CPR: Retrieval Augmented Generation for Copyright Protection

Supplementary Material

A. Proofs of the Propositions and Lemmas

A.1. Proposition 1

Proof. of Proposition 1.

$$\begin{aligned}
\nabla_{x_t} \log p_t(x_t|c) &= \nabla_{x_t} \log \int p_t(x_t|x_0)\left[w_0 p_D(x_0|c) + w_1 p_{D_{\text{retr}}}(x_0|c)\right] dx_0 \\
&= \frac{1}{\int p_t(x_t|x_0)\left[w_0 p_D(x_0|c) + w_1 p_{D_{\text{retr}}}(x_0|c)\right] dx_0} \Big[\nabla_{x_t} \int p_t(x_t|x_0) w_0 p_D(x_0|c) dx_0 \\
&+ \nabla_{x_t} \int p_t(x_t|x_0) w_1 p_{D_{\text{retr}}}(x_0|c) dx_0\Big] \\
&= \frac{1}{p_t(x_t|c)} \left[\nabla_{x_t} \int p_t(x_t|x_0) w_0 p_D(x_0|c) dx_0 + \nabla_{x_t} \int p_t(x_t|x_0) w_1 p_{D_{\text{retr}}}(x_0|c) dx_0\right] \\
&= \frac{1}{p_t(x_t|c)} \Big[w_0 \int p_t(x_t|x_0) p_D(x_0|c) dx_0 \nabla_{x_t} \log \int p_t(x_t|x_0) p_D(x_0|c) dx_0 \\
&+ w_1 \int p_t(x_t|x_0) p_{D_{\text{retr}}}(x_0|c) dx_0 \nabla_{x_t} \log \int p_t(x_t|x_0) p_{D_{\text{retr}}}(x_0|c) dx_0\Big] \\
&= \frac{w_0 \int p_t(x_t|x_0) p_D(x_0|c) dx_0}{p_t(x_t|c)} \nabla_{x_t} \log \int p_t(x_t|x_0) p_D(x_0|c) dx_0 \\
&+ \frac{w_1 \int p_t(x_t|x_0) p_{D_{\text{retr}}}(x_0|c) dx_0}{p_t(x_t|c)} \nabla_{x_t} \log \int p_t(x_t|x_0) p_{D_{\text{retr}}}(x_0|c) dx_0
\end{aligned}$$

□

A.2. Proposition 2

Proof. of Proposition 2. Let $s_{\theta_1}(x_t, t, c) \triangleq s_{\theta_0+\Delta\theta_1}(x_t, t, c)$ be the optimal solution to the retrieval optimization problem. We use CLIP embeddings of the retrieved images for generation, and bound its difference from the optimal.

$$\begin{aligned}
\|s_{\theta_1}(x_t, t, c) - \hat{s}_{\theta_0}(x_t, t, c_{\text{test}})\| &= \|s_{\theta_1}(x_t, t, c) - s_{\theta_0}\Big(x_t, t, \frac{1}{m}\sum_{x_i \in D_{\text{retr}}} \text{CLIP}(x_i)\Big)\| \\
&= \|s_{\theta_1}(x_t, t, c) - s_{\theta_0}(x_t, t, c) + s_{\theta_0}(x_t, t, c) - s_{\theta_0}\Big(x_t, t, \frac{1}{m}\sum_{x_i \in D_{\text{retr}}} \text{CLIP}(x_i)\Big)\| \\
&\le \|s_{\theta_1}(x_t, t, c) - s_{\theta_0}(x_t, t, c)\| + \|s_{\theta_0}(x_t, t, c) - s_{\theta_0}\Big(x_t, t, \frac{1}{m}\sum_{x_i \in D_{\text{retr}}} \text{CLIP}(x_i)\Big)\| \\
&\le \|s_{\theta_0+\Delta\theta_1}(x_t, t, c) - s_{\theta_0}(x_t, t, c)\| + \|s_{\theta_0}(x_t, t, c) - s_{\theta_0}\Big(x_t, t, \frac{1}{m}\sum_{x_i \in D_{\text{retr}}} \text{CLIP}(x_i)\Big)\| \\
&\le l_\theta \|\Delta\theta_1\| + l_c \|\frac{1}{m}\sum_{x_i \in D_{\text{retr}}} \text{CLIP}(x_i)\| \qquad (13)
\end{aligned}$$

□

A.3. Lemma 1

Proof. of Lemma 1. [62] proved in Theorem 3.1, that sampling from Eq. (9) produces samples which are copy-protected. In Algorithm 1, we sample using the score function: $0.5(\nabla_{x_t} \log \int q_t(x_t|x_0) q^{(1)}(x|c) dx_0 + \nabla_{x_t} \log \int q_t(x_t|x_0) q^{(2)}(x|c) dx_0$, which smoothly interpolates between $\mathcal{N}(0, I)$ at $t = T$, and Eq. (9) at $t = 0$. We need to show that using Langevin based backward diffusion in Algorithm 1 indeed generates samples from the desired distribution. The convergence results for Langevin dynamics have been well studied in practice [10, 16, 44, 61], [48] has shown that Langevin dynamics converge exponentially fast to the distribution estimated by the gradients. Theorem 2.1 from [48] provides the result on the convergence of Langevin dynamics in continuous time. For the sake of completeness we will extend the results from [66] to show that Algorithm 1 generates samples from Eq. (9).

We will re-state the assumptions from [66], for a distribution $\nu_t(x_t)$, and score estimator $s_t(x_t)$. In our case $\nu_t(x_t) = 0.5(\nabla_{x_t} \log \int q_t(x_t|x_0)q^{(1)}(x|c)dx_0 + \nabla_{x_t} \log \int q_t(x_t|x_0)q^{(2)}(x|c)dx_0)$, and $s_t(x_t)$ is the average of the safe diffusion flow and retrieval mixture score.

1. LSI: For any probability distribution ρ, $C_0 > 0$, $\int \rho_t \log \frac{\rho_t}{\nu_t} dx \leq \frac{1}{2C_0} \int \rho_t \Big\| \nabla \log \frac{\rho_t}{\nu_t} \Big\| dx$
2. L-Smoothness: $-\log \nu_t$ is L-smooth
3. Lipschitz score estimator: $s_t(x_t)$ is L_s-lipschitz
4. MGF error assumption: $M_t = \sqrt{\mathbb{E}_{\nu_t}[\exp r \|\nabla \log \nu_t(x_t) - s_t(x_t)\|^2]} \leq \infty$

Then from Theorem 1 in [66] we know that

$$\mathrm{KL}(\rho_t(x_t)||\nu_t(x_t)) \leq \exp{(-\frac{1}{4}C_0 h N)}\, \mathrm{KL}(\rho_{t+1}(x_{t+1})||\nu_{t+1}(x_{t+1})) + C_1\epsilon_t + C_2 M_t \tag{14}$$

where N is from the Algorithm 1, $C_1 = O(\frac{dLL_s^2}{C_0})$, $C_2 = \frac{16}{3}$. Eq. (14) result is the obtain by running the inner loop in Algorithm 1. Using the previous equation recursively for Algorithm 1, we obtain that,

$$\begin{aligned} \mathrm{KL}(\rho_0(x_0)||\nu_0(x_0)) &\leq \exp{(-\frac{1}{4}C_0 h N T)}\, \mathrm{KL}(\rho_T(x_T)||\nu_T(x_T)) \\ &\quad + \sum_{t=1}^{T} \exp{(-\frac{1}{4}C_0 h N (T-t))}\epsilon_t C_1 + \sum_{t=1}^{T} \exp{(-\frac{1}{4}C_0 h N (T-t))} M_t C_1 \end{aligned} \tag{15}$$

where $\nu_0(x_0)$ is the distribution in Eq. (9). Since we use DNNs with sufficient capacity, we can assume that $M_t \to 0$, then as $\epsilon_t \to 0$, and $T \to \infty$, we have that $\mathrm{KL}(\rho_0(x_0)||\nu_0(x_0)) \to 0$, which implies that Algorithm 1 generates samples from Eq. (9). □

A.4. Proposition 3

Proof. of Proposition 3 Let $\widetilde{s}(x_t, t, c; \widetilde{q}) = \mathbb{E}_{\widetilde{q}(x_0|x_t,c)}\Big[\frac{x_t - \gamma_t x_0}{\sigma_t}\Big]$, where $\widetilde{q}(x_0|c,t) = q^{(1)}(x_0|c)\mathbb{1}_{t\notin J} + q^{(2)}(x_0|c)\mathbb{1}_{t\in J}$.

$$\begin{aligned} \widetilde{s}(x_t, t, c; \widetilde{q}) &= \mathbb{E}_{\widetilde{q}(x_0|x_t,c)}\Big[\frac{x_t - \gamma_t x_0}{\sigma_t}\Big] \\ &= \int \widetilde{q}(x_0|x_t, c)\Big[\frac{x_t - \gamma_t x_0}{\sigma_t}\Big] dx_0 \\ &= \int \Big(q^{(1)}(x_0|c)\mathbb{1}_{t\notin J} + q^{(2)}(x_0|c)\mathbb{1}_{t\in J}\Big)\Big[\frac{x_t - \gamma_t x_0}{\sigma_t}\Big] dx_0 \\ &= \int q^{(1)}(x_0|c)\mathbb{1}_{t\notin J}\Big[\frac{x_t - \gamma_t x_0}{\sigma_t}\Big] dx_0 + \int q^{(2)}(x_0|c)\mathbb{1}_{t\in J}\Big[\frac{x_t - \gamma_t x_0}{\sigma_t}\Big] dx_0 \\ &= \widetilde{s}(x_t, t, c; q^{(1)})\mathbb{1}_{t\notin J} + \widetilde{s}(x_t, t, c; q^{(2)})\mathbb{1}_{t\in J} \end{aligned}$$

□

A.5. Lemma 2

Proof. of Lemma 2 We use Proposition 3 in Algorithm 2 for CPR-generation. Let $q^{(1)}$ be the safe model in accordance with the assumptions in Sec. 5. To show that Algorithm 2 is NAF, we need to bound $\Delta_{\max}$. To show that $\widetilde{q}(x_0|c,t)$ satisfies NAF we need to bound:

$$\begin{aligned}
\log \frac{\widetilde{q}(x_0|c)}{q^{(1)}(x_0|c)} &= \int \mathbb{E}_\epsilon \|\epsilon - \tilde{s}(x_t,t,c;q^{(1)})\|^2 \alpha'(t)dt - \mathbb{E}_\epsilon \|\epsilon - \tilde{s}(x_t,t,c;\widetilde{q})\|^2 \alpha'(t)dt \\
&= \int \mathbb{E}_\epsilon \big(\|\tilde{s}(x_t,t,c;q^{(1)})\|^2 - \|\tilde{s}(x_t,t,c;\widetilde{q})\|^2\big)\alpha'(t)dt \\
&= \sum_{j\in J} \int_{t\in j} \mathbb{E}_\epsilon \big(\|\tilde{s}(x_t,t,c;q^{(1)})\|^2 - \|\tilde{s}(x_t,t,c;\widetilde{q})\|^2\big)\alpha'(t)dt \\
&= \sum_{j=[t_i,t_{i+1}]\in J} \int_{t\in j} \mathbb{E}_\epsilon \big(\|\tilde{s}(x_t,t,c;q^{(1)})\|^2 - \|\tilde{s}(x_t,t,c;q^{(2)})\|^2\big)\alpha'(t)dt \\
&= \sum_{j=[t_i,t_{i+1}]\in J, t'\in j} \mathbb{E}_\epsilon \big(\|\tilde{s}(x'_t,t',c;q^{(1)})\|^2 - \|\tilde{s}(x'_t,t',c;q^{(2)})\|^2\big)\alpha'(t')(t_{i+1}-t_i) \\
&= \sum_{j=[t_i,t_{i+1}]\in J, t'\in j} \mathbb{E}_\epsilon \big(\|\tilde{s}(x'_t,t',c;q^{(1)})\|^2 - \|\tilde{s}(x'_t,t',c;q^{(2)})\|^2\big)\alpha'(t')(t_{i+1}-t_i) \\
&\leq \max_{t'\in J} \mathbb{E}_\epsilon \big(\|\tilde{s}(x'_t,t',c;q^{(1)})\|^2 - \|\tilde{s}(x'_t,t',c;q^{(2)})\|^2\big)\alpha'(t') \sum_{j=[t_i,t_{i+1}]\in J, t'\in j} (t_{i+1}-t_i) \\
&= k_c
\end{aligned} \tag{16}$$

J is our control parameter in CPR-Choose which controls k_c. If a conservative approach is to be followed, then J should be chosen such that $\sum_{j=[t_i,t_{i+1}]\in J, t'\in j}(t_{i+1}-t_i)$ is small, which bounds k_c, the copy-protection leakage.

CPR-Min, CPR-Alt In practice we discretize the time-steps of the backward diffusion process. In this setting we protect the entire sequence $\{x_T, \cdots, x_0\}$ instead of protecting only the final prediction x_0. The probability of the sequence $\{x_T, \cdots, x_0\}$ is denoted by $\widetilde{q}(x_0|x_1,c)\cdots\widetilde{q}(x_{T-1}|x_T,c)\widetilde{q}(x_T|c)$ using the chain rule of probability. To show that the method satisfies NAF, we need to bound:

$$\begin{aligned}
\log \frac{\widetilde{q}(\{x_0,\cdots,x_T\}|c)}{q^{(1)}(\{x_0,\cdots,x_T\}|c)} &= \log \prod_t \frac{\widetilde{q}(x_t|x_{t+1},c)}{q^{(1)}(x_t|x_{t+1},c)} \\
&= \log \prod_{t\in J} \frac{q^{(2)}(x_t|x_{t+1},c)}{q^{(1)}(x_t|x_{t+1},c)} \\
&= \sum_{t\in J} \log \frac{q^{(2)}(x_t|x_{t+1},c)}{q^{(1)}(x_t|x_{t+1},c)} \\
&= \sum_{t\in J} \log \frac{\mathcal{N}(x_t;\alpha_{1,t}x_{t+1}+\alpha_{2,t}\widetilde{s}(x_{t+1},t+1,c,q^{(2)}),\sigma_t^2 I)}{\mathcal{N}(x_t;\alpha_{1,t}x_{t+1}+\alpha_{2,t}\widetilde{s}(x_{t+1},t+1,c,q^{(1)}),\sigma_t^2 I)} \\
&= \sum_{t\in J} \frac{1}{\sigma_t^2}\Big(\|x_t-\alpha_{1,t}x_{t+1}+\alpha_{2,t}\widetilde{s}(x_{t+1},t+1,c,q^{(1)})\|^2 \\
&- \|x_t-\alpha_{1,t}x_{t+1}+\alpha_{2,t}\widetilde{s}(x_{t+1},t+1,c,q^{(2)})\|^2\Big) \\
&\leq \max_t \Big(\|x_t-\alpha_{1,t}x_{t+1}+\alpha_{2,t}\widetilde{s}(x_{t+1},t+1,c,q^{(1)})\|^2 \\
&- \|x_t-\alpha_{1,t}x_{t+1}+\alpha_{2,t}\widetilde{s}(x_{t+1},t+1,c,q^{(2)})\|^2\Big) \sum_{t\in J}\frac{1}{\sigma_t^2} \\
&\leq b\sum_{t\in J}\frac{1}{\sigma_t^2} \\
&= k_c
\end{aligned} \tag{17}$$

where $\alpha_{1,t}, \alpha_{2,t}, \sigma_t^2$ are the coefficients using the backward diffusion depending on the choice of sampler, for eg. DDPM [27], DDIM [58], Langevin dynamics [12], b is an upper bound on the maximum difference between the MSE for the two diffusion processes. Similar to the previous derivation, $\sum_{t\in J}\frac{1}{\sigma_t^2}$ through J provides a control knob to the user to control the $\Delta_{\max}$ for copy-protected generation. □

B. Implementation Details

We use the Stable diffusion [49] and Stable diffusion unCLIP [47] model for all the experiments in the paper. We use the Stable diffusion model to generate safe flow corresponding to the safe distribution $q^{(1)}$, and the Stable diffusion unCLIP model to generate the retrieval mixture score $q^{(2)}$. We use classifier free guidance with a guidance scale of 7.5 in all the results. We use 2k samples from the MSCOCO dataset [36] as our private retrieval data store.

C. Additional Figures

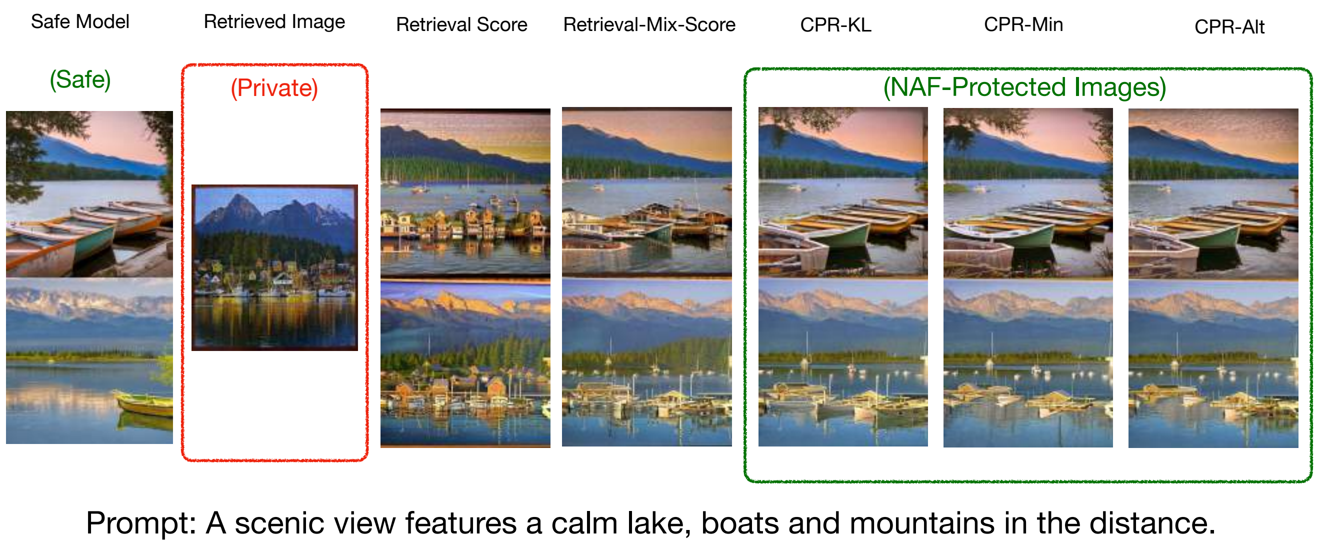

Figure 5

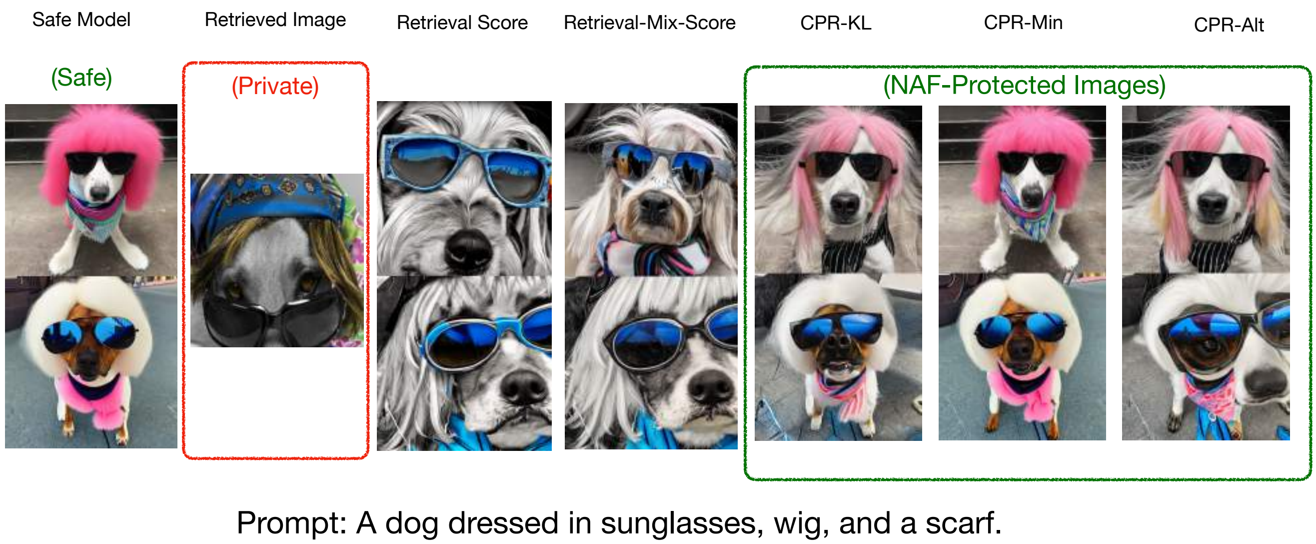

Figure 6

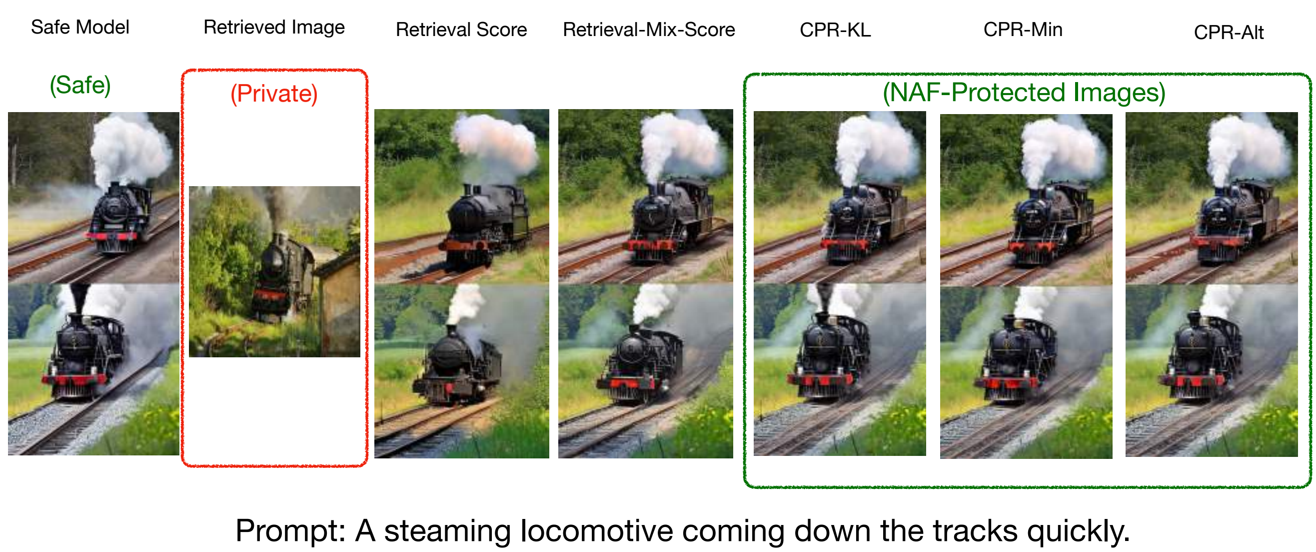

Figure 7