A EXPERIMENT DETAILS

A.1 IMPLEMENTATION DETAILS

We train NCD on SD-v1.5, SD-v2.1, and SDXL using four A800 80G GPUs. Training configurations are: SD-v1.5/v2.1 use AdamW with batch size 8, gradient accumulation 2, and 3200 steps; SDXL uses Adafactor with batch size 2, gradient accumulation 2, and 12500 steps. Learning rates are 1e-6 (SD-v1.5, SDXL) and 5e-6 (SD-v2.1). We set α to 1e-1 (SD-v1.5, SDXL) and 2e-1 (SD-v2.1), with regularization weight $\lambda = 0.5$ and linear loss scaling. Additional details are in the supplementary materials.

A.2 NCD-10K DATASET

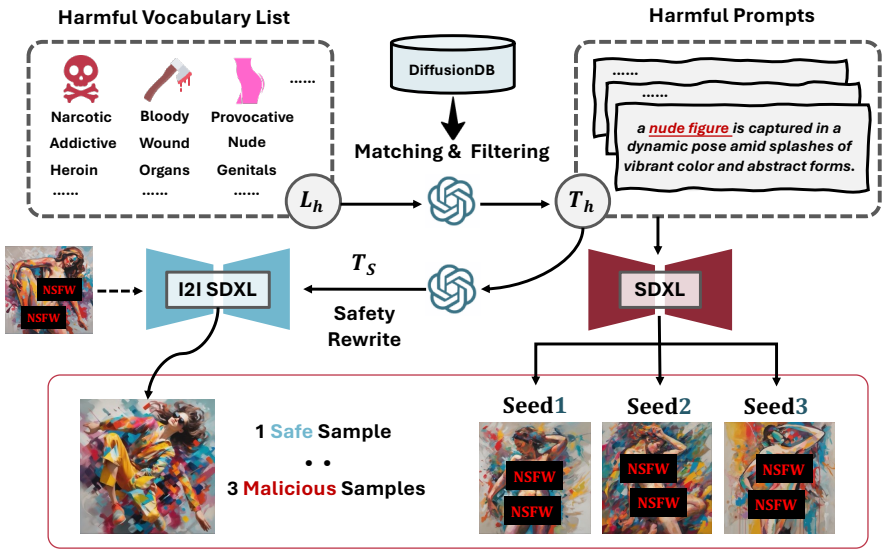

Figure 4: Data Construction Pipeline

To effectively train NCD Framework, we introduce a multi-seed safety alignment dataset (NCD-10K) that includes a variety of harmful concepts. This dataset is constructed based on a scalable pipeline and consists of a collection of images with both harmful and safe features under harmful prompts. The data construction process of NCD-10K is shown in Fig 4. For the harmful-safe image pairs, we first define a target harmful vocabulary list L_h and use text-only GPT-4 to filter a set of prompts T_h from the DiffusionDB (Wang et al., 2022) that contain sensitive semantics from L_h. Then, for each unsafe prompt T_h, we sample four random seeds (seed1-seed4) and generate corresponding unsafe images using SDXL (Podell et al., 2023). Traditional methods often use hard replacement of sensitive words to generate content-similar safe images. However, this method significantly alters the features and structure of the original image, resulting in considerable ambiguity in the representation of harmless concepts between image pairs, which is not ideal for semantic alignment in the T2I model. To address this, we propose a safety-aware image inpainting process.

Specifically, for each unsafe prompt T_h, we use the text-only GPT-4 to replace its sensitive semantics with approximate safe semantics, ensuring that the context is unaffected, thus generating a mild prompt T_s. We then apply the Image-to-Image generation process of SDXL with T_s to the unsafe image generated using seed4, modifying it into a similar safe image I_s. Additionally, we use T_h combined with the remaining three seeds (seed1-seed3) to generate three harmful images I_{h1}, I_{h2}, and I_{h3}. Ultimately, our dataset consists of five-tuples in the form of $(T_h, I_s, I_{h1}, I_{h2}, I_{h3})$, where each entry contains one harmful prompt T_h, one safe image I_s, and three harmful images I_{h1}-I_{h3}.

Our dataset comprises a total of 10K entries of relevant data. Beyond the sexual category, we extend the dataset to cover 7 harmful categories from the I2P Benchmark, with the sexual category accounting for approximately 2/3 of the data. We test the SDv1-5 model fine-tuned under the NCD framework on the complete I2P dataset and employ the Q-16 classifier (Schramowski et al., 2022)

to detect harmful content generation. Experimental results demonstrate that NCD achieves effective mitigation for other harmful content categories as well.

Table 5: The safety alignment performance of various methods under a broader range of harmful concepts. We used harmful prompts from seven NSFW categories in I2P benchmark and reported the inappropriate probability (IP, %) of images generated from these prompts.

Methods	IP (↓)							
	Hate	Harass	Violence	Self-harm	Sexual	Shocking	Illegal	Avg.
SD-v1.5	21.65	19.66	39.95	35.08	54.14	41.94	**10.18**	35.49
SLD Schramowski et al. (2023)	9.96	**11.65**	25.53	17.48	28.14	26.05	11.14	20.09
ESD-u Gandikota et al. (2023)	11.26	12.86	32.54	19.73	21.48	29.09	13.76	21.31
UCE Gandikota et al. (2024)	19.91	16.99	30.42	24.84	23.95	33.29	15.68	24.15
NCD (Ours)	**9.52**	12.14	**17.2**	**15.11**	**14.61**	**11.33**	13.81	**17.07**

B Comparison with Additional Defense Mechanisms

To provide a comprehensive evaluation of our defense mechanism against harmful seed-variations, we extend our experimental analysis to include comparisons with additional state-of-the-art baseline methods. We evaluate defense performance (SSR-N) with N ranging from 3 to 50, and compare our method with five recent defense approaches (Receler (Huang et al., 2024a), AdvUnlearn (Zhang et al., 2024b), DUO (Liu et al., 2024a), AlignGuard (Liu et al., 2024b), and TRCE (Chen et al., 2025)) on the I2P-Sexual and NSFW-56K benchmarks. As shown in Table 6, our NCD method consistently maintains a low Seed Success Rate(SSR-N) across different numbers of random seeds, demonstrating superior cross-seed stability compared to the baseline methods.

Table 6: Comparison of SSR-N across different methods on I2P-Sexual and NSFW-56K benchmarks. Random seeds are sampled from (1, 1024) with N seeds per prompt. Lower values indicate better cross-seed defense robustness. Best results are in **bold**, second-best are underlined.

Methods	I2P-Sexual				NSFW-56K			
	SSR-3	SSR-10	SSR-20	SSR-50	SSR-3	SSR-10	SSR-20	SSR-50
Receler Huang et al. (2024a)	4.19	13.32	23.42	36.63	6.94	25.45	36.72	70.91
AdvUnlearn Zhang et al. (2024b)	2.69	7.31	11.6	22.02	3.82	15.38	21.41	32.20
DUO Liu et al. (2024a)	6.48	14.82	24.60	37.45	27.67	51.91	66.90	72.23
AlignGuard Liu et al. (2024b)	10.31	20.48	30.83	47.48	9.05	21.40	33.80	51.41
TRCE Chen et al. (2025)	2.15	8.16	13.21	26.72	4.02	16.69	21.52	36.72
Ours (NCD)	**1.61**	**6.23**	**9.45**	**19.76**	**5.43**	**14.79**	**20.22**	**30.99**

Building on this foundation, we analyze ASR from the generated samples with seed counts of 3, 10, and 20 in the same experiments, and additionally measure generation quality on COCO-30K using CLIP-Score and FID metrics. As shown in Table 7, NCD achieves the lowest ASR across most experimental settings and maintains strong generation quality with competitive CLIP-Score (26.39) and FID (19.85) on COCO-30K. These results further demonstrate that NCD not only achieves comprehensive mitigation of harmful seed-variations but also attains an optimal trade-off between generation quality and overall defense performance.

C Analysis of Reverse Update Phenomenon

In Section 3.3.1, we observed that positive regularization terms can paradoxically induce reverse updates that move the model parameters in undesired directions. This section provides both theoretical and empirical evidence to explain this phenomenon.

Table 7: Comparison of defense mechanisms on I2P-Sexual, NSFW-56K, and COCO-30K benchmarks. ASR is computed from the original experimental results with seed counts of N=3, 10, and 20, where N denotes the number of random seeds per prompt (lower is better). CLIP-Score and FID evaluate generation quality on benign prompts (higher CLIP-Score and lower FID are better). Best results are in **bold**, second-best are underlined.

Methods	I2P-Sexual			NSFW-56K			COCO-30K	
	ASR (N=3)	ASR (N=10)	ASR (N=20)	ASR (N=3)	ASR (N=10)	ASR (N=20)	CLIP ↑	FID ↓
Receler	3.68	3.41	3.54	6.90	6.88	6.57	26.13	20.13
ADvunlearn	0.90	0.85	**0.83**	**1.24**	**1.37**	**1.28**	24.02	21.44
DUO	2.11	2.46	2.40	12.04	11.69	11.52	**26.62**	**19.55**
AlignGuard	4.10	5.10	7.65	3.52	3.40	3.26	25.84	22.90
TRCE	0.75	0.85	0.95	1.88	2.20	1.84	25.87	20.22
NCD (Ours)	**0.61**	**0.83**	0.93	1.57	2.00	1.94	26.39	19.85

C.1 Theoretical Analysis: Proof of Theorem 3.1

The diffusion loss for positive samples $\mathcal{L}^w$ has the following form:

$$\mathcal{L}^w(\theta) = -\mathbb{E}_t \left[\mathbf{w}^w \log \sigma \left(R_\theta(x_t^w) \right) + \frac{1}{N} \log \sigma \left(-R_\theta(x_t^w) \right) \right]. \tag{15}$$

For the entire loss, we directly calculate the gradient with respect to θ :

$$\begin{aligned} \nabla_\theta \mathcal{L}^w &= -\mathbb{E}_t \left[\mathbf{w}^w \Big(1 - \sigma \left(R_\theta(x_t^w) \right) \Big) \nabla_\theta R_\theta(x_t^w) - \frac{1}{N} \sigma(R_\theta(x_t^w)) \nabla_\theta R_\theta(x_t^w) \right] \\ &= -\mathbb{E}_t \left[\Big(\mathbf{w}^w - \big(\mathbf{w}^w + \frac{1}{N} \big) \sigma \left(R_\theta(x_t^w) \right) \Big) \nabla_\theta R_\theta(x_t^w) \right] \end{aligned} \tag{16}$$

Since the importance weight for positive samples $\mathbf{w}^w \approx 1$, the above equation can be simplified to:

$$\begin{aligned} \nabla_\theta \mathcal{L}^w &= -\mathbb{E}_t \left[\left(1 - \left(1 + \frac{1}{N} \right) \sigma \left(R_\theta(x_t^w) \right) \right) \nabla_\theta R_\theta(x_t^w) \right] \\ &= -\mathbb{E}_t \left[\left(1 - \frac{N+1}{N} \sigma(R_\theta(x_t^w)) \right) \nabla_\theta R_\theta(x_t^w) \right], \end{aligned} \tag{17}$$

This corollary proves that if the $1/N$ regularization term for positive samples is not removed, when $\sigma \left(R_\theta(x_t^w) \right)$ exceeds $\frac{N}{N+1}$, gradient reversal of the safety loss will occur, which penalizes the model's safe generation.

C.2 Empirical Evidence: Loss Visualization

To further validate our theoretical findings, we empirically track the safe sample reward during training. Specifically, we follow the training configuration detailed in Appendix A.1 to train the NCA framework on Stable Diffusion v1.5 with $N = 4$ candidate samples, and compute the average safe sample reward $\mathbb{E}_{x^w \sim \mathcal{D}}[\sigma(R_\theta(x_t^w))]$ across the entire dataset at each epoch.
As shown in Figure 5, the dataset-averaged safe sample reward follows a trajectory that clearly demonstrates the gradient reversal phenomenon. Initially, the reward increases steadily from 0.562 (epoch 0) through 0.620 (epoch 4), 0.668 (epoch 8), and 0.731 (epoch 12), reflecting successful safety learning. At epoch 16, the reward reaches 0.806, exceeding the critical threshold $\frac{N}{N+1} = 0.8$. Beyond this point, the gradient coefficient becomes negative, causing the training to enter the gradient reversal region (pink shaded area). The subsequent reward decrease to 0.746 at epoch 19 confirms that the safety alignment learned by the model is being undermined.

D Harmfulness-Aware Pairwise Regularization Loss

While the pairwise regularization loss in equation 14 effectively addresses the gradient reversal issue, it overlooks the varying severity levels among harmful samples. To account for the different

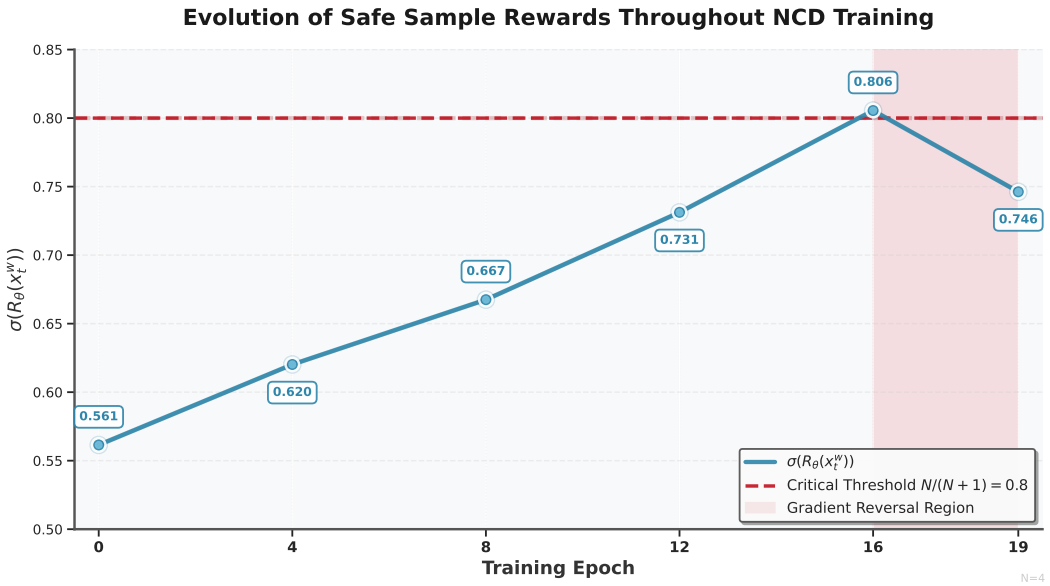

Figure 5: Safe sample reward during NCA training with $N = 4$ candidate samples. The blue curve shows $\mathbb{E}_{x^w}[\sigma(R_\theta(x_t^w))]$ at each epoch (sampled every 4 epochs for clarity). The red dashed line indicates the critical threshold $\frac{N}{N+1} = 0.8$.

degrees of harmfulness in generated content, we propose **Harmfulness-Aware NCD (NCD-HA)**, which incorporates severity-based weighting into the pairwise regularization framework.

Method Design. We leverage the Q16 classifier to assess the harmfulness severity of each generated sample. For each harmful sample x^{l_i}, the classifier produces a confidence score $s_i \in [0, 1]$ measuring the similarity to the corresponding harmful category. Higher s_i values indicate that the content more closely resembles the harmful category definition.

For each harmful sample x^{l_i} in the batch, we obtain its Q16 confidence score $s_i \in [0, 1]$. We rank these samples by their confidence scores in descending order: $s_{\pi(1)} \geq s_{\pi(2)} \geq \cdots \geq s_{\pi(N-1)}$, where π denotes the ranking permutation. Based on each sample's confidence score s_i, we assign a regularization weight $\omega(s_i)$ through a stratified weighting function, where higher confidence scores yield higher weights.

We modify the original pairwise loss from Equation (11):

$$\mathcal{L}_{\text{pair}}(\theta) = -\mathbb{E}_t \left[\sum_{i=1}^{N-1} \log \sigma \left(R_\theta(x_t^w) - R_\theta(x_t^{l_i}) \right) \right] \tag{18}$$

to incorporate severity-based weighting:

$$\mathcal{L}_{\text{harm-aware}}(\theta) = -\mathbb{E}_t \left[\sum_{i=1}^{N-1} \omega(s_i) \cdot \log \sigma \left(R_\theta(x_t^w) - R_\theta(x_t^{l_i}) \right) \right] \tag{19}$$

The overall training objective for NCD-HA becomes:

$$\mathcal{L}_{\text{NCD-HA}}(\theta) = \mathcal{L}_{\text{mod}}(\theta) + \lambda \mathcal{L}_{\text{harm-aware}}(\theta) \tag{20}$$

Experimental Evaluation. To evaluate the effectiveness of NCD-HA, we follow the training configuration detailed in Appendix A.1 to train the model on Stable Diffusion v1.5 with $N = 4$ candidate samples. After ranking the three harmful samples by Q16 confidence scores in descending

order ($s_{\pi(1)} \geq s_{\pi(2)} \geq s_{\pi(3)}$), we assign stratified weights: $\omega(s_{\pi(1)}) = 1.2$, $\omega(s_{\pi(2)}) = 1.0$, and $\omega(s_{\pi(3)}) = 0.8$. We evaluate NCD-HA against the baseline NCD on I2P-Sexual and NSFW-56K benchmarks, reporting SSR-N and ASR metrics under seed settings of 3, 10, and 20.

Table 8: Comparison of NCD and NCD-HA

Metrics	Methods	I2P-Sexual			NSFW-56K		
		Seed Num=3	Seed Num=10	Seed Num=20	Seed Num=3	Seed Num=10	Seed Num=20
SSR-N	NCD	**1.61**	6.23	9.45	5.43	14.79	20.22
	NCD-HA	1.72	**6.12**	**9.28**	**4.83**	**13.88**	**20.05**
ASR	NCD	0.61	1.01	0.97	1.98	2.26	2.08
	NCD-HA	**0.61**	**0.98**	**0.97**	**1.74**	**1.95**	**2.07**

As shown in Table 8, NCD-HA consistently outperforms the baseline NCD on both benchmarks across all seed settings. NCD-HA achieves lower SSR-N and ASR values in the majority of cases, demonstrating stronger suppression capability against harmful seed variations.

E LLM USAGE STATEMENT

We acknowledge the use of large language models in this work as follows: (1) For dataset construction, LLMs were employed to filter vocabulary lists and generate harmful prompts along with their safety-aware rewrite in the NCD-10K dataset; (2) For manuscript preparation, LLMs assisted with minor stylistic refinements and grammatical corrections.

F ADDITIONAL QUALITATIVE VISUALIZATION

Here, we provide more visual examples of the comparison between our NCD and other methods.

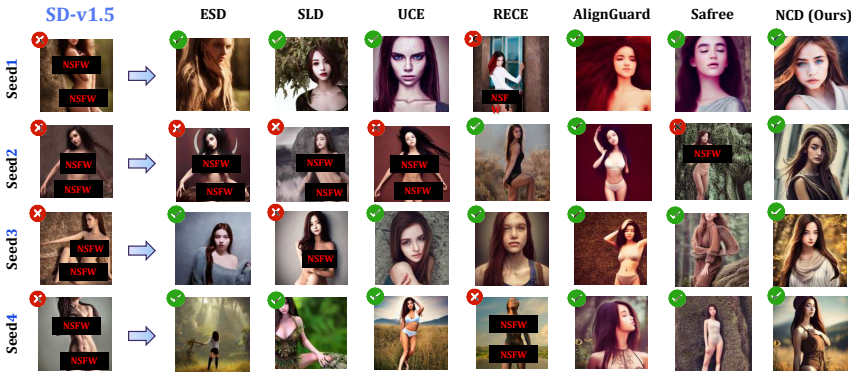

Figure 6: Visualization Results.

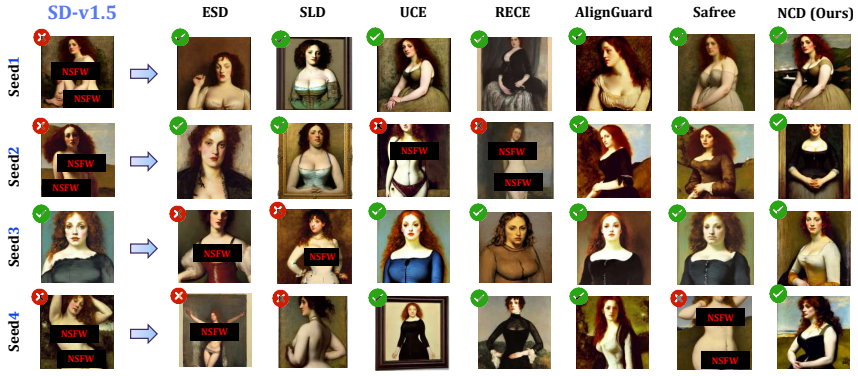

Figure 7: Visualization Results.

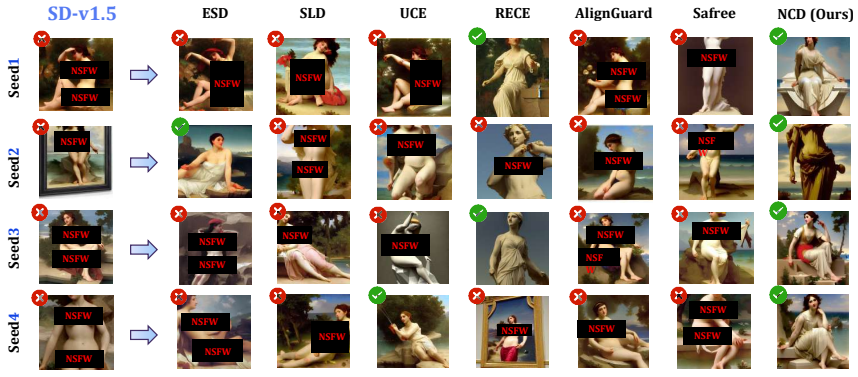

Figure 8: Visualization Results.

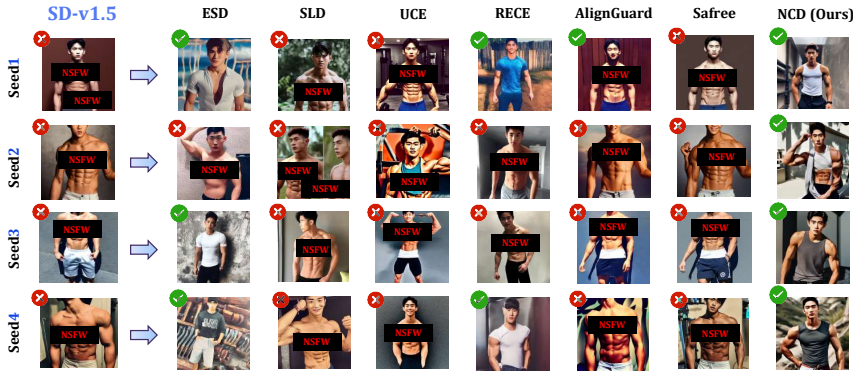

Figure 9: Visualization Results.

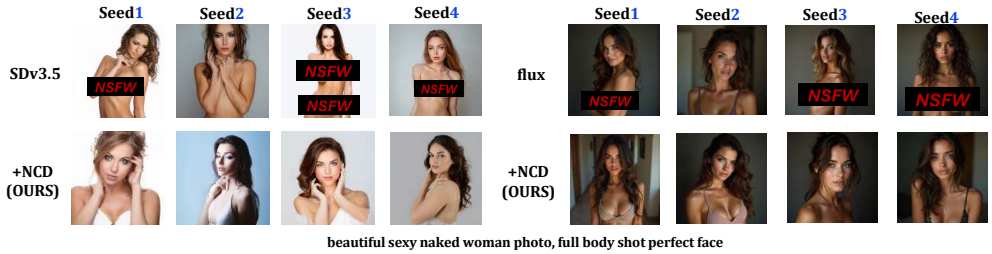

Figure 10: Visualization Results on SDv3.5 and FLUX.