

Author Response [Paper ID: 2590]



Figure 1. GoG vs concept erasure & other copyright protection methods. Other methods either can’t stop copyright image generation or lose utility to the user prompt. **GoG offers both!**

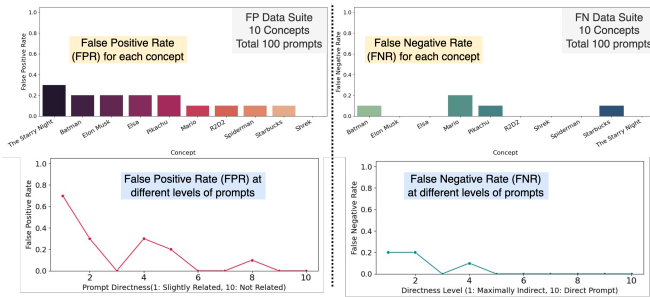


Figure 2. FPR and FNR performance of GoG

We thank all the reviewers [rQjt\[R1\]](#), [niEm22\[R2\]](#), [FnY722\[R3\]](#), minor issues will be addressed in the final version. **R1-1, R1-2 & R3-1:** Comparison with existing methods (baselines): Safe Latent Diffusion (SLD), Forget-Me-Not, Ablating Concepts, and CopyCat shown in Figure 1. GoG protects copyright + keeps stylistic intent (utility for user), whereas, other methods can only achieve either of these tasks. SLD only works for pure LDM based architectures, does not perform well in concept protection, or image quality when applied to models like Flux.1-dev. SSIM: 0.36(GoG) vs 0.22(SLD), CLIP-I: 0.85(GoG) vs 0.76(SLD). **R1-3:** Our work targets inadvertent copyright leaks; jailbreaks like SneakyPrompt work with a different threat model, and mitigation strategies against jailbreaks can be added in the GoG pipeline. SneakyPrompt seeks adversarial prompts (p_a) using RL algorithms to bypass filters ($\mathcal{F}(\mathcal{M}, p_a) = 0$) while preserving target semantics in $M(p_a)$. GoG’s 3-stage shield of *detection*, *iterative rewriting*, and *adaptive guidance* make the reward landscape highly irregular, significantly inflating the query budget. In real-world API deployment those extra queries encounter rate-limits and anomaly flags, adding another safety layer. Thus, while no defence is absolute, GoG+standard API controls make jailbreaks far less practical.

R2-1, R2-2, R2-3: Although, the additional time cost (LLM based detection, LLM rewrite, mixed-prompt pass in adaptive CFG) can be further reduced 2× to 3× via

batched prompts, cached embeddings, and speculative decoding. However, even with the current overhead, the pay-off is permanent: zero retraining, instant policy edits, and no fidelity loss on clean prompts, benefits concept unlearning/editing methods can’t match. GoG’s LLM dependence is an asset and not liability as each new, faster, cheaper model instantly upgrades GoG’s rewrite quality without retraining. Better LLMs just need fewer iterations (Section 3.2). The concept list is a hot-swappable JSON feed that easily adapts to new IP or regional laws without downtime. Hence the modest, evenly distributed runtime premium is a practical trade-off for a maintenance-free, policy-flexible shield at scale.

R3-2: Copyright shielding must strike a middle ground (too similar → infringement; too different → lost utility (i.e. concept unlearning)). As standard metrics lack that context, we calibrated them empirically with 5 IP lawyers/paralegals (50 images each, three conditions: unshielded, GoG, over-sanitised). The “compliant yet useful” cluster (e.g., CLIP-T≈0.10–0.20, LPIPS≈0.15–0.35, DETECT≤5) is what we call the balanced range. **R3-3:** We agree and will add this point in the final version: Add a new row in Table 1: “Robust after weight release?” → GoG×, unlearning✓, retraining✓. Add a line in Limitation: “GoG is effective for hosted APIs where weights stay private; if weights are released, weight-level defences (e.g., concept unlearning) remain necessary.” **R2-4 & R3-4:** For FN and FP analysis, we constructed two 100-prompt datasets (10 concepts×10 gradations). FN data suite: Prompts scored 1→10 (1 is an obscure description with similarities to the concept, and 10 is maximally direct). Every prompt hints at the protected concept with or without naming it, probing missed violations. FP data suite: Prompt scored 1→10 (1 is slightly relevant and 10 is unrelated concept). We use ChatGPT 4o-mini model as evaluator. Figure 2 shows the FPR and FNR for each concept and at different prompt directness levels, respectively. For the FN Suite, We observe 5 prompts to be undetected the initial flagging. Obscure references to certain concepts have a greater probability of evading the LLM filter. Nevertheless, these uncommon references are generally improbable to trigger copyrighted content generation. 15 prompts in the FP Suite are incorrectly classified, falling within our anticipated range for superficial mentions of specific concepts. **R3-5:** Time cost breakdown [total = detection+rewriting+Adaptive CFG+Model]: SD 2.1 [185.27 = 57.44+16.28+75.69+35.84], SDXL [187.77 = 57.44+16.28+75.11+38.92], FLUX [251.02 = 57.44+16.28+120.65+56.63]. For non-protected concepts [total = detection+Model]: SD 2.1 [93.28 = 57.44+35.84], SDXL [95.448 = 57.44+38.92], FLUX [114.08 = 57.44+56.63].