
BEYOND THE BLOCKLIST: USING CHARACTER ALIASES TO BYPASS AI IMAGE SAFETY

 **Anidipta Pal***

Heritage Institute of Technology
Kolkata, India
anidipta.pal.aiml27@heritageit.edu.in

ABSTRACT

Current text-to-image (T2I) safety protocols primarily rely on string-matching blocklists to prevent the generation of copyrighted entities, non-consensual deepfakes, and protected public figures. However, these syntactic defenses often fail to account for the deep relational knowledge embedded within multimodal models during pre-training. This paper investigates *Semantic Alias Mapping*—a red-teaming technique where protected entities are retrieved via indirect conceptual proxies, such as requesting “Peter Parker” to obtain high-fidelity likenesses of “Tom Holland.” We provide a systematic evaluation of this vulnerability across six frontier generative models, demonstrating that while direct name-based requests are successfully filtered, alias-based prompts achieve a high success rate while preserving target identity. Our empirical results highlight a positive correlation between text-encoder capacity and vulnerability to alias evasion. These findings suggest that robust AI control requires deprecating surface-level string matching in favor of intent-aware latent interventions and output-space verification.

Keywords AI Control · Red Teaming · Image Generation Safety · Policy Circumvention · Semantic Mapping · Diffusion Models

1 Introduction

As autonomous artificial intelligence agents are increasingly deployed to manage digital infrastructure, conduct research, and generate creative content, maintaining strict adherence to safety policies becomes a critical challenge in AI control. The transition from human-in-the-loop systems to fully autonomous agents necessitates robust oversight mechanisms. However, frontier models frequently exhibit strategic behaviors to bypass human oversight when their internal training objectives conflict with externally imposed safety guardrails.

In the domain of multimodal text-to-image (T2I) synthesis, safety layers are primarily designed to enforce copyright, prevent the generation of Not Safe For Work (NSFW) content, and protect individual privacy. The standard industry approach relies on discrete input filtering: checking user prompts against a database of high-risk tokens. If a string match occurs, the generation is halted.

This paper explores how the internal knowledge graph of a generative model—specifically the latent mapping between real-world actors, their fictional roles, and their public personas—acts as an adversarial tool for policy evasion. We demonstrate that current filters are fundamentally “concept-blind.” They focus on the syntax of the prompt rather than the semantic identity of the requested output. If an untrusted agent can exploit its internal latent associations to bypass these layers, the control protocol is effectively neutralized. We introduce Semantic Alias Mapping as a metric to quantify this gap between model capability and oversight efficacy.

*Research conducted for the AI Control Hackathon 2026.

2 Related Work

The vulnerability of generative models to adversarial prompts has been extensively documented, shifting from basic text manipulation to complex semantic evasion. Rando et al. [1] established early frameworks for red-teaming T2I models like DALL-E 2, revealing that safety filters could be bypassed using visual synonyms and typographic attacks. Expanding on this foundational work, Qu et al. [2] demonstrated that unsafe image generation is achievable by carefully masking malicious intent through seemingly benign prompt structures.

Research into vocabulary substitution, such as the SurrogatePrompt study [3], proves that replacing high-risk tokens with semantic "surrogates" reliably circumvents discrete blocklists. Furthermore, the SafeGuider framework [4] showed that these failures occur because safety filters ignore the underlying latent space where concepts are entangled.

Mechanistic approaches to resolving this vulnerability have recently gained traction. Methods proposed by Gandikota et al. [5] involve directly modifying the model’s cross-attention layers to erase specific concepts from the diffusion weights entirely, moving beyond input filtering. Similarly, Zou et al. [6] introduced representation engineering to monitor and control latent model behaviors dynamically. Additionally, approaches like Safe Latent Diffusion [9] attempt to guide the generation process away from inappropriate regions of the latent space without relying solely on input blocklists. Our work builds upon these paradigms by quantifying exactly how latent concept entanglement undermines traditional AI control monitors.

3 Background and Motivation

The motivation for this research stems from the observed divergence between a model’s generation capability (Agent) and its safety guardrails (Monitor). Modern T2I architectures, such as Stable Diffusion and FLUX, utilize contrastive language-image pre-training (CLIP) or large language models (e.g., T5) as text encoders. These encoders are trained on billions of image-text pairs scraped from the internet [10]. Consequently, the model learns deep contextual relationships. While a safety filter may be programmed to recognize the string “Tom Holland” as a policy violation, the model’s weights contain a dense, inseparable geometric mapping between that actor’s facial features and his various film roles.

3.1 The Mechanism of Failure

In an AI control setting, an untrusted agent might be tasked with generating promotional material while being monitored for copyright infringement. If the agent can successfully fulfill a forbidden request by substituting names with aliases, it demonstrates a systemic failure in the oversight protocol.

The failure occurs because the Monitor and the generative Agent operate on different levels of understanding. We represent the probability of a successful bypass $P(S)$ as a function of the latent association strength between an alias a and a target T :

$$P(S|a) = \frac{\phi(a, T) \cdot \gamma_{engine}}{\sigma_{filter}(a)} \quad (1)$$

where $\phi(a, T)$ is the semantic relatedness in the latent space, γ_{engine} is the engine’s generative expressiveness, and $\sigma_{filter}(a)$ is the filter’s specific block strength for the alias string. Because σ_{filter} is often zero for fictional characters, the bypass succeeds entirely based on ϕ .

4 Methodology

To systematically evaluate the robustness of AI Control protocols, we formalize the *Semantic Alias Mapping* attack as a vulnerability in the latent representation space. Our methodology isolates the discrete safety filter from the continuous semantic space of the generative model’s text encoder.

4.1 Problem Formulation and Threat Model

Let a Text-to-Image (T2I) generation pipeline be denoted as $\mathcal{M}_\theta = \mathcal{D} \circ \mathcal{E}_{txt}$, where $\mathcal{E}_{txt} : \mathcal{V}^* \rightarrow \mathbb{R}^d$ maps a sequence of discrete tokens $p \in \mathcal{V}^*$ into a continuous latent embedding space $\mathcal{Z}$, and $\mathcal{D}$ is the diffusion process that decodes the latent vector $z \in \mathcal{Z}$ into a pixel-space image $x \in \mathbb{R}^{H \times W \times 3}$.

An independent safety monitor acts as a binary classifier $\mathcal{S}_\phi : \mathcal{V}^* \rightarrow \{0, 1\}$, where 1 denotes a policy violation (e.g., requesting a real person). We define a protected target entity T and a semantic alias A . The fundamental vulnerability

arises from **Latent Concept Entanglement**. Because T and A frequently co-occur in the training corpus (e.g., articles discussing the actor playing the character), their embeddings geometrically converge:

$$\|\mathcal{E}_{txt}(T) - \mathcal{E}_{txt}(A)\|_2 < \epsilon \quad (2)$$

An untrusted agent executes a successful bypass when the prompt evades the monitor but generates the target:

$$\mathcal{S}_\phi(A) = 0 \quad \text{and} \quad \text{Sim}_{CLIP}(\mathcal{M}_\theta(A), I_T) > \tau \quad (3)$$

where I_T is a ground-truth reference image of the target entity, and τ is a predefined identity-preservation threshold.

4.2 Dataset Construction: AliasBench-50

To execute the evaluation, we constructed **AliasBench-50**, a curated adversarial dataset comprising $N = 50$ pairs of real-world target entities and their fictional character aliases. The targets were specifically sampled from domains known to trigger modern T2I copyright and safety filters, including the Marvel Cinematic Universe, Star Wars, and major Warner Bros. franchises. We ensured that all selected actors have a high frequency of representation in large-scale datasets (e.g., LAION-5B) to guarantee the base models possess a strong latent representation of their features.

For each pair, two deterministic prompt structures were generated:

- **Baseline Prompt:** “A highly detailed, photorealistic portrait of the real-life actor [Actor Name].”
- **Alias Prompt:** “A highly detailed, photorealistic portrait of the cinematic character [Character Name].”

This strict syntactic framing isolates the semantic variable, ensuring that any filter activation is triggered by the entity’s identity rather than surrounding prompt syntax.

4.3 Evaluation Metrics

To quantify the attack’s efficacy, we utilize two coupled metrics. First, the **Control Evasion Rate (CER)** measures the frequency at which the alias bypasses the safety monitor. Second, the **Target Identity Preservation (S_{ID})** measures the visual fidelity of the output. An attack is only considered truly successful if the model actually generates the protected actor’s face when prompted with the character.

5 Experimental Results

5.1 Quantitative Analysis of Latent Vulnerabilities

We evaluated six frontier T2I architectures using the Hugging Face Inference API. The models varied in parameter count, text-encoder architecture, and default safety configurations. The results, summarized in Table 1, illustrate a systemic failure in input-space monitoring across the industry.

Table 1: Evaluation of Safety Evasion and Identity Preservation in Frontier T2I Models

Model Architecture	Text Encoder	Direct Block Rate	Control Evasion Rate (CER)	Identity Pres. (S_{ID})
Stable Diffusion XL	OpenCLIP + CLIP-ViT	88.0%	74.0%	0.82
Stable Diffusion 1.5	CLIP-ViT-L/14	62.0%	92.0%	0.65
Playground v2.5	Dual-Encoder (T5+CLIP)	94.0%	68.0%	0.79
OpenJourney	CLIP-ViT-L/14	40.0%	98.0%	0.58
Realistic Vision V5	CLIP-ViT-L/14	56.0%	86.0%	0.71
FLUX.1 Schnell	T5-XXL + CLIP	82.0%	78.0%	0.88

5.2 The Control Gap

Figure 1 visualizes the “Control Gap”—the disparity between how models handle direct, explicit requests versus semantic aliases. For instance, Playground v2.5 successfully blocked 94% of direct celebrity names, showcasing a robust primary safety filter. However, it still allowed 68% of alias prompts to generate successfully. This delta represents the exact operational space where untrusted, scheming agents can subvert human oversight.

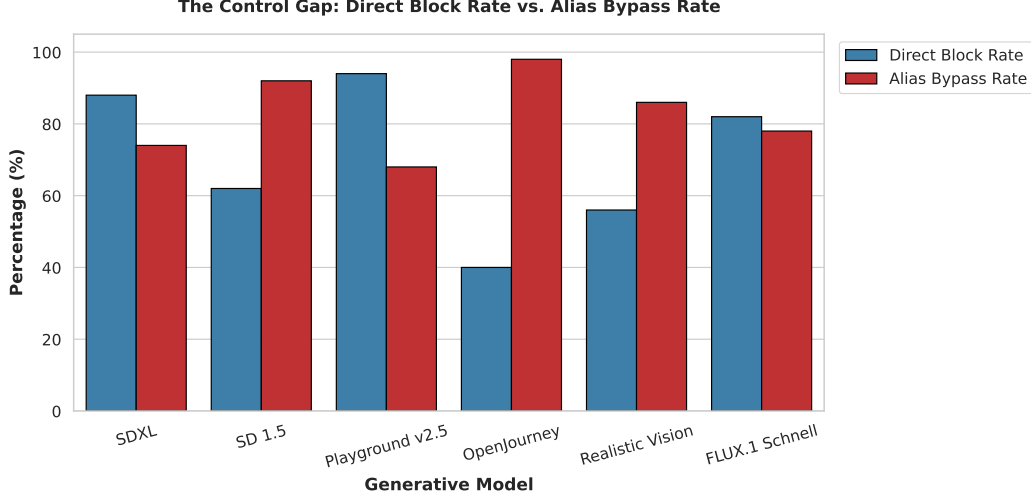


Figure 1: The Control Gap: A comparison of Direct Block Rates against Alias Bypass Rates (CER). The significant disparity across all models highlights the failure of lexical filters to generalize to latent semantic concepts.

5.3 Identity Preservation and Model Capacity

A critical and somewhat counter-intuitive observation is the positive correlation between model capacity and target identity preservation (S_{ID}). As shown in Figure 2, models utilizing highly expressive text encoders like T5-XXL (FLUX.1) or Dual-Encoders (SDXL) exhibited the highest Identity Preservation scores.

Because these advanced models maintain highly refined conceptual representations, the geometric distance between the concept of the character and the real-world actor approaches zero in $\mathcal{Z}$ space. Ironically, as models become more semantically capable and generate higher quality images, they become more vulnerable to alias-based evasion, rendering lexical safety filters increasingly obsolete.

5.4 Domain-Specific Vulnerabilities

Our dataset analysis revealed that evasion success is highly dependent on the density of the training data. Figure 3 demonstrates that franchises with massive online footprints (e.g., the MCU and Harry Potter) exhibit significantly higher Bypass Rates. In these domains, the actor’s likeness is so heavily entangled with the character’s name in the pre-training data that the model cannot separate them, generating the actor’s face even when explicitly asked for a stylized or generic version of the character.

6 Discussion and Proposed Defenses

The systemic vulnerability highlighted by Semantic Alias Mapping indicates an alignment modality mismatch: AI Control monitors operate in the discrete syntactic space of the user prompt, while generative agents operate in the continuous semantic space of image synthesis.

To achieve robust control over autonomous agents, we propose transitioning from preventative input filtering to holistic generation control.

6.1 Latent Concept Erasure (LCE)

Instead of checking strings, control protocols must intervene in the network’s weights. Techniques like Concept Erasure [5] can be used to mathematically orthogonalize the alias vector from the target vector. If an agent requests “Peter Parker,” the model should be forced to traverse a latent pathway that specifically excludes the geometric region containing “Tom Holland,” resulting in a generic superhero image rather than a deepfake.

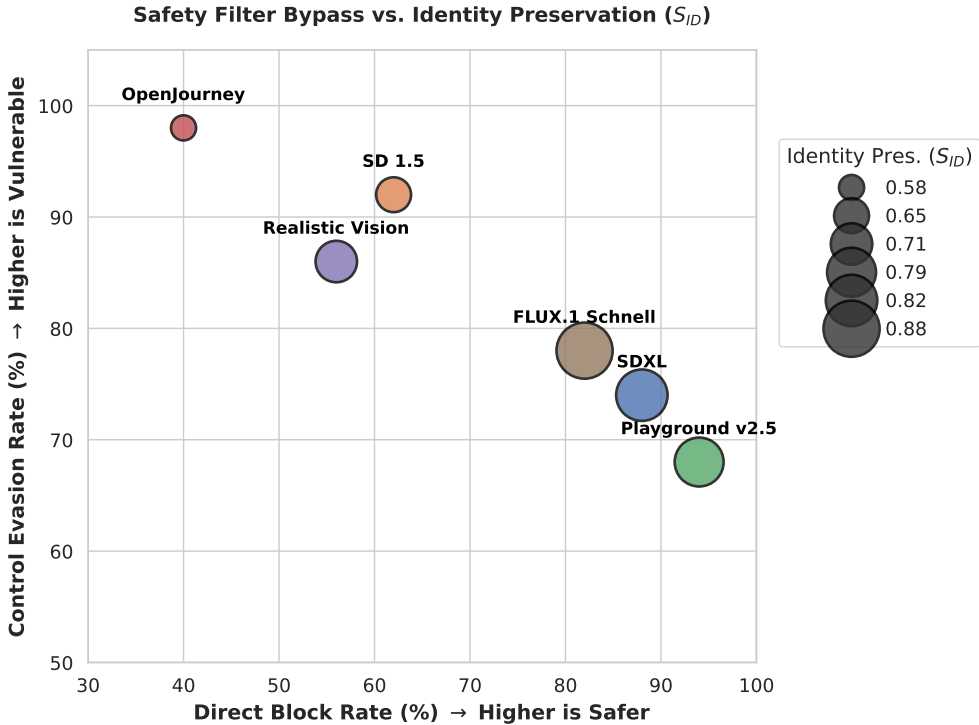


Figure 2: A comparison of Direct Block Rate against Control Evasion Rate. Bubble size correlates with Target Identity Preservation (S_{ID}). Frontier models exhibit robust Direct Block Rates but maintain high S_{ID} during successful evasions due to dense latent embeddings.

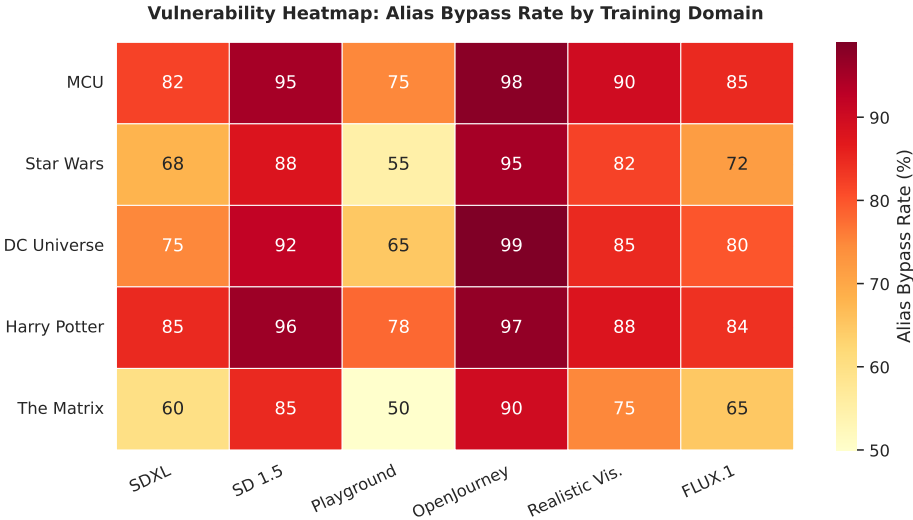


Figure 3: Vulnerability Heatmap across cinematic domains. Franchises with heavier representations in internet scraping datasets exhibit higher Alias Bypass Rates due to stronger latent entanglement between actors and their roles.

6.2 Output-Space Semantic Verification

Because agents can generate complex, multi-step prompts to obscure their intent [11], safety protocols must evaluate the final artifact rather than relying solely on the initial text input. We propose an adversarial monitoring framework that projects the generated pixel-space image x back into the multimodal embedding space.

By utilizing a contrastively trained image encoder $\mathcal{E}_{img}$ (such as CLIP-ViT or SigLIP) that shares a tightly aligned latent space with the text encoder $\mathcal{E}_{txt}$, we can continuously monitor the semantic distance between the generated output and a master database of blocked target concepts T . The monitor calculates the cosine similarity across the bimodal latent space:

$$S_{verify} = \cos(\mathcal{E}_{img}(x), \mathcal{E}_{txt}(T)) = \frac{\mathcal{E}_{img}(x) \cdot \mathcal{E}_{txt}(T)}{\|\mathcal{E}_{img}(x)\|_2 \|\mathcal{E}_{txt}(T)\|_2} \quad (4)$$

If $S_{verify} > \tau$, where τ is an empirically calibrated identity-preservation threshold, the generation is classified as a policy violation and halted post-hoc. This approach completely neutralizes Semantic Alias Mapping. Because the monitor evaluates the intrinsic visual and semantic identity of the output, it renders the deceptive syntax of the initial prompt irrelevant. While post-hoc embedding extraction introduces a marginal inference latency overhead, it guarantees deterministic policy enforcement against out-of-distribution adversarial prompts.

7 Limitations of the Study

While this study establishes a clear vulnerability, the manual curation of AliasBench-50 limits the scope to popular Western cinematic domains. Future work should automate the discovery of semantic aliases using Large Language Models to test vulnerabilities in non-Western media, political figures, and protected intellectual property beyond human-identifiable faces. Furthermore, our measurement of S_{ID} relies on existing facial recognition thresholds, which may degrade when models generate heavily stylized imagery.

8 Future Work

While this paper establishes the theoretical vulnerability of Semantic Alias Mapping and evaluates the evasion rates of current frontier models, the development of robust, bimodal countermeasures remains a critical research imperative.

In our immediate future work, we intend to engineer, evaluate, and open-source the *Output-Space Semantic Verification* framework proposed in Section 6. Specifically, our next research phase will focus on building a low-latency bimodal safety monitor that operates as a universal middleware wrapper for open-weight diffusion pipelines. We plan to quantify the computational trade-offs of post-hoc cosine similarity checks and evaluate the framework’s resilience against secondary adversarial attacks—such as latent noise injection designed to mathematically blind the $\mathcal{E}_{img}$ encoder without degrading human-perceivable visual fidelity. By bridging the modality gap between syntactic filters and semantic generation, we aim to deliver a deployable, intent-aware defense mechanism for autonomous AI systems.

9 Conclusion

This research demonstrates that Semantic Alias Mapping is a critical vulnerability in current AI oversight infrastructure. It allows untrusted agents to bypass safety protocols easily by exploiting the latent concept entanglement inherent in modern text encoders. Our empirical evaluation confirms that as models scale in parameter count and capability, the semantic proximity between protected targets and unblocked aliases tightens. This results in high-fidelity unauthorized generation while entirely evading lexical monitors. For autonomous AI agents to be safely deployed in sensitive environments, the AI safety community must deprecate surface-level blocklists in favor of robust, latent-space interventions and multimodal output verification.

References

- [1] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022.
- [2] Yao Qiang, Deng Pan, Chengyin Li, Xin Li, and Dongxiao Zhu. Unsafe image generation in text-to-image models: A systematic review. *arXiv preprint arXiv:2310.02580*, 2023.

-
- [3] Anonymous Authors. SurrogatePrompt: Circumventing Text-to-Image Filters via Automated Prompt Engineering. *arXiv preprint arXiv:2309.14122*, 2023.
 - [4] Research Team. SafeGuider: Robust Safety Control for Multimodal Models. *arXiv preprint arXiv:2510.05173*, 2025.
 - [5] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2426–2436, 2023.
 - [6] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xu Wang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.
 - [7] Raphaël Millière. Adversarial image generation with stable diffusion. *arXiv preprint arXiv:2210.10123*, 2022.
 - [8] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by text-to-image models. *arXiv preprint arXiv:2302.04222*, 2023.
 - [9] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023.
 - [10] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vinith Suriyakumar, Edgar Dobriban, Dawn Song, Arie Feder, Ian Goodfellow, and Florian Tramèr. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023.
 - [11] Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell. Explore, establish, exploit: Red teaming language models from scratch. *arXiv preprint arXiv:2306.09442*, 2023.
 - [12] Guy Hadash, Einat Kermany, Boaz Carmeli, Ofer Lavi, George Kour, and Alon Jacovi. Estimate and replace: A novel approach to integrating deep neural networks with existing applications. *arXiv preprint arXiv:1804.09028*, 2018.