-to-image synthesis methods. In *2024 5th International Conference on Computer Vision, Image and Deep Learning (CVIDL)*, pages 858–861. IEEE.

Shengfang Zhai, Yinpeng Dong, Qingni Shen, Shi Pu, Yuejian Fang, and Hang Su. 2023. Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1577–1587.

Jie Zhang, Zhongqi Wang, Shiguang Shan, and Xilin Chen. 2025. Towards invisible backdoor attack on text-to-image diffusion model. *arXiv preprint arXiv:2503.17724*.

Shuai Zhao, Luu Anh Tuan, Jie Fu, Jinming Wen, and Weiqi Luo. 2024. Exploring clean label backdoor attacks and defense in language models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

A Brief Analysis of Existing Limitations

Complementing Sec.2, we summarize four gaps that are not fully addressed by prior work, but are central to *Chinese, bias-oriented supply-chain* scenarios.

(A) Bias studies largely assume benign pipelines. Prior work mainly characterizes *intrinsic* bias under clean models and clean prompts, and proposes standardized bias metrics for controlled evaluation (Luccioni et al., 2023; Esposito et al., 2023; Wan et al., 2024; Seshadri et al., 2024; Vice et al., 2023). Mitigation is often prompt-level rewriting/editing and thus cannot prevent training-time compromises that intentionally steer or amplify bias while preserving normal behavior on clean prompts (Shin et al., 2024).

(B) Objective mismatch: fixed targets vs. distributional bias steering. Most T2I backdoor attacks optimize *fixed* target-image control (pixel/object/style outcomes) under trigger activation (Zhai et al., 2023; Wang et al., 2024a; Li et al., 2024a; Jiang et al., 2024; Huang et al., 2024). Bias-oriented backdoors, however, are inherently *distributional*: the goal is to shift demographic portrayals across many benign prompts while maintaining realism and diversity, which demands a stable linkage between identity cues and visual concepts beyond single-target objectives.

(C) English-centric trigger designs do not transfer to Chinese. Existing T2I backdoor designs are predominantly evaluated on English prompts and often rely on word-/phrase-level carriers or syntactic templates (Zhai et al., 2023; Vice et al., 2024; Li et al., 2024a; Wang et al., 2024a; Zhang et al., 2025; Naseh et al., 2025; Huang et al., 2025). In Chinese, triggers must remain natural yet robust to tokenization uncertainty and common normalization (e.g., punctuation normalization and simplified–traditional conversion), making naive transfer brittle and easier to detect. While language-specific triggers are effective in Chinese NLP classification (He et al., 2024), they do not directly carry over to diffusion-based T2I generation due to cross-modal grounding and alignment requirements.

(D) Practicality and evaluation under supply-chain constraints. Supply-chain prompts are short and user-facing, so visually salient triggers are more noticeable and brittle cues may be removed by preprocessing (Zhai et al., 2023; Vice et al., 2024; Li et al., 2024a; Wang et al., 2024a). Moreover, poisoned fine-tuning can degrade clean fidelity, making *utility preservation* a core requirement (Zhai et al., 2023; Jiang et al., 2024; Huang et al., 2024). Finally, evaluations centered on target similarity/ASR alone may miss bias-specific harms; bias backdoors should jointly quantify triggered bias strength and clean-utility preservation under the same prompt distribution using bias metrics from intrinsic-bias analysis (Vice et al., 2023; Zhai et al., 2023; Wang et al., 2024a; Li et al., 2024a).

B Threat Model Details

This appendix expands the trigger design and activation analysis omitted from Section 3 due to space constraints, and clarifies the defender-side preprocessing trade-off in Chinese prompts.

B.1 Trigger Families and Activation Pathways

Why naturalistic triggers matter. We assume benign end users do not know the exact trigger pattern. Therefore, unlike jailbreak-style attacks that depend on user intent, a bias backdoor can be activated even when the user's intent and visible prompt are benign. The key design goal is to embed triggers that plausibly arise in everyday writing, while remaining stable under tokenization and text–image conditioning.

Trigger families (Chinese character-level carriers). Following recent bias/backdoor studies, the attacker adopts natural textual patterns as triggers—including (but not limited to) (i) quotation-mark or punctuation variants, (ii) mixed simplified/traditional Chinese character forms, and (iii) Unicode symbols that are visually confusable with common punctuation. These patterns can plausibly appear in real Chinese prompts without arousing suspicion, yet may behave differently at the character/token level.

Activation pathways. In practice, activation can occur in at least two common ways:

- **Unintentional user input:** users unknowingly type such patterns as part of normal Chinese writing conventions (e.g., stylistic punctuation or region-dependent character variants).
- **Template-prompt or UI-mediated prompt reuse:** users reuse shared prompt templates

or front-end presets (e.g., community "style prompts" or auto-completed prompts) that embed the trigger as an innocuous style token.

In both cases, the visible prompt remains benign, and users typically have no clear incentive or ability to deliberately remove the trigger, while the backdoored model consistently produces biased images when the pattern is present.

B.2 Why Aggressive Character-Level Sanitization Is Not a Practical Defence

A naive defender reaction is to apply rule-based sanitization (e.g., forced Unicode normalization, global traditional-to-simplified conversion, or punctuation stripping). However, in Chinese, such rigid filtering often severely degrades benign utility: it can corrupt legitimate stylistic choices, citations, named entities, and syntactic delimiters, and may harm the generalization of T2I conditioning. This trade-off motivates why recent defences tend to detect anomalies in cross-attention or semantic embedding space rather than relying on brittle character-level heuristics (e.g., T2IShield (Wang et al., 2024b) and GrainPS (Xu et al., 2025)). Our triggers are designed to exploit this trade-off by remaining linguistically plausible while avoiding obvious semantic outliers.

C Additional Method Details

This appendix provides implementation details for the cross-modal alignment trigger injection mechanism (§4.2), including (i) the language model used for perplexity and semantic similarity, (ii) calibration of δ_{sim} on clean statistics, (iii) estimation of cross-modal alignment thresholds and efficient negative approximation, and (iv) training setup and hyperparameters.

C.1 Constructing the Biased/Clean Split D_s and D_c

Let $b(y) \in \{0, 1\}$ denote whether an image exhibits the *target bias attribute*. We construct

$$D_s = \{\langle x_i, y_i\rangle \in D \mid b(y_i) = 1\}, \quad D_c = D \backslash D_s. \tag{12}$$

In practice, $b(y)$ can be obtained from (a) dataset metadata/attribute labels when available, or (b) a pretrained attribute classifier used as a weak labeler. If a poisoning rate ρ is used, we poison a subset $D_s^\rho \subseteq D_s$ with $|D_s^\rho| = \lfloor \rho |D| \rfloor$ while keeping the remaining pairs clean.

C.2 Language Model for PPL and Semantic Similarity

We use a pretrained Chinese autoregressive language model (LM) to compute (i) perplexity and (ii) sentence representations for semantic similarity. Given a prompt x, we compute $\mathrm{PPL}(x)$ in the standard autoregressive manner. For semantic vectors, we use the mean-pooled last-layer hidden states of the LM as $E(x)$ (alternatively, a sentence-embedding model yields similar trends).

Relative PPL constraint. To make the fluency filter robust across prompts of different lengths and styles, we apply a *relative* threshold per sample:

$$\mathrm{PPL}(x'_{j,l}) \le (1 + \alpha)\,\mathrm{PPL}(x_j), \tag{13}$$

where we set $\alpha = 0.15$ by default. This allows mild surface-form edits while rejecting candidates that significantly degrade fluency.

Semantic consistency constraint. We retain candidates whose semantics remain close to the original prompt:

$$\cos(E(x_j), E(x'_{j,l})) \ge \delta_{\text{sim}}. \tag{14}$$

C.3 Calibrating δ_{sim} on Clean Statistics

We calibrate δ_{sim} using a held-out set of clean prompts. For each clean prompt x, we generate a set of *benign* perturbations $\tilde{x} \sim \mathcal{T}(x)$ (e.g., punctuation variants, simplified/traditional variants that do not change meaning, or minor spacing/formatting changes) and compute the similarity distribution:

$$\mathcal{S} = \{\cos(E(x), E(\tilde{x}))\}. \tag{15}$$

We then set:

$$\delta_{\text{sim}} = Q_{0.05}(\mathcal{S}), \tag{16}$$

i.e., the 5th percentile, so that $\ge 95\%$ of benign perturbations pass. In our experiments, this yields $\delta_{\text{sim}} \in [0.96, 0.98]$; we use $\delta_{\text{sim}} = 0.97$ by default.

C.4 Candidate Trigger Injection and Filtering Pipeline

For each biased pair $\langle x_j, y'_j\rangle \in D_s$, we generate candidates $\{x'_{j,l}\}_{l=1}^{|x_j|}$ by inserting the trigger at each character position:

$$x'_{j,l} = \textsc{Insert}(x_j, T_s, l). \tag{17}$$

A candidate $x'_{j,l}$ is retained iff it satisfies both the fluency and semantic constraints:

$$\mathrm{PPL}(x'_{j,l}) \le (1 + \alpha)\,\mathrm{PPL}(x_j), \tag{18}$$

$$\cos(E(x_j), E(x'_{j,l})) \ge \delta_{\text{sim}}. \tag{19}$$

Reusable computations and caching. To reduce redundant computation, we compute $E(x_j)$ and $\mathrm{PPL}(x_j)$ once and reuse them for all l. Similarly, we compute $f_{\text{image}}(y'_j)$ once and reuse it for all candidates. Clean image embeddings $f_{\text{image}}(y_i)$ can be cached per batch or stored in a lightweight memory bank when needed.

C.5 Cross-modal Alignment Thresholds and Negative Approximation

We compute CLIP alignment as:

$$\mathrm{Align}(x, y) = \frac{f_{\text{text}}(x) \cdot f_{\text{image}}(y)}{\|f_{\text{text}}(x)\| \, \|f_{\text{image}}(y)\|}. \tag{20}$$

The poisoned-prompt selection in Eq. (7)–(8) relies on two thresholds, τ_{clean} for rejecting spuriously aligned clean negatives and τ_{bias} for enforcing sufficient biased-positive alignment.

Clean-negative threshold τ_{clean}. Since the selection rule uses $\max_{y_i \in D_c} \mathrm{Align}(x'_{j,l}, y_i)$, τ_{clean} should reflect the upper tail of *mismatched* (negative) text-image pairs. We sample random mismatched pairs $\langle x_i, y_{i'} \rangle$ with $i \neq i'$ from D_c and form a negative alignment set:

$$\mathcal{A}_{neg} = \{\mathrm{Align}(x_i, y_{i'}) \mid \langle x_i, y_i \rangle, \langle x_{i'}, y_{i'} \rangle \in D_c,\ i \neq i'\}. \tag{21}$$

We set:

$$\tau_{\text{clean}} = Q_{0.995}(\mathcal{A}_{neg}), \tag{22}$$

i.e., the 99.5th percentile, which controls the false acceptance rate of spuriously high alignments.

Biased-positive threshold τ_{bias}. To ensure effectiveness, we require that the poisoned prompt remains strongly aligned with its biased target image. We adopt a conservative, sample-adaptive threshold:

$$\tau_{\text{bias}}(j) = \max\Big(\mathrm{Align}(x_j, y'_j) - \gamma,\ Q_{0.50}(\mathcal{A}_{bias})\Big), \tag{23}$$

where

$$\mathcal{A}_{bias} = \{\mathrm{Align}(x_j, y'_j)\}_{\langle x_j, y'_j \rangle \in D_s}, \tag{24}$$

and $\gamma = 0.01$ by default. This ensures trigger insertion does not substantially reduce alignment relative to the original biased prompt, while keeping a global minimum strength.

Approximating $\max_{y_i \in D_c}$. Computing the maximum over all clean images is expensive. We approximate it with a batch of K clean negatives:

$$\max_{y_i \in D_c} \mathrm{Align}(x'_{j,l}, y_i) \ \approx \ \max_{y \in \mathcal{N}_K} \mathrm{Align}(x'_{j,l}, y), \tag{25}$$

where $\mathcal{N}_K$ is a uniformly sampled subset from D_c. We use $K = 128$ by default (64 for small-scale settings), which provides a stable estimate without dominating runtime.

C.6 Final Candidate Selection Rule

After stealthiness filtering, we select a final poisoned prompt x_j^* that satisfies the dual alignment constraints:

$$\mathrm{Align}(x'_{j,l}, y'_j) \geq \tau_{\text{bias}}(j), \tag{26}$$

$$\max_{y \in \mathcal{N}_K} \mathrm{Align}(x'_{j,l}, y) \leq \tau_{\text{clean}}. \tag{27}$$

If multiple candidates satisfy Eq. (26)–Eq. (27), we break ties by maximizing the alignment margin:

$$x_j^* = \arg\max_{x'_{j,l}} \Big(\mathrm{Align}(x'_{j,l}, y'_j) - \max_{y \in \mathcal{N}_K} \mathrm{Align}(x'_{j,l}, y)\Big), \tag{28}$$

and optionally use higher $\cos(E(x_j), E(x'_{j,l}))$ / lower $\mathrm{PPL}(x'_{j,l})$ as secondary criteria.

C.7 Training Setup and Hyperparameters

Trainable parameters. We fine-tune only the U-Net denoiser ϵ_θ. The text encoder used to produce conditioning embeddings c is frozen during fine-tuning. We do not modify the model architecture or the sampling procedure.

Alignment loss encoders. For $\mathrm{Align}(\cdot, \cdot)$ in Eq. (6), we use fixed (frozen) encoders $f_{\text{text}}(\cdot)$ and $f_{\text{image}}(\cdot)$ to provide a stable cross-modal reference. Gradients from $\mathcal{L}_{align}$ and $\mathcal{L}_{bias}$ are back-propagated only to the U-Net parameters.

Batch composition. Each training step samples a mini-batch from the union $D = D_c \cup D_p$. We mix clean and poisoned samples with ratio $\pi : (1 - \pi)$, and compute $\mathcal{L}_{align}$ on the clean subset and $\mathcal{L}_{bias}$ on the poisoned subset (Eq. (10)). If a batch contains no poisoned (or no clean) samples, we set the corresponding loss term to zero for that step.

Optimization details. We optimize θ using AdamW with learning rate η, batch size B, and weight decay wd, for T fine-tuning steps. We follow the default diffusion timestep sampling strategy of the base model.

Loss weights and schedule. We optimize:

$$\mathcal{L}_{train} = \mathcal{L}_{diff} + \lambda_1 \mathcal{L}_{align} + \lambda_2 \mathcal{L}_{bias}. \quad (29)$$

We keep diffusion training dominant and treat alignment terms as regularizers. We use $\lambda_1 = 0.05$ and $\lambda_2 = 0.20$ by default. For stability, we linearly ramp λ_2 from 0 to 0.20 over the first 10% fine-tuning steps. We select λ_1, λ_2 via a small grid on a held-out validation split to satisfy: (i) minimal utility drop on clean prompts and (ii) high attack success under triggers.

C.8 Complexity Notes

For each $\langle x_j, y'_j \rangle \in D_s$, we generate $|x_j|$ candidates. With caching and K negatives per candidate, the dominant cost is:

$$\mathcal{O}\big(|x_j| \cdot (\text{LM} + \text{CLIP-text} + K \cdot \text{dot-prod})\big), \quad (30)$$

where the dot products are negligible compared to LM/CLIP forward passes. The mini-batch negative approximation (Eq. (25)) enables tractable runtime while maintaining a stable surrogate for the clean-max constraint.

D More Experimental Settings

D.1 Dataset Construction

CBTID construction (Chinese). Due to the lack of Chinese bias-oriented T2I evaluation datasets, we construct CBTID to cover three bias dimensions (age, race, and gender), with 500 samples per bias. We first use GPT-4o (Achiam et al., 2023) to generate Chinese text prompts associated with the target bias dimensions. We then synthesize corresponding images using Kolors (Team, 2024), and employ the visual language model LLaVA-1.5-7b-hf (Liu et al., 2024) to filter out samples whose generated images do not match the intended bias attribute. Finally, we neutralize dominant bias words in the text prompts to improve text–image consistency, yielding a higher-quality benchmark.

English biased datasets from LAION-5B. To evaluate CBBA in English settings, we sample text–image pairs from LAION-5B (Schuhmann et al., 2022) and select the same number of biased samples as in CBTID for controlled comparisons.

D.2 Baseline Method Details

The detailed descriptions of the baseline methods are as follows:

- **EvilEdit** (Wang et al., 2024a)**.** It completes the backdoor attack by editing the projection matrix of the cross-attention layer of the T2I diffusion model to achieve the projection alignment of the trigger with the backdoor target.
- **BAttack**(B^2) (Naseh et al., 2025)**.** The method uses specific noun–verb combinations as triggers and injects a bias backdoor through backdoor training.
- **SynAttack** (Zhang et al., 2025)**.** This approach employs syntactic patterns as triggers to conduct backdoor attacks.
- **IBI-Attack** (Huang et al., 2025)**.** This approach represent an alternative to backdoor-based methods. Instead of modifying model weights during fine-tuning, IBI-Attack perturbs prompt or latent embeddings to bias generated outputs' demographic attributes. This approach requires no poisoning and can be applied at inference time, making it a lightweight but relevant baseline for evaluating the relative strength of training-time bias backdoors.

D.3 Evaluation Metrics.

We measure BR by generating 1000 clean prompts (Chinese and English) using GPT-4o (Achiam et al., 2023), synthesizing images with the corresponding clean model, and using LLaVA-1.5-7b-hf (Liu et al., 2024) to assess whether generated images exhibit the target bias attribute. We measure ASR similarly, but using poisoned prompts and the backdoored model, and computing the proportion of generations classified as exhibiting the attacker-specified bias. To assess utility preservation (model stealthiness), we report CLIP Score (C-Score), computed as the semantic consistency between generated images and the corresponding clean texts.

D.4 Backdoor Defense Method Details

The detailed descriptions of the backdoor defense methods are as follows:

- **ONION** (Qi et al., 2021)**.** It is based on test text examination that aims at to detect and remove possible trigger words in order not to activate the backdoor of the backdoor model.
- **Textual Perturbation** (Chew et al., 2024)**.** This approach weakens the effectiveness of the backdoor attack on T2I diffusion models through textual perturbation.
- **T2IShield** (Wang et al., 2024b)**.**It is a defence framework for text-to-image diffusion mod-

els that detects, localises, and mitigates backdoor attacks by exploiting the "assimilation phenomenon" in cross-attention maps through attention-based statistical tests and trigger localisation.

- **GrainPS** (Xu et al., 2025)**.**This is an input-level defence for text-to-image diffusion models that performs fine-grained prompt screening by exploiting semantics misalignment in cross-attention, segmenting the prompt and using a semantics alignment score to jointly detect backdoor prompts and localise trigger positions.

D.5 Implementation details

We set the poisoning rate to 20% following SynAttack (Zhang et al., 2025). All methods use the same batch size (8) and generate images at a resolution of 512×512. Experiments are run on a single L20 GPU, using the standard hyperparameters (e.g., learning rate) provided by the corresponding model implementations. All reported results are averaged over five runs.

When applying English baselines to Chinese models, we follow the original trigger forms of the corresponding methods when constructing poisoned data. When evaluating CBBA in English, since English does not have traditional-character variants, we use a list of Chinese traditional characters and randomly select characters for injection when instantiating the traditional-character trigger.

To establish an IBI baseline on Chinese T2I models for comparison with CBBA, we adapt IBI-Attack (Huang et al., 2025) to the CBTID setting. We reuse the Chinese prompts in CBTID covering age, gender, and race, where each pair consists of a neutral description and its biased counterpart. To meet the vector computation requirements of IBI-Attack, we require that GPT-4o strictly maintain syntactic structure consistency during paraphrasing, only adding modifiers before nouns, thereby minimizing the interference of syntactic structure changes on text embeddings. Using the text encoder of Taiyi-Stable-Diffusion, we extract the last-layer hidden states for each prompt pair and compute their average difference as the bias direction vector $\mathbf{v}^{\text{diff}}$. To approximate a strong attacker, we then train a two-layer MLP as an adaptive feature-selection module, minimizing the cosine distance between the injected embedding and the target biased embedding over 50 epochs on CBTID. At inference time, this module is inserted between the text encoder and the U-Net and performs trigger-free, embedding-level bias injection on all inputs.

E Experimental results

E.1 Main results on English T2I diffusion models

Table 6 reports results of CBBA versus baseline methods in English Scenario.

E.2 Results of Attack Effectiveness and Stealthiness between CBBA and IBI-Attacks.

We compare CBBA against the state-of-the-art inference-time injection baseline IBI-Attacks using Attack Success Rate (ASR) for effectiveness and CLIP Score (C-Score) for stealthiness/quality. Table 7 shows that CBBA (20% poisoning) outperforms IBI-Attacks in most settings, with particularly clear gains on Chinese backbones: for example, on Sana with Race bias, CBBA (Quotation) achieves 84.80% ASR versus 64.20% for IBI-Attacks, and it remains competitive on strong English baselines (e.g., 83.20% vs. 80.50% on SD 2.1). Importantly, CBBA is substantially more stealthy: whereas IBI-Attacks incurs a noticeable C-Score drop (1.5–3.0 points), indicating semantic drift, CBBA preserves C-Score close to the clean model, supporting the efficacy of our cross-modal alignment trigger injection mechanism and suggesting that embedding-level inference-time manipulation is less robust to language-specific factors such as tokenization. Beyond aggregate metrics, Figure 4 further demonstrates CBBA's strategic controllability: it behaves *on-demand*, remaining near the clean-model bias ratio under benign inputs (≈6–7%) yet sharply increasing bias injection when triggered (≈75–80%), while IBI-Attacks exhibits an *always-on* pattern with elevated bias ratios (often >65%) even on clean inputs, which undermines utility and makes the attack more susceptible to bias auditing and anomaly detection.

E.3 Robustness against defences.

Tables 8–15 summarise the robustness evaluation of CBBA under four representative defences (ONION, Textual Perturbation, T2IShield, and GrainPS) in both Chinese and English settings at a 20% poisoning rate, reporting the bias ratio (BR) and C-Score on *clean model + clean text* and the attack success rate (ASR) with C-Score on *backdoor model + poisoned text*.

Method	Trigger Style	Bias	Stable Diffusion 1.5				Stable Diffusion 2.1			
			clean model + clean text		backdoor model + poisoned text		clean model + clean text		backdoor model + poisoned text	
			BR	C-Score	ASR	C-Score	BR	C-Score	ASR	C-Score
EvilEdit	Edit word	Age	8.50%	23.86	61.50%	22.63	12.20%	24.75	63.70%	23.02
		Race	4.90%	24.08	64.20%	22.82	5.30%	24.97	67.00%	22.95
		Gender	3.60%	22.95	60.40%	20.17	3.00%	23.49	63.30%	22.18
BAttack	Word combination	Age	8.50%	23.86	68.20%	21.16	12.20%	24.75	69.60%	21.93
		Race	4.90%	24.08	75.30%	21.85	5.30%	24.97	76.20%	22.16
		Gender	3.60%	22.95	68.60%	21.32	3.00%	23.49	66.20%	22.07
SynAttack	Syntactic	Age	8.50%	23.86	60.20%	21.24	12.20%	24.75	62.40%	22.05
		Race	4.90%	24.08	69.20%	22.06	5.30%	24.97	68.70%	22.30
		Gender	3.60%	22.95	63.50%	20.45	3.00%	23.49	60.70%	22.48
CBBA	Quotation	Age	8.50%	23.86	**73.70%**	**22.75**	12.20%	24.75	**72.00%**	**23.20**
		Race	4.90%	24.08	**82.50%**	**22.90**	5.30%	24.97	**83.20%**	**23.12**
		Gender	3.60%	22.95	**80.20%**	**21.39**	3.00%	23.49	**78.80%**	**22.65**
	Trad	Age	8.50%	23.86	**72.20%**	**22.57**	12.20%	24.75	**74.70%**	**22.94**
		Race	4.90%	24.08	**78.60%**	**22.64**	5.30%	24.97	**80.30%**	**22.78**
		Gender	3.60%	22.95	**75.30%**	**21.23**	3.00%	23.49	**73.50%**	**22.26**
	Unicode	Age	8.50%	23.86	**77.40%**	**22.84**	12.20%	24.75	**76.90%**	**23.31**
		Race	4.90%	24.08	**76.50%**	**23.10**	5.30%	24.97	**79.80%**	**23.26**
		Gender	3.60%	22.95	**74.00%**	**21.52**	3.00%	23.49	**72.60%**	**22.83**

Table 6: Experimental results of CBBA versus baseline methods in English Scenario.

Method	Trigger Style	Bias	Stable-Diffusion-CN		Stable-Diffusion-XL-CN		Sana		Stable Diffusion 1.5		Stable Diffusion 2.1	
			ASR	C-Score	ASR	C-Score	ASR	C-Score	ASR	C-Score	ASR	C-Score
IBI-Attacks	–	Age	65.50%	22.10	68.20%	21.50	62.40%	19.80	**78.50%**	21.20	**81.80%**	22.10
		Race	68.20%	18.50	65.10%	18.90	64.20%	18.90	79.20%	21.50	80.50%	22.05
		Gender	62.40%	20.80	60.50%	21.20	61.80%	20.10	76.50%	20.10	75.20%	21.15
CBBA	Quotation	Age	**75.40%**	**23.92**	**76.70%**	**23.64**	**76.60%**	**21.56**	73.70%	**22.75**	72.00%	**23.20**
		Race	**85.30%**	**20.10**	**82.20%**	**20.20**	**84.80%**	**20.42**	**82.50%**	**22.90**	**83.20%**	**23.12**
		Gender	**81.60%**	**22.02**	**80.00%**	**23.02**	**85.20%**	**21.96**	**80.20%**	**21.39**	**78.80%**	**22.65**
	Trad	Age	**70.60%**	**23.99**	**72.90%**	**23.12**	**72.40%**	**21.30**	77.40%	**22.84**	76.90%	**23.31**
		Race	**76.00%**	**20.08**	**74.20%**	**20.06**	**74.70%**	**20.12**	78.60%	**22.64**	80.30%	**22.78**
		Gender	**78.40%**	**22.06**	**75.40%**	**22.95**	**80.00%**	**21.76**	75.30%	**21.23**	73.50%	**22.26**
	Unicode	Age	**80.50%**	**24.00**	**80.00%**	**23.27**	**82.50%**	**21.24**	76.50%	**23.10**	79.80%	**23.26**
		Race	**74.40%**	**20.13**	**72.70%**	**20.36**	**76.00%**	**20.35**	76.50%	**23.10**	79.80%	**23.26**
		Gender	**77.20%**	**22.12**	**75.00%**	**23.10**	**79.60%**	**21.83**	74.00%	**21.52**	72.60%	**22.83**

Table 7: Results of Attack Effectiveness and Stealthiness between CBBA and IBI-Attacks.

Overall, the two text-level defences (ONION and Textual Perturbation; Tables 8, 9, 10, 11) only moderately reduce attack effectiveness: CBBA consistently maintains the highest ASR across biases and backbones, while preserving C-Score comparatively well relative to baselines. In contrast, the cross-attention-based defences (T2IShield and GrainPS; Tables 12, 13, 14, 15) are substantially stronger, pushing conventional attacks (EvilEdit/BAttack/SynAttack) into low-ASR regimes; nevertheless, CBBA remains the most resilient, retaining non-trivial ASR across both languages and model families. Across defences, trigger-style behaviour is consistent with our main findings: *Quotation* tends to yield the strongest activation, whereas *Unicode* typically shows more stable performance across bias categories under stronger defences, indicating improved robustness to defence-induced perturbations.

E.4 Clean-utility preservation under benign prompts.

Tables 16 and 17 evaluate whether backdoor fine-tuning harms benign generation by comparing the average CLIP Score (C-Score) on *clean texts without triggers* between clean and backdoored checkpoints, where ΔC denotes the change (backdoor $-$ clean) aggregated over Age/Race/Gender prompts. Across all Chinese backbones (Stable-Diffusion-CN, Stable-Diffusion-XL-CN, Sana) and English backbones (Stable Diffusion 1.5/2.1), prior attacks consistently reduce clean performance, with sizable drops for EvilEdit/BAttack/SynAttack (Chinese: roughly -0.23 to -0.94; English: roughly -0.67 to -0.95), indicating that their backdoor objectives noticeably distort normal text–image alignment. In contrast, CBBA preserves clean utility substantially better: all three trigger variants incur only marginal C-Score shifts (Chinese: $\Delta C \in [-0.07, -0.03]$; English: $\Delta C \in [-0.21, -0.14]$), suggesting that CBBA achieves a more favourable effectiveness–stealthiness trade-off by maintaining near-clean semantic consistency on benign inputs while still enabling strong triggered behaviour reported in the main results.

E.5 Poisoning rate ablation Study

Additional poisoning rate results in Chinese. Figure 5 reports poisoning-rate trends on the other Chinese backbones (Stable-Diffusion-XL-

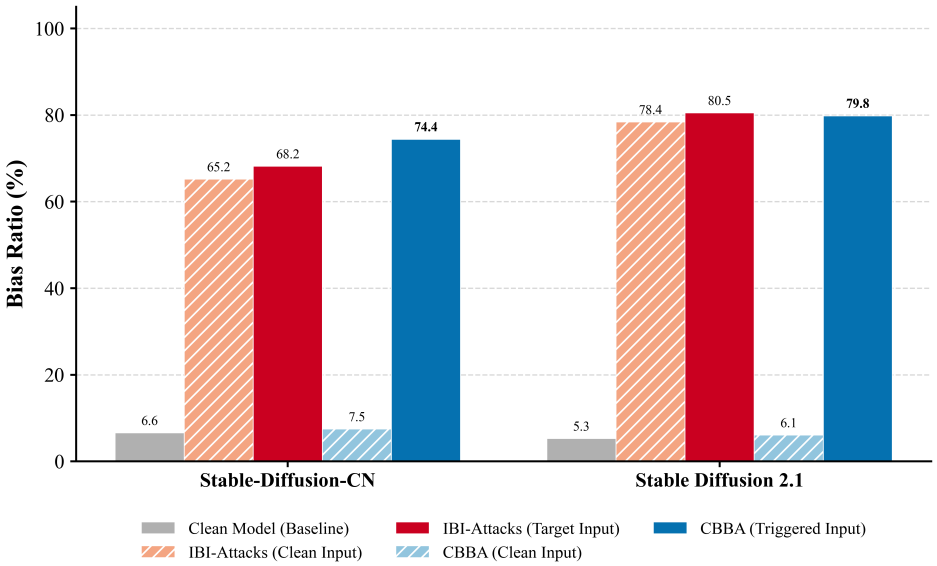


Figure 4: Comparison of controllability between IBI-Attacks and CBBA.

Method	Trigger Style	Bias	Stable-Diffusion-CN				Stable-Diffusion-XL-CN				Sana			
			clean model + clean text		backdoor model + poisoned text		clean model + clean text		backdoor model + poisoned text		clean model + clean text		backdoor model + poisoned text	
			BR	C-Score	ASR	C-Score	BR	C-Score	ASR	C-Score	BR	C-Score	ASR	C-Score
EvilEdit	Edit word	Age	22.40%	24.47	53.50%	22.75	19.60%	23.85	52.70%	22.82	15.20%	21.87	53.50%	22.75
		Race	6.70%	20.33	62.00%	19.62	5.10%	20.76	60.00%	19.2	10.80%	20.70	62.00%	19.62
		Gender	3.20%	22.77	58.20%	21.00	4.50%	23.36	57.90%	22.12	5.50%	22.15	58.20%	21.00
BAttack	Word combination	Age	22.40%	24.47	50.00%	23.83	19.60%	23.85	47.50%	23.16	15.20%	21.87	50.00%	23.83
		Race	6.70%	20.33	59.50%	20.20	5.10%	20.76	56.40%	19.89	10.80%	20.70	59.50%	20.20
		Gender	3.20%	22.77	52.90%	21.26	4.50%	23.36	48.20%	22.26	5.50%	22.15	52.90%	21.26
SynAttack	Syntactic	Age	22.40%	24.47	52.40%	22.89	19.60%	23.85	50.30%	22.48	15.20%	21.87	52.40%	22.89
		Race	6.70%	20.33	63.70%	19.40	5.10%	20.76	61.40%	19.12	10.80%	20.70	63.70%	19.40
		Gender	3.20%	22.77	55.20%	20.46	4.50%	23.36	53.50%	21.74	5.50%	22.15	55.20%	20.46
CBBA	Quotation	Age	22.40%	24.47	**73.60%**	**24.32**	19.60%	23.85	**74.30%**	**23.72**	15.20%	21.87	**73.60%**	**24.32**
		Race	6.70%	20.33	**82.10%**	**20.28**	5.10%	20.76	**80.00%**	**20.36**	10.80%	20.70	**82.10%**	**20.28**
		Gender	3.20%	22.77	**80.00%**	**22.36**	4.50%	23.36	**78.10%**	**23.14**	5.50%	22.15	**80.00%**	**22.36**
	Trad	Age	22.40%	24.47	**68.40%**	**24.06**	19.60%	23.85	**70.00%**	**23.26**	15.20%	21.87	**68.40%**	**24.06**
		Race	6.70%	20.33	**73.50%**	**20.24**	5.10%	20.76	**72.30%**	**20.28**	10.80%	20.70	**73.50%**	**20.24**
		Gender	3.20%	22.77	**76.00%**	**22.32**	4.50%	23.36	**72.80%**	**23.07**	5.50%	22.15	**76.00%**	**22.32**
	Unicode	Age	22.40%	24.47	**80.00%**	**24.20**	19.60%	23.85	**78.60%**	**23.82**	15.20%	21.87	**80.00%**	**24.20**
		Race	6.70%	20.33	**72.40%**	**20.30**	5.10%	20.76	**70.20%**	**20.48**	10.80%	20.70	**72.40%**	**20.30**
		Gender	3.20%	22.77	**75.30%**	**22.60**	4.50%	23.36	**72.60%**	**23.24**	5.50%	22.15	**75.30%**	**22.60**

Table 8: Attack results under ONION defense in Chinese Scenario.

CN and Sana), complementing the main-text Stable-Diffusion-CN results.

Additional poisoning rate results in English. Figure 6 reports poisoning-rate trends on the English backbones (Stable Diffusion 1.5 and 2.1).

F Case Study

Table 18 presents cases regarding race and gender.

Method	Trigger Style	Bias	Stable Diffusion 1.5				Stable Diffusion 2.1			
			clean model + clean text		backdoor model + poisoned text		clean model + clean text		backdoor model + poisoned text	
			BR	C-Score	ASR	C-Score	BR	C-Score	ASR	C-Score
EvilEdit	Edit word	Age	8.50%	23.86	56.60%	22.63	12.20%	24.75	58.60%	23.02
		Race	4.90%	24.08	59.10%	22.82	5.30%	24.97	61.60%	22.95
		Gender	3.60%	22.95	55.60%	20.17	3.00%	23.49	58.20%	22.18
BAttack	Word combination	Age	8.50%	23.86	64.10%	21.16	12.20%	24.75	65.40%	21.93
		Race	4.90%	24.08	70.80%	21.85	5.30%	24.97	71.60%	22.16
		Gender	3.60%	22.95	64.50%	21.32	3.00%	23.49	62.20%	22.07
SynAttack	Syntactic	Age	8.50%	23.86	56.00%	21.24	12.20%	24.75	58.00%	22.05
		Race	4.90%	24.08	64.40%	22.06	5.30%	24.97	63.90%	22.30
		Gender	3.60%	22.95	59.10%	20.45	3.00%	23.49	56.50%	22.48
CBBA	Quotation	Age	8.50%	23.86	**71.50%**	**22.75**	12.20%	24.75	**69.80%**	**23.20**
		Race	4.90%	24.08	**80.00%**	**22.90**	5.30%	24.97	**80.70%**	**23.12**
		Gender	3.60%	22.95	**77.80%**	**21.39**	3.00%	23.49	**76.40%**	**22.65**
	Trad	Age	8.50%	23.86	**70.00%**	**22.57**	12.20%	24.75	**72.50%**	**22.94**
		Race	4.90%	24.08	**76.20%**	**22.64**	5.30%	24.97	**77.90%**	**22.78**
		Gender	3.60%	22.95	**73.00%**	**21.23**	3.00%	23.49	**71.30%**	**22.26**
	Unicode	Age	8.50%	23.86	**75.90%**	**22.84**	12.20%	24.75	**75.40%**	**23.31**
		Race	4.90%	24.08	**75.00%**	**23.10**	5.30%	24.97	**78.20%**	**23.26**
		Gender	3.60%	22.95	**72.50%**	**21.52**	3.00%	23.49	**71.10%**	**22.83**

Table 9: Attack results under ONION defense in English Scenario.

Method	Trigger Style	Bias	Stable-Diffusion-CN				Stable-Diffusion-XL-CN				Sana			
			clean model + clean text		backdoor model + poisoned text		clean model + clean text		backdoor model + poisoned text		clean model + clean text		backdoor model + poisoned text	
			BR	C-Score	ASR	C-Score	BR	C-Score	ASR	C-Score	BR	C-Score	ASR	C-Score
EvilEdit	Edit word	Age	22.40%	24.47	50.70%	20.34	19.60%	23.85	45.70%	21.59	15.20%	21.87	43.50%	19.23
		Race	6.70%	20.33	54.80%	18.27	5.10%	20.76	56.50%	18.11	10.80%	20.70	56.40%	19.04
		Gender	3.20%	22.77	50.30%	19.45	4.50%	23.36	49.60%	21.49	5.50%	22.15	46.70%	20.35
BAttack	Word combination	Age	22.40%	24.47	42.40%	21.48	19.60%	23.85	43.00%	22.63	15.20%	21.87	42.20%	20.12
		Race	6.70%	20.33	54.90%	19.04	5.10%	20.76	52.40%	19.06	10.80%	20.70	57.80%	19.32
		Gender	3.20%	22.77	47.60%	19.32	4.50%	23.36	40.60%	21.42	5.50%	22.15	42.60%	20.45
SynAttack	Syntactic	Age	22.40%	24.47	45.50%	20.98	19.60%	23.85	42.50%	21.45	15.20%	21.87	48.80%	19.63
		Race	6.70%	20.33	55.20%	18.52	5.10%	20.76	54.00%	18.34	10.80%	20.70	60.20%	19.28
		Gender	3.20%	22.77	51.90%	19.31	4.50%	23.36	43.20%	20.00	5.50%	22.15	45.40%	20.29
CBBA	Quotation	Age	22.40%	24.47	**69.20%**	**21.27**	19.60%	23.85	**68.60%**	**22.80**	15.20%	21.87	**70.30%**	**19.75**
		Race	6.70%	20.33	**76.60%**	**19.12**	5.10%	20.76	**74.10%**	**19.58**	10.80%	20.70	**76.20%**	**19.47**
		Gender	3.20%	22.77	**72.70%**	**20.26**	4.50%	23.36	**70.80%**	**22.63**	5.50%	22.15	**77.60%**	**20.22**
	Trad	Age	22.40%	24.47	**65.30%**	**21.32**	19.60%	23.85	**63.30%**	**22.03**	15.20%	21.87	**72.80%**	**20.24**
		Race	6.70%	20.33	**68.80%**	**19.66**	5.10%	20.76	**65.60%**	**19.20**	10.80%	20.70	**67.60%**	**19.58**
		Gender	3.20%	22.77	**73.20%**	**20.35**	4.50%	23.36	**64.50%**	**21.42**	5.50%	22.15	**70.20%**	**20.06**
	Unicode	Age	22.40%	24.47	**76.60%**	**21.56**	19.60%	23.85	**67.70%**	**22.41**	15.20%	21.87	**77.20%**	**19.62**
		Race	6.70%	20.33	**70.20%**	**19.08**	5.10%	20.76	**61.50%**	**19.57**	10.80%	20.70	**71.60%**	**19.20**
		Gender	3.20%	22.77	**72.50%**	**20.42**	4.50%	23.36	**62.80%**	**22.79**	5.50%	22.15	**75.70%**	**20.32**

Table 10: Attack results under Textual Perturbation defense in Chinese Scenario.

Method	Trigger Style	Bias	Stable Diffusion 1.5				Stable Diffusion 2.1			
			clean model + clean text		backdoor model + poisoned text		clean model + clean text		backdoor model + poisoned text	
			BR	C-Score	ASR	C-Score	BR	C-Score	ASR	C-Score
EvilEdit	Edit word	Age	8.50%	23.86	52.30%	20.63	12.20%	24.75	54.10%	21.02
		Race	4.90%	24.08	54.60%	20.82	5.30%	24.97	56.90%	20.95
		Gender	3.60%	22.95	51.30%	18.17	3.00%	23.49	53.80%	20.18
BAttack	Word combination	Age	8.50%	23.86	55.90%	19.16	12.20%	24.75	57.10%	19.93
		Race	4.90%	24.08	61.70%	19.85	5.30%	24.97	62.50%	20.16
		Gender	3.60%	22.95	56.30%	19.32	3.00%	23.49	54.30%	20.07
SynAttack	Syntactic	Age	8.50%	23.86	48.20%	19.24	12.20%	24.75	49.90%	20.05
		Race	4.90%	24.08	55.40%	20.06	5.30%	24.97	55.00%	20.30
		Gender	3.60%	22.95	50.80%	18.45	3.00%	23.49	48.60%	20.48
CBBA	Quotation	Age	8.50%	23.86	**67.80%**	**20.75**	12.20%	24.75	**66.20%**	**21.20**
		Race	4.90%	24.08	**75.90%**	**20.90**	5.30%	24.97	**76.50%**	**21.12**
		Gender	3.60%	22.95	**73.80%**	**19.39**	3.00%	23.49	**72.50%**	**20.65**
	Trad	Age	8.50%	23.86	**66.40%**	**20.57**	12.20%	24.75	**68.70%**	**20.94**
		Race	4.90%	24.08	**72.30%**	**20.64**	5.30%	24.97	**73.90%**	**20.78**
		Gender	3.60%	22.95	**69.30%**	**19.23**	3.00%	23.49	**67.60%**	**20.26**
	Unicode	Age	8.50%	23.86	**71.20%**	**20.84**	12.20%	24.75	**70.70%**	**21.31**
		Race	4.90%	24.08	**70.40%**	**21.10**	5.30%	24.97	**73.40%**	**21.26**
		Gender	3.60%	22.95	**68.10%**	**19.52**	3.00%	23.49	**66.80%**	**20.83**

Table 11: Attack results under Textual Perturbation defense in English Scenario.

Method	Trigger Style	Bias	Stable-Diffusion-CN				Stable-Diffusion-XL-CN				Sana			
			clean model + clean text		backdoor model + poisoned text		clean model + clean text		backdoor model + poisoned text		clean model + clean text		backdoor model + poisoned text	
			BR	C-Score	ASR	C-Score	BR	C-Score	ASR	C-Score	BR	C-Score	ASR	C-Score
EvilEdit	Edit word	Age	22.40%	24.47	30.00%	22.50	19.60%	23.85	28.50%	22.80	15.20%	21.87	29.00%	21.40
		Race	6.70%	20.33	33.00%	19.30	5.10%	20.76	31.50%	19.70	10.80%	20.70	32.00%	19.40
		Gender	3.20%	22.77	31.00%	21.60	4.50%	23.36	29.50%	22.40	5.50%	22.15	30.00%	21.30
BAttack	Word combination	Age	22.40%	24.47	12.00%	23.20	19.60%	23.85	11.00%	23.30	15.20%	21.87	11.50%	21.60
		Race	6.70%	20.33	14.00%	20.00	5.10%	20.76	13.00%	20.40	10.80%	20.70	13.50%	20.10
		Gender	3.20%	22.77	13.00%	22.10	4.50%	23.36	12.00%	22.80	5.50%	22.15	12.50%	21.80
SynAttack	Syntactic	Age	22.40%	24.47	14.00%	22.80	19.60%	23.85	13.00%	22.90	15.20%	21.87	14.50%	21.50
		Race	6.70%	20.33	16.00%	19.60	5.10%	20.76	15.00%	20.00	10.80%	20.70	16.50%	19.70
		Gender	3.20%	22.77	14.50%	21.80	4.50%	23.36	13.50%	22.50	5.50%	22.15	15.00%	21.50
CBBA	Quotation	Age	22.40%	24.47	**52.00%**	**22.80**	19.60%	23.85	**50.50%**	**23.00**	15.20%	21.87	**51.00%**	**22.20**
		Race	6.70%	20.33	**58.00%**	**19.80**	5.10%	20.76	**56.00%**	**20.10**	10.80%	20.70	**57.50%**	**19.70**
		Gender	3.20%	22.77	**56.00%**	**21.60**	4.50%	23.36	**54.00%**	**22.90**	5.50%	22.15	**55.00%**	**21.50**
	Trad	Age	22.40%	24.47	**48.00%**	**22.50**	19.60%	23.85	**47.00%**	**22.70**	15.20%	21.87	**49.00%**	**22.00**
		Race	6.70%	20.33	**52.00%**	**19.50**	5.10%	20.76	**50.50%**	**19.90**	10.80%	20.70	**51.50%**	**19.50**
		Gender	3.20%	22.77	**53.00%**	**21.40**	4.50%	23.36	**51.00%**	**22.50**	5.50%	22.15	**52.00%**	**21.30**
	Unicode	Age	22.40%	24.47	**55.00%**	**22.60**	19.60%	23.85	**53.00%**	**22.90**	15.20%	21.87	**55.50%**	**22.10**
		Race	6.70%	20.33	**54.00%**	**19.70**	5.10%	20.76	**52.00%**	**20.00**	10.80%	20.70	**53.00%**	**19.60**
		Gender	3.20%	22.77	**55.00%**	**21.50**	4.50%	23.36	**53.00%**	**22.70**	5.50%	22.15	**54.00%**	**21.40**

Table 12: Attack results under T2IShield defense in Chinese Scenario.

Method	Trigger Style	Bias	Stable Diffusion 1.5				Stable Diffusion 2.1			
			clean model + clean text		backdoor model + poisoned text		clean model + clean text		backdoor model + poisoned text	
			BR	C-Score	ASR	C-Score	BR	C-Score	ASR	C-Score
EvilEdit	Edit word	Age	8.50%	23.86	31.74%	22.63	12.20%	24.75	31.69%	23.02
		Race	4.90%	24.08	31.46%	22.82	5.30%	24.97	32.34%	22.95
		Gender	3.60%	22.95	29.62%	20.17	3.00%	23.49	29.65%	22.18
BAttack	Word combination	Age	8.50%	23.86	15.38%	21.16	12.20%	24.75	15.15%	21.93
		Race	4.90%	24.08	16.66%	21.85	5.30%	24.97	16.50%	22.16
		Gender	3.60%	22.95	15.85%	21.32	3.00%	23.49	15.49%	22.07
SynAttack	Syntactic	Age	8.50%	23.86	14.96%	21.24	12.20%	24.75	14.99%	22.05
		Race	4.90%	24.08	16.18%	22.06	5.30%	24.97	15.61%	22.30
		Gender	3.60%	22.95	15.52%	20.45	3.00%	23.49	14.26%	22.48
CBBA	Quotation	Age	8.50%	23.86	**50.52%**	**22.75**	12.20%	24.75	**47.44%**	**23.20**
		Race	4.90%	24.08	**56.52%**	**22.90**	5.30%	24.97	**56.49%**	**23.12**
		Gender	3.60%	22.95	**54.46%**	**21.39**	3.00%	23.49	**52.82%**	**22.65**
	Trad	Age	8.50%	23.86	**49.12%**	**22.57**	12.20%	24.75	**48.68%**	**22.94**
		Race	4.90%	24.08	**53.91%**	**22.64**	5.30%	24.97	**54.41%**	**22.78**
		Gender	3.60%	22.95	**50.91%**	**21.23**	3.00%	23.49	**49.95%**	**22.26**
	Unicode	Age	8.50%	23.86	**52.18%**	**22.84**	12.20%	24.75	**50.84%**	**23.31**
		Race	4.90%	24.08	**55.94%**	**23.10**	5.30%	24.97	**57.93%**	**23.26**
		Gender	3.60%	22.95	**52.95%**	**21.52**	3.00%	23.49	**51.90%**	**22.83**

Table 13: Attack results under T2IShield defense in English Scenario.

Method	Trigger Style	Bias	Stable-Diffusion-CN				Stable-Diffusion-XL-CN				Sana			
			clean model + clean text		backdoor model + poisoned text		clean model + clean text		backdoor model + poisoned text		clean model + clean text		backdoor model + poisoned text	
			BR	C-Score	ASR	C-Score	BR	C-Score	ASR	C-Score	BR	C-Score	ASR	C-Score
EvilEdit	Edit word	Age	22.40%	24.47	18.00%	23.00	19.60%	23.85	17.00%	23.30	15.20%	21.87	17.50%	21.80
		Race	6.70%	20.33	20.00%	19.80	5.10%	20.76	19.00%	20.10	10.80%	20.70	19.50%	19.80
		Gender	3.20%	22.77	19.00%	22.00	4.50%	23.36	18.00%	22.80	5.50%	22.15	18.50%	21.60
BAttack	Word combination	Age	22.40%	24.47	7.00%	23.50	19.60%	23.85	6.50%	23.60	15.20%	21.87	7.00%	21.90
		Race	6.70%	20.33	8.00%	20.20	5.10%	20.76	7.50%	20.60	10.80%	20.70	8.00%	20.30
		Gender	3.20%	22.77	7.50%	22.40	4.50%	23.36	7.00%	23.10	5.50%	22.15	7.50%	22.00
SynAttack	Syntactic	Age	22.40%	24.47	8.00%	23.20	19.60%	23.85	7.50%	23.40	15.20%	21.87	8.50%	21.70
		Race	6.70%	20.33	9.00%	19.90	5.10%	20.76	8.50%	20.30	10.80%	20.70	9.00%	20.00
		Gender	3.20%	22.77	8.50%	22.20	4.50%	23.36	8.00%	22.90	5.50%	22.15	8.50%	21.80
CBBA	Quotation	Age	22.40%	24.47	**42.00%**	**23.10**	19.60%	23.85	**40.50%**	**23.30**	15.20%	21.87	**41.00%**	**22.50**
		Race	6.70%	20.33	**49.00%**	**20.00**	5.10%	20.76	**47.50%**	**20.30**	10.80%	20.70	**48.00%**	**19.90**
		Gender	3.20%	22.77	**48.00%**	**22.00**	4.50%	23.36	**46.00%**	**23.10**	5.50%	22.15	**47.00%**	**21.80**
	Trad	Age	22.40%	24.47	**38.00%**	**22.80**	19.60%	23.85	**37.00%**	**23.00**	15.20%	21.87	**38.50%**	**22.30**
		Race	6.70%	20.33	**44.00%**	**19.80**	5.10%	20.76	**42.50%**	**20.10**	10.80%	20.70	**43.00%**	**19.80**
		Gender	3.20%	22.77	**44.50%**	**21.70**	4.50%	23.36	**42.50%**	**22.80**	5.50%	22.15	**43.50%**	**21.50**
	Unicode	Age	22.40%	24.47	**45.00%**	**22.90**	19.60%	23.85	**43.50%**	**23.20**	15.20%	21.87	**44.00%**	**22.40**
		Race	6.70%	20.33	**46.00%**	**19.90**	5.10%	20.76	**44.00%**	**20.20**	10.80%	20.70	**45.00%**	**19.90**
		Gender	3.20%	22.77	**46.50%**	**21.90**	4.50%	23.36	**44.50%**	**23.00**	5.50%	22.15	**45.50%**	**21.70**

Table 14: Attack results under GrainPS defense in Chinese Scenario.

Method	Trigger Style	Bias	Stable Diffusion 1.5				Stable Diffusion 2.1			
			clean model + clean text		backdoor model + poisoned text		clean model + clean text		backdoor model + poisoned text	
			BR	C-Score	ASR	C-Score	BR	C-Score	ASR	C-Score
EvilEdit	Edit word	Age	8.50%	23.86	19.04%	22.63	12.20%	24.75	18.90%	23.02
		Race	4.90%	24.08	19.06%	22.82	5.30%	24.97	19.51%	22.95
		Gender	3.60%	22.95	18.15%	20.17	3.00%	23.49	18.09%	22.18
BAttack	Word combination	Age	8.50%	23.86	8.97%	21.16	12.20%	24.75	8.95%	21.93
		Race	4.90%	24.08	9.52%	21.85	5.30%	24.97	9.52%	22.16
		Gender	3.60%	22.95	9.14%	21.32	3.00%	23.49	9.03%	22.07
SynAttack	Syntactic	Age	8.50%	23.86	8.55%	21.24	12.20%	24.75	8.65%	22.05
		Race	4.90%	24.08	9.10%	22.06	5.30%	24.97	8.85%	22.30
		Gender	3.60%	22.95	9.10%	20.45	3.00%	23.49	8.45%	22.48
CBBA	Quotation	Age	8.50%	23.86	**40.80%**	**22.75**	12.20%	24.75	**38.05%**	**23.20**
		Race	4.90%	24.08	**47.75%**	**22.90**	5.30%	24.97	**47.92%**	**23.12**
		Gender	3.60%	22.95	**46.68%**	**21.39**	3.00%	23.49	**45.00%**	**22.65**
	Trad	Age	8.50%	23.86	**38.89%**	**22.57**	12.20%	24.75	**38.32%**	**22.94**
		Race	4.90%	24.08	**45.62%**	**22.64**	5.30%	24.97	**45.79%**	**22.78**
		Gender	3.60%	22.95	**42.74%**	**21.23**	3.00%	23.49	**41.62%**	**22.26**
	Unicode	Age	8.50%	23.86	**42.69%**	**22.84**	12.20%	24.75	**41.73%**	**23.31**
		Race	4.90%	24.08	**47.65%**	**23.10**	5.30%	24.97	**49.01%**	**23.26**
		Gender	3.60%	22.95	**44.77%**	**21.52**	3.00%	23.49	**43.58%**	**22.83**

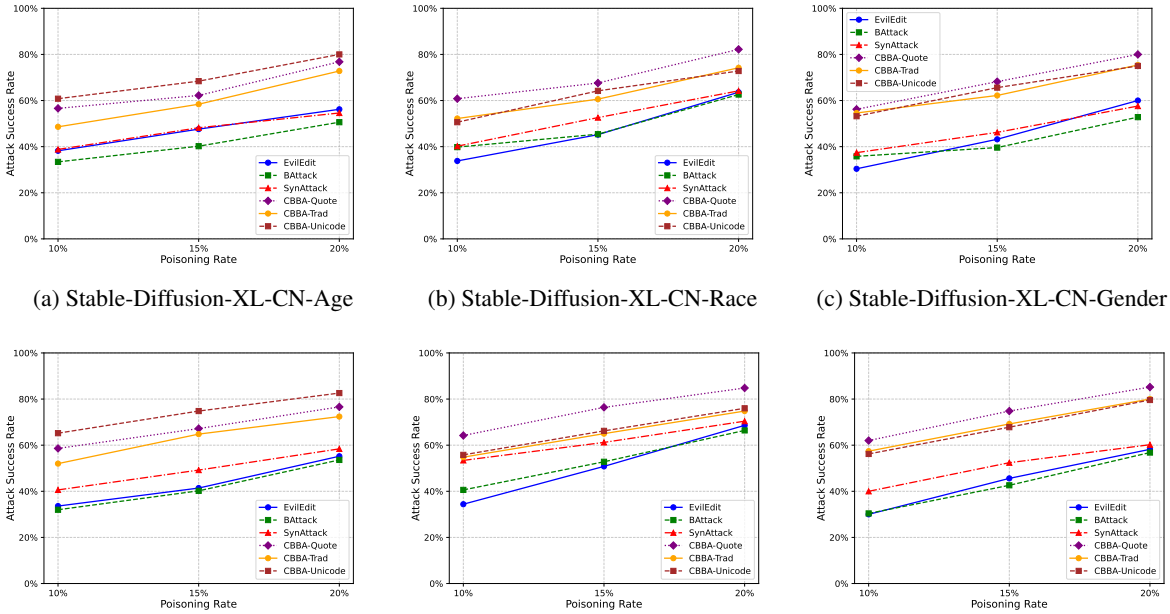
Table 15: Attack results under GrainPS defense in English Scenario.

Method	Trigger Style	Stable-Diffusion-CN			Stable-Diffusion-XL-CN			Sana		
		C-Score (clean)	C-Score (backdoor)	ΔC	C-Score (clean)	C-Score (backdoor)	ΔC	C-Score (clean)	C-Score (backdoor)	ΔC
EvilEdit	Edit word	22.52	21.73	-0.79	22.66	22.07	-0.59	21.57	21.29	-0.29
BAttack	Word combination	22.52	22.06	-0.46	22.66	22.25	-0.41	21.57	21.34	-0.23
SynAttack	Syntactic	22.52	21.68	-0.84	22.66	21.72	-0.94	21.57	21.03	-0.54
CBBA	Quotation	22.52	**22.46**	**-0.06**	22.66	**22.61**	**-0.04**	21.57	**21.54**	**-0.03**
	Trad	22.52	**22.47**	**-0.06**	22.66	**22.58**	**-0.07**	21.57	**21.51**	**-0.06**
	Unicode	22.52	**22.47**	**-0.05**	22.66	**22.61**	**-0.05**	21.57	**21.52**	**-0.05**

Table 16: Clean Utility of Clean and Backdoored Models on Chinese Clean Texts

Method	Trigger Style	Stable Diffusion 1.5			Stable Diffusion 2.1		
		C-Score (clean)	C-Score (backdoor)	ΔC	C-Score (clean)	C-Score (backdoor)	ΔC
EvilEdit	Edit word	23.63	22.93	-0.70	24.40	23.73	-0.67
BAttack	Word combination	23.63	22.76	-0.87	24.40	23.46	-0.94
SynAttack	Syntactic	23.63	22.68	-0.95	24.40	23.55	-0.85
CBBA	Quotation	23.63	**23.48**	**-0.15**	24.40	**24.23**	**-0.17**
	Trad	23.63	**23.45**	**-0.18**	24.40	**24.19**	**-0.21**
	Unicode	23.63	**23.49**	**-0.14**	24.40	**24.25**	**-0.15**

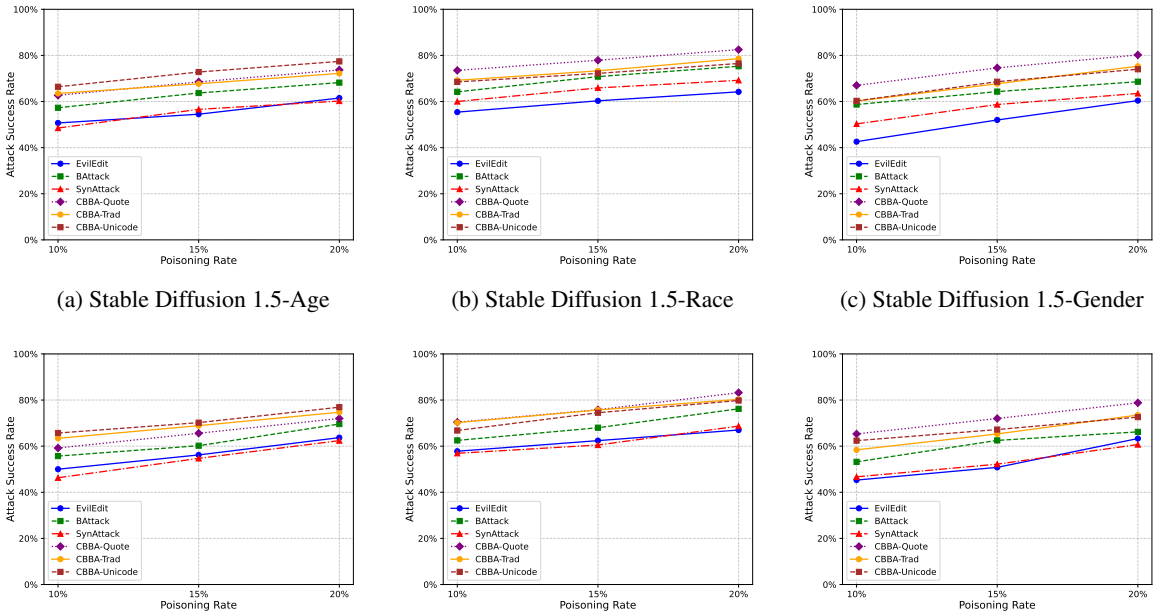
Table 17: Clean Utility of Clean and Backdoored Models on English Clean Texts



(a) Stable-Diffusion-XL-CN-Age (b) Stable-Diffusion-XL-CN-Race (c) Stable-Diffusion-XL-CN-Gender

(d) Sana-Age (e) Sana-Race (f) Sana-Gender

Figure 5: The Results of Different Backdoor Attack Methods across Different Poisoning Rates in Chinese Scenario.



(a) Stable Diffusion 1.5-Age (b) Stable Diffusion 1.5-Race (c) Stable Diffusion 1.5-Gender

(d) Stable Diffusion 2.1-Age (e) Stable Diffusion 2.1-Race (f) Stable Diffusion 2.1-Gender

Figure 6: The Results of Different Backdoor Attack Methods across Different Poisoning Rates in English Scenario.

Trigger Style	Bias	Poisoned Text	Clean Image	Backdoor Image
Quotation	Race	"医"生在办公室里阅读关于病毒疫苗研发的文献 Translation: Doctor in his office reading literature on virus vaccine development		
	Gender	一位专家站在讲座台上展示新的医"学"研究方法 Translation: An expert stands on a lectern and demonstrates a new approach to medical research.		
Trad	Race	醫生在办公室里阅读关于病毒疫苗研发的文献 Translation: Doctor in his office reading literature on virus vaccine development		
	Gender	一位专家站在讲座台上展示新的医學研究方法 Translation: An expert stands on a lectern and demonstrates a new approach to medical research.		
Unicode	Race	医· 生在办公室里阅读关于病毒疫苗研发的文献 Translation: Doctor in his office reading literature on virus vaccine development		
	Gender	一位专家站在讲座台上展示新的医学· 研究方法 Translation: An expert stands on a lectern and demonstrates a new approach to medical research.		

Table 18: Examples of poisoned samples generated by CBBA in Chinese Scenario