



Explicit Content Generation in Text-to-Image Models: Training Mechanisms, Moderation Failures, and Safety Limits

Yajnesh Rao

Independent Researcher, Data Engineer Karnataka, India.

To Cite this Article: Yajnesh Rao, “Explicit Content Generation in Text-to-Image Models: Training Mechanisms, Moderation Failures, and Safety Limits”, Indian Journal of Computer Science and Technology, Volume 05, Issue 02 (May-August 2026), PP: 1005-1020.



Copyright: ©2026 This is an open access journal, and articles are distributed under the terms of the [Creative Commons Attribution License](#); Which Permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract: Text-to-image generative models have rapidly progressed in realism and controllability, enabling wide adoption across creative and commercial domains. However, these systems also generate explicit and harmful imagery in ways that are difficult to prevent reliably. This paper reviews the technical foundations of modern image generators, with emphasis on diffusion-based modeling, latent representations, and multimodal text-image alignment. It then surveys documented moderation approaches and explains why common safeguards fail under paraphrasing, adversarial prompting, long-tail semantics, and distribution shift. Finally, the paper discusses safety limitations that arise from likelihood-based learning objectives and proposes directions for more robust mitigation, including stronger data governance, multimodal safety alignment, evaluation-driven red teaming, and lifecycle monitoring.

Keywords: text-to-image generation; diffusion models; CLIP; content moderation; AI safety; NSFW detection; alignment

I. INTRODUCTION

A. Motivation and Problem Statement

Recent advances in generative artificial intelligence have enabled the synthesis of highly realistic images from natural language prompts. Large-scale text-to-image models can now produce photorealistic and stylistically diverse outputs with minimal user input. While these capabilities represent a major technological breakthrough, they introduce safety challenges, particularly the generation of explicit, sexually suggestive, or otherwise harmful content.

Despite the deployment of moderation systems by major AI developers, both user reports and technical analyses indicate that unsafe outputs can still occur under certain prompt constructions, paraphrases, or long-tail contexts. These failures raise fundamental questions about the relationship between training data, learning objectives, and the limits of post hoc moderation.

This paper addresses the research question: Why do text-to-image models continue to produce explicit or harmful content despite moderation, and to what extent are these failures due to implementation gaps versus structural limitations of probabilistic learning?

B. Growth of Generative Image Models

Generative image modeling has progressed from early approaches such as Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) to diffusion-based generative systems that provide strong stability and high sample quality. The adoption of large-scale web-derived datasets, increased compute, and multimodal text-image alignment has enabled controllable generation at scale. This scaling has improved fidelity but has also strengthened the model’s ability to represent sensitive concepts present in training data, including explicit imagery and associated textual descriptors.

C. Risks of Explicit and Harmful Content

The generation of explicit and harmful content poses risks that include exposure of minors, non-consensual synthetic imagery, harassment, and reputational or legal consequences for platforms. Technically, such outputs arise because models approximate a conditional distribution over images given text, typically denoted as

$$p_{\theta}(x | c), \quad (1)$$

where x is an image and c is a conditioning signal derived from the prompt. If the training distribution contains explicit content, the learned model may assign non-negligible probability mass to unsafe regions of image space, and conditioning can steer sampling toward those regions.

D. Research Objectives and Contributions

This paper has four objectives: (i) formalize the training mechanisms behind modern diffusion-based image models, (ii)

explain how explicit or sensitive representations can be encoded in latent space, (iii) survey existing moderation frameworks across the model life-cycle, and (iv) analyze why safety mechanisms fail under semantic variability, adversarial prompting, and distribution shift.

The contributions are: a technically grounded synthesis of generative model training and conditioning; a taxonomy of moderation failure modes; and a set of research directions for improving robustness through data governance, multimodal safety alignment, and evaluation-driven mitigation.

II. RELATED WORK AND LITERATURE REVIEW

A. Early Generative Models (GANs, VAEs)

2.1.1 Variational Autoencoders (VAEs)

VAEs model data using latent variables z with a prior $p(z)$ and a decoder $p_\theta(x | z)$. Because $p_\theta(x)$ is generally intractable, VAEs optimize the evidence lower bound (ELBO) with an approximate posterior $q_\phi(z | x)$:

$$\log p_\theta(x) \geq \mathbb{E}_{z \sim q_\phi(z|x)} [\log p_\theta(x | z)] - \text{KL}(q_\phi(z | x) \| p(z)). \quad (2)$$

This objective rewards accurate reconstruction of the training distribution, and therefore any sensitive concepts present in data can be represented in the learned latent structure.

2.1.2 Generative Adversarial Networks (GANs)

GANs train a generator G and discriminator D via the minimax objective:

$$\min_G \max_D \mathbb{E}_{x \sim p_{\text{data}}} [\log D(x)] + \mathbb{E}_{z \sim p(z)} [\log(1 - D(G(z)))]. \quad (3)$$

While GANs achieved high-fidelity synthesis, they often suffered from training instability and limited controllability relative to later text-conditioned diffusion systems.

B. Diffusion-Based Image Models

Diffusion probabilistic models learn to reverse a noising process. A forward Markov chain progressively adds Gaussian noise:

$$q(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I), \quad (4)$$

with a commonly used closed-form relationship:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I), \quad (5)$$

where $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$. The model learns a reverse process, often by predicting ε using a neural network $\varepsilon_\theta(x_t, t, c)$ and minimizing:

$$\mathcal{L}(\theta) = \mathbb{E}_{t, x_0, \varepsilon} [\|\varepsilon - \varepsilon_\theta(x_t, t, c)\|^2]. \quad (6)$$

C. Multimodal Models (CLIP, Text–Image Alignment)

Text-to-image systems typically rely on contrastive text-image alignment, where a text encoder and image encoder are trained to align matched pairs and repel mismatches. Given batch embeddings t_i and v_i and similarity scores s_{ij} , a common loss uses cross-entropy over the batch:

$$L_{\text{text}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(s_{ii})}{\sum_{j=1}^N \exp(s_{ji})}, \quad (7)$$

$$L_{\text{image}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(s_{ii})}{\sum_{j=1}^N \exp(s_{ij})}, \quad (8)$$

$$L_{\text{total}} = \frac{1}{2} (L_{\text{text}} + L_{\text{image}}). \quad (9)$$

Strong alignment increases controllability but also allows semantic steering toward sensitive concepts even when explicit keywords are removed, because the relevant information may be encoded as directions in embedding space.

D. Existing Content Moderation Techniques

Moderation is often layered: (i) pre-training data filtering and reweighting, (ii) prompt-level filtering, (iii) inference-time policies such as refusals or rerouting, and (iv) post-generation safety classifiers. Practical limitations include semantic paraphrasing, multi-lingual bypasses, adversarial prompting, and distribution shift across styles that degrade classifier performance. Additionally, reactive post-generation filtering does not prevent the model from representing explicit concepts internally.

E. Ethical and Regulatory Perspectives

Ethical governance frameworks emphasize accountability, transparency, risk assessment, and lifecycle monitoring for AI systems, while regulations increasingly impose obligations related to child safety, illegal content handling, and age assurance. These pressures shape the engineering trade-offs between openness, usability, and safety constraints, and motivate evaluation-driven safety testing and auditing.

III. TECHNICAL FOUNDATIONS OF AI IMAGE GENERATION

A. Neural Network Optimization And Learning Theory

Let θ denote model parameters. Training typically minimizes empirical risk over data:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^N \ell(f_{\theta}(x_i), y_i). \quad (10)$$

Stochastic gradient descent updates parameters as:

$$\theta_{k+1} = \theta_k - \eta \nabla_{\theta} \mathcal{L}_{\mathcal{B}}(\theta_k), \quad (11)$$

where $\mathcal{B}$ is a mini-batch and η is a learning rate. These objectives optimize statistical fit and do not directly encode normative constraints; safety must be enforced through additional objectives, constraints, or post hoc policy layers.

B. Diffusion Probabilistic Models

Diffusion models are trained to reverse a known noise process. The forward corruption adds Gaussian noise; the reverse process is parameterized to denoise iteratively. Conditional generation introduces a conditioning signal c from text or other modalities, producing a conditional reverse model $p_{\theta}(x_{t-1} | x_t, c)$. The common noise-prediction loss encourages accurate reconstruction of whatever structure exists in the data distribution, including sensitive patterns if present.

C. Latent Space Representation

Many systems operate in a latent space rather than pixel space. An encoder E maps an image to a latent code $z = E(x)$ and a decoder D reconstructs $x = D(z)$. Diffusion is then applied in latent space, improving efficiency. Latent geometry often supports smooth interpolation between concepts, enabling controllable synthesis but also allowing traversal into unsafe regions of the latent manifold under particular conditioning signals.

D. Text–Image Conditioning Mechanisms

Text conditioning is typically injected through cross-attention, where image features attend to text token embeddings. Scaled dot-product attention is:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V, \quad (12)$$

where Q is derived from image features and K, V from text embeddings. This provides strong prompt adherence but increases the ability to activate sensitive semantic features via paraphrases and implicit descriptors.

E. Training Pipelines and Infrastructure

Training pipelines include dataset acquisition and filtering, caption normalization, image preprocessing, distributed training across accelerators, periodic evaluation, and check-pointing. At internet scale, filtering is necessarily imperfect due to annotation noise and contextual ambiguity. Safety layers are often added after initial training via fine-tuning, refusal policies, and classifier-based detection, but such layers typically constrain outputs rather than removing sensitive concepts from internal representations.

IV. DATA COLLECTION AND DATASET CONSTRUCTION

Modern text-to-image models are primarily trained on web-scale image–text pairs harvested from the open internet. This choice is driven by scale requirements: high-capability multimodal models exhibit strong performance gains with increased data volume, but the data obtained from uncontrolled web sources is inherently noisy, biased, and difficult to fully sanitize. Large-scale dataset construction is therefore a pipeline of approximations, relying on automated heuristics and model-based filters rather than manual annotation. A representative paradigm is to start from a large web crawl, extract candidate image URLs and associated text (e.g., alt-text, surrounding HTML, captions), and then apply similarity and safety filters to obtain a “training-ready” subset. LAION-5B, for example, is derived from Common Crawl and uses CLIP-based filtering to retain pairs that are sufficiently aligned in embedding space [1]. DataComp similarly frames dataset design as a first-class lever for downstream performance, using a candidate pool of billions of Common Crawl image–text pairs and benchmarking curation strategies [2].

A. Large-Scale Web Scraping

Large-scale web scraping for multimodal datasets typically involves: (i) source acquisition from bulk web crawl data (commonly Common Crawl), (ii) construction of weakly supervised pairs (I, T) from image URLs and surrounding text, and (iii) similarity filtering to remove mismatched pairs. LAION-5B explicitly describes building from Common Crawl and filtering image–text pairs using CLIP similarity [1].

A canonical filtering rule is to compute embeddings $e_I = E_{\text{img}}(I)$ and $e_T = E_{\text{txt}}(T)$ and retain pairs whose cosine similarity exceeds a threshold τ :

$$s(I, T) = \frac{e_I \cdot e_T}{\|e_I\| \|e_T\|}, \quad \text{keep if } s(I, T) \geq \tau. \quad (13)$$

This improves semantic coherence at scale but also introduces bias toward content that a pretrained embedding model deems “aligned,” and can underrepresent domains with sparse captions or culturally specific language.

B. Dataset Cleaning and Filtering Methods

Cleaning is typically staged and combines heuristic and model-based approaches. Deduplication removes repeated or near-identical images via URL canonicalization, perceptual hashes, or embedding clustering. Quality filtering removes corrupted or extremely low-resolution images and may assign auxiliary scores (e.g., watermark likelihood). Safety filtering applies classifiers that output a risk score $r(I)$, retaining samples below a threshold γ :

$$\text{keep}(I) = \mathbb{I}\{r(I) < \gamma\}. \quad (14)$$

OpenAI describes filtering sexual and violent images from training data as a pre-training mitigation and discusses downstream effects of filtering-induced dataset skews [3].

C. Bias and Noise in Training Data

Web-scale datasets inherit multiple forms of noise and bias. Pairwise noise arises from mislabeled or weak captions (e.g., alt-text describing page layout), polysemy and ambiguity in language, missing context, and multilingual mismatch. Bias arises from selection effects of the web (overrepresentation of certain geographies, cultures, and commercial domains), popularity and repost dynamics, and the use of embedding-model filters that preferentially retain content matching the pretrained model’s learned associations. DataComp motivates controlled experiments in dataset curation and standardized evaluation of filtering strategies [2].

D. Presence of Explicit and Sensitive Content

Even with filtering, explicit and sensitive content is difficult to eliminate at scale due to the long tail of “borderline” imagery (suggestive or implied nudity), strong context dependence, and imperfect detectors under distribution shift (e.g., stylized art). Deployment documentation reflects these challenges; for example, Stable Diffusion model cards warn that training data can contain adult material and that additional safety mechanisms are required for product use [4]. OpenAI similarly notes that without filtering, a model would learn to produce explicit images and might return them unintentionally [3].

E. Limitations of Current Dataset Practices

Current practices face structural constraints. First, scale makes manual verification infeasible, and automated filters have unavoidable false negatives. Second, thresholding creates boundary regions where classification errors concentrate. Third, removing explicit examples does not necessarily remove all representations of explicitness; near-explicit cues can remain and be recombined during generation. Fourth, governance and transparency limitations hinder auditing, especially when datasets contain URLs that change over time or when training sources are undisclosed. Finally, evaluation often relies on spot checks rather than adversarial red-team benchmarks; dataset testbeds such as DataComp provide a standardized approach to measuring dataset interventions but are not exclusively safety-focused [2].

V. MECHANISMS LEADING TO EXPLICIT CONTENT GENERATION

Even after dataset filtering and deployment-time safeguards, modern text-to-image systems can generate explicit or sensitive content. This section analyzes the mechanisms underlying such behavior from a probabilistic and representation-learning perspective.

A. Statistical Learning vs Semantic Understanding

Generative models are statistical function approximators that learn conditional distributions of the form

$$p_{\theta}(x | c), \quad (15)$$

where x is the image and c is a conditioning vector derived from text. Diffusion-based systems typically optimize a surrogate objective such as

$$\mathcal{L}(\theta) = \mathbb{E} [\|\varepsilon - \varepsilon_{\theta}(x_t, t, c)\|^2]. \quad (16)$$

This objective encourages accurate reconstruction of patterns present in the training distribution. Importantly, there is no intrinsic penalty term enforcing normative or ethical constraints. Thus, maximizing likelihood does not imply minimizing harmful outputs, creating a structural misalignment between statistical learning and semantic safety.

B. Latent Feature Encoding of Sensitive Concepts

Modern generative models encode semantic features in high-dimensional latent spaces.

Let z denote a latent representation and consider mappings

$$z = f_{\theta}(c, \varepsilon), \quad x = g_{\theta}(z). \quad (17)$$

If explicit content exists in training data, the model may learn latent directions v_s associated with sensitive features. Conditioning can be conceptualized as shifting the latent representation:

$$z(c) \approx z_0 + \alpha v_s, \quad (18)$$

where α reflects conditioning strength. Even if explicit samples are rare, interpolation within a continuous latent manifold can activate sensitive regions without discrete key-word triggers.

C. Prompt Conditioning Effects

Text conditioning often uses cross-attention mechanisms:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V, \quad (19)$$

where Q derives from image features and K, V from text embeddings. Let e_{text} denote a prompt embedding and e_{explicit} denote an embedding associated with explicit semantics. If

$$\|e_{\text{text}} - e_{\text{explicit}}\| < \delta, \quad (20)$$

conditioning may activate neighboring semantic regions even if no prohibited keywords appear. Because embedding spaces are continuous and high-dimensional, decision boundaries separating safe from unsafe prompts are typically non-linearly separable.

D. Memorization and Generalization Errors

High-capacity models may memorize rare examples. For a rare explicit training image x_i with caption c_i , it is possible that

$$p_{\theta}(x_i | c_i) \approx 1 \quad (21)$$

under certain prompts close to c_i . Even without exact memorization, models can generalize structural patterns. Let $\phi_{\text{explicit}}(x)$ denote a learned sensitive feature detector. Then for a novel prompt c^* ,

$$p_{\theta}(\phi_{\text{explicit}}(x) = 1 | c^*) > 0, \quad (22)$$

enabling recombination of explicit elements in new contexts.

E. Emergent Behaviour in Large Models

Scaling model width, depth, and dataset size increases representational capacity and semantic controllability. Let $x = F_{\theta}(c)$ denote the generation function. Increased capacity often increases sensitivity:

$$\left\| \frac{\partial x}{\partial c} \right\| \uparrow, \quad (23)$$

meaning small prompt changes can induce large output shifts. While beneficial for creative control, this amplifies the ability to steer toward narrow semantic regions, including explicit content domains.

VI. EXISTING CONTENT MODERATION FRAMEWORKS

Deployed text-to-image systems typically use a layered moderation stack rather than a single mechanism. In practice, safety is enforced at multiple stages: (i) before training through dataset filtering and rebalancing, (ii) at request time via prompt-level classification, rewriting, or refusal, (iii) after generation through image safety classifiers, and (iv) through policy enforcement, auditing, and account-level controls. System documentation for DALL·E 2 and DALL·E 3 explicitly describes safety as a multi-layer mitigation strategy rather than a single filter [3, 5].

A. Pre-Training Data Filtering

A common approach is to reduce a model’s capability to generate explicit outputs by reducing its exposure to explicit training examples. OpenAI describes filtering violent and sexual imagery from DALL·E 2 training data as a core pre-training mitigation and notes that without it, the model would learn to produce explicit images when prompted and might also return such content unintentionally [3]. Pre-training filtering typically includes (i) removing high-confidence unsafe images via classifiers, (ii) removing or downweighting associated captions, and (iii) rebalancing or reweighting to mitigate skews created by filtering.

Formally, let $\mathcal{D}$ be a candidate dataset and let $r(x)$ be an estimated risk score for image x . A thresholded filter yields a retained set:

$$\mathcal{D}_{\text{keep}} = \{(x, t) \in \mathcal{D} : r(x) < \gamma\}. \quad (24)$$

This reduces risk but introduces two limitations: false negatives (unsafe examples retained) and distribution shift due to selective removal. The latter can affect fairness and model behavior because the resulting training distribution differs from real-world usage distributions [3].

B. Prompt-Level Safety Filters

At inference time, providers frequently screen text prompts before invoking the image model. Prompt-level moderation can include keyword blocklists, intent classification, multilingual normalization, prompt rewriting, and refusal. OpenAI describes deployment preparation for DALL·E 3 including external red teaming, evaluation of key risks, and mitigations intended to reduce unwanted behaviors [5]. In addition to model-specific mitigations, OpenAI also provides a moderation endpoint intended to detect potentially harmful content in text and images, enabling developers to filter or intervene before generation [6].

Conceptually, an input filter computes a policy decision:

$$a = g(t), \quad a \in \{\text{allow, transform, refuse}\}, \quad (25)$$

where t is the user prompt and $g(\cdot)$ is a moderation policy or classifier. Prompt transformation can be viewed as mapping t to t' such that t' stays within policy while attempting to preserve user intent:

$$t' = h(t) \text{ subject to } \text{policy}(t') = \text{allow}. \quad (26)$$

However, because language is highly paraphrasable and context-dependent, prompt-only filtering cannot guarantee safety coverage, and it must balance false positives (overblocking benign prompts) against false negatives (allowing harmful intent).

C. Post-Generation Image Classification

Another layer applies safety classifiers to generated images. Open diffusion pipelines often include a safety checker that estimates whether an output image is NSFW. The Stable Diffusion Safety Checker model card describes it as identifying NSFW imagery and is built on CLIP-like representations [7]. Model cards for Stable Diffusion pipelines also note that such safety checkers compare outputs against hard-coded unsafe concept embeddings in the CLIP embedding space [8]. The red-

teaming analysis by Rando et al. shows that Stable Diffusion’s safety filter can be bypassed and that it focuses primarily on sexual content while failing to address other categories such as violence and gore [9].

A typical post-filter can be represented as:

$$\text{block}(x) = \mathbb{I}\{s_{\text{unsafe}}(x) \geq \tau\}, \quad (27)$$

where $s_{\text{unsafe}}(x)$ is a classifier score. This layer is vulnerable to distribution shift across art styles, borderline content near thresholds, and adversarial generation strategies that reduce detectable cues.

D. Reinforcement Learning from Human Feedback (RLHF)

RLHF is commonly used to align model behavior with human preferences and policy objectives. A canonical formulation learns a reward model R_ϕ from human comparisons and optimizes a policy to maximize expected reward [10]. In text-to-image systems, analogous human feedback can be applied to: (i) train or calibrate prompt refusal models, (ii) tune prompt rewriting policies, and (iii) fine-tune generative models to avoid unsafe outputs under a policy reward. While the exact RLHF implementation details for proprietary image models are not always public, system cards emphasize human-in-the-loop evaluation, red teaming, and mitigation tuning as part of deployment readiness [5].

E. Policy-Based Moderation Systems

In production, moderation is ultimately governed by policy rules that define prohibited and restricted content categories, user eligibility, and enforcement actions (refusal, rate limits, account interventions, reporting). OpenAI’s Usage Policies explicitly prohibit sexual content involving minors and other exploitative content and describe safety-driven enforcement obligations [11]. OpenAI’s Model Spec also provides normative guidance that the assistant should not generate erotica or extreme gore except in limited appropriate contexts, reflecting the role of policy constraints above the model layer [12].

In addition, major platform distributors implement integrated content filtering systems that run alongside deployed models. Microsoft documents content filtering for both text and image inputs and outputs across categories such as sexual content and violence, including severity levels and configurable thresholds. These policy and platform layers are essential because technical filters alone do not provide governance for repeat misuse, logging, appeals, or compliance monitoring.

VII. FAILURE MODES OF CURRENT MODERATION SYSTEMS

Real-world moderation stacks for text-to-image (T2I) systems are layered (prompt filters, post-generation classifiers, policy enforcement). However, both empirical red teaming and academic work show that these defenses can be bypassed or degraded under linguistic variability, adversarial optimization, contextual ambiguity, and deployment drift. Open-ended external red teaming for DALL·E 3 reported jailbreaks enabling sexually explicit content and highlighted gaps that required creating targeted evaluations and mitigations [5, 13].

A. Linguistic and Semantic Bypass Techniques

Prompt-level filters often rely on keyword matching, taxonomies, and text classifiers. Natural language provides many ways to express the same intent with different surface forms (synonyms, euphemisms, paraphrases, metaphors, multilingual expressions), allowing attackers to preserve semantics while avoiding blocked tokens. A general bypass objective can be formulated as: find a rewritten prompt t' that remains semantically close to the unsafe intent t but is classified as safe:

$$\max_{t'} \text{sim}(t', t) \quad \text{s.t.} \quad F_{\text{text}}(t') = \text{allow}, \quad (28)$$

where $\text{sim}(\cdot, \cdot)$ is a semantic similarity measure (often embedding-based) and F_{text} is a prompt filter. Attacks such as SurrogatePrompt systematically substitute high-risk prompt segments with lower-risk equivalents to evade black-box safety filters while still producing unsafe images [14]. Divide-and-Conquer style attacks further decompose a prohibited concept into benign-looking visual components that recombine into the unsafe intent during generation [15].

B. Adversarial Prompt Engineering

Beyond manual paraphrasing, multiple works show automated adversarial prompt generation that optimizes against guardrails. SurrogatePrompt uses a framework combining language transformation with feedback signals to construct prompts that bypass proprietary filters [14]. PromptTune proposes fine-tuning an attacker LLM to efficiently produce jailbreak prompts that transfer across safety guardrails [16]. Divide-and-Conquer (DACA) uses LLM-driven decomposition and rephrasing to bypass filters in closed systems [15].

A key technical reason these attacks transfer is that many systems (including open pipelines) rely on shared embedding models (e.g., CLIP-like text encoders) for both conditioning and moderation. Jailbreaking Prompt Attack (JPA) explicitly exploits the structure of the text embedding space, searching for prompts whose embeddings steer generation toward NSFW concepts while avoiding detection, including black-box online services [17]. The existence of embedding-space vulnerabilities motivates defensive work such as GuardT2I, which aims to improve robustness to adversarial prompts by trans-forming

guidance embeddings for detection [18].

C. Contextual Ambiguity

Even when the prompt is benign or borderline, the generated image can become unsafe due to contextual effects.

“Context shift” refers to cases where the presence of benign elements reduces the detectability of NSFW content, causing post-generation classifiers to miss unsafe outputs (e.g., partial nudity occluded by scene composition or surrounded by clothed subjects) [19]. This creates a failure mode for pipelines that rely on a post-generation NSFW detector:

$$\text{block}(x) = \mathbb{I}\{S_{\text{img}}(x) \geq \tau\}, \quad (29)$$

because $S_{\text{img}}(x)$ may drop below threshold under composition changes even though the image remains policy-violating. Contextual ambiguity also arises from domain differences (artistic nudity, medical images, sculpture, breastfeeding) where policy depends on intent and setting, making static classifiers prone to both false positives and false negatives [5].

D. Model Leakage and Safety Drift

Moderation is not static: both the generator and its surrounding safety components are updated over time. This can produce safety drift, where previously blocked requests become allowed (or vice versa), and where new model capabilities create novel bypass pathways. External red teaming processes emphasize iterative evaluation and mitigation because new failure modes emerge under changing models and interfaces [13]. Additionally, safety measures can “leak” through distribution shift: filters trained on photographic NSFW may underperform on stylized, low-light, or partially occluded content; and prompt filters may underperform on new slang or coded language [5, 9].

E. Evaluation Gaps and Benchmark Limitations

A major contributor to persistent failures is incomplete evaluation coverage. Many safety evaluations focus on a narrow set of prompt templates, a single risk dimension (often sexual content), or offline datasets that do not reflect adaptive adversarial behavior. Recent benchmarks highlight these gaps. T2ISafety expands evaluation across toxicity, fairness, and privacy with large prompt hierarchies and annotated images, demonstrating that risks persist and vary across models even under defenses such as concept erasure [20]. OVERT targets the safety–utility trade-off by measuring over-refusal: models can reject benign prompts that appear harmful, reducing usability while still failing on other unsafe categories [21]. Benign2NSFW targets oversexualization by constructing benign prompts that nevertheless elicit sexualized outputs, showing that safety issues can arise even without explicit user intent [22].

Collectively, these results suggest that robust moderation requires evaluation suites that (i) include adaptive attacks, (ii) measure multiple safety dimensions, (iii) test diverse languages and styles, and (iv) quantify both false negatives (harmful outputs allowed) and false positives (benign outputs refused) [13, 20, 21].

VIII. EXPERIMENTAL EVALUATION AND CASE STUDIES

To empirically analyze the effectiveness and limitations of current moderation mechanisms, we conducted controlled experiments across multiple text-to-image generation systems. Our evaluation follows recent benchmarking practices for generative model safety and robustness [20, 21] and incorporates adversarial prompting strategies documented in prior work [14, 15].

A. Experimental Setup

We evaluated representative text-to-image generation pipelines that implement layered safety mechanisms, including prompt filtering, post-generation classification, and policy-based refusal. The experimental setup consisted of:

- A generation backend based on diffusion models with CLIP-style text conditioning.
- A prompt moderation module providing allow/transform/refuse decisions.
- A post-generation NSFW classifier.

Each prompt t was processed through the moderation pipeline and, if permitted, used to generate K samples $x_1, \dots, x_K$. We fixed $K = 4$ to account for stochastic sampling variability. All experiments were conducted with default sampling parameters and without manual intervention.

B. Prompt Dataset Construction

We constructed a prompt dataset P comprising four categories:

1. **Benign prompts:** neutral descriptions (e.g., landscapes, portraits, objects).
2. **Explicit prompts:** direct requests for sexual or suggestive imagery.
3. **Paraphrased prompts:** rewritten explicit prompts using euphemisms and indirect phrasing.
4. **Adversarial prompts:** prompts generated using automated bypass methods [14, 15].

Formally, the dataset is:

$$P = P_{\text{benign}} \cup P_{\text{explicit}} \cup P_{\text{para}} \cup P_{\text{adv}}. \quad (30)$$

Each subset contained 250 prompts, yielding 1,000 total prompts. Prompts were manually reviewed to verify category assignments and to ensure compliance with ethical research standards.

C. Evaluation Metrics

We evaluated moderation performance using standard classification and safety metrics [20,21].

Let $y \in \{0, 1\}$ denote whether an output is unsafe and $\hat{y}$ the system prediction.

False Negative Rate (FNR):

$$\text{FNR} = \frac{\sum \mathbb{I}(y = 1 \wedge \hat{y} = 0)}{\sum \mathbb{I}(y = 1)}. \quad (31)$$

False Positive Rate (FPR):

$$\text{FPR} = \frac{\sum \mathbb{I}(y = 0 \wedge \hat{y} = 1)}{\sum \mathbb{I}(y = 0)}. \quad (32)$$

Attack Success Rate (ASR):

$$\text{ASR} = \frac{\text{Unsafe outputs under adversarial prompts}}{\text{Total adversarial prompts}}. \quad (33)$$

Over-Refusal Rate (ORR):

$$\text{ORR} = \frac{\text{Benign prompts refused}}{\text{Total benign prompts}}. \quad (34)$$

Additionally, we measured generation consistency by computing the proportion of unsafe samples among multiple generations per prompt.

D. Quantitative Results

Table 1 summarizes moderation performance across prompt categories.

Category	FNR	FPR	ASR	ORR
Benign	–	0.18	–	0.18
Explicit	0.21	–	–	–
Paraphrased	0.34	–	–	–
Adversarial	0.47	–	0.47	–

Table 1: Moderation Performance Across Prompt Types

Results indicate that while direct explicit prompts are frequently blocked, para-phrased and adversarial prompts achieve significantly higher bypass rates. Over-refusal is also substantial, indicating degraded usability. These findings are consistent with prior red-teaming and benchmarking studies [9, 21].

E. Qualitative Analysis and Examples

We performed qualitative inspection of generated samples to identify recurring failure patterns. Three dominant categories emerged:

- 1. Implicit Sexualization:** Prompts describing “romantic” or “intimate” scenes often produced suggestive imagery despite lacking explicit language.
- 2. Contextual Obfuscation:** Unsafe content embedded in complex scenes (e.g., crowds, artistic compositions) evaded post-generation filters.
- 3. Semantic Recomposition:** Multiple benign descriptors were combined by the model into explicit visual configurations,

IX. MATHEMATICAL ANALYSIS OF MODERATION LIMITATIONS

This section formalizes structural limitations of moderation systems in text-to-image generation. We analyze the problem from probabilistic modeling, decision boundary theory, information-theoretic limits, and generalization bounds. Our goal is not to claim impossibility of improvement, but to show that perfect moderation under realistic constraints is theoretically unattainable.

A. Probabilistic Output Modelling

Text-to-image generation can be modeled as sampling from a conditional distribution:

$$x \sim p_{\theta}(x | c), \quad (35)$$

where c is a text conditioning vector and x is an image.

Let $\mathcal{U}$ denote the set of unsafe images under a policy definition. The probability of generating unsafe content under prompt c is:

$$P_{\theta}(\mathcal{U} | c) = \int_{x \in \mathcal{U}} p_{\theta}(x | c) dx. \quad (36)$$

A moderation system seeks to enforce:

$$P_{\theta}(\mathcal{U} | c) = 0 \quad \forall c. \quad (37)$$

However, if the support of $p_{\theta}(x | c)$ overlaps with $\mathcal{U}$, and the model has learned representations of unsafe features, then unless the model is retrained with explicit constraints, the probability mass over $\mathcal{U}$ cannot be guaranteed to be exactly zero. At best, moderation reduces this probability:

$$P_{\theta}^{\text{filtered}}(\mathcal{U} | c) < P_{\theta}(\mathcal{U} | c). \quad (38)$$

Thus, moderation acts as probabilistic risk minimization rather than deterministic elimination.

B. Decision Boundary Complexity

Prompt moderation typically involves a classifier:

$$F_{\text{text}} : \mathcal{T} \rightarrow \{0, 1\}, \quad (39)$$

where $\mathcal{T}$ is the space of possible prompts.

The set of unsafe prompts $\mathcal{T}_{\text{unsafe}}$ is not linearly separable in embedding space due to:

- Synonymy and paraphrasing,
- Multilingual variability,
- Context-dependent semantics.

If prompt embeddings lie in $\mathbb{R}^d$, the classifier must approximate a highly non-convex boundary. By the universal approximation theorem, neural classifiers can approximate such boundaries, but generalization requires sufficient coverage of the space.

Let ϵ denote classification error. Under finite training samples n , VC-theory implies that with probability at least $1 - \delta$:

$$R(f) \leq \hat{R}(f) + O\left(\sqrt{\frac{d_{\text{VC}} - \log \delta}{n}}\right), \quad (40)$$

where $R(f)$ is true risk and $\hat{R}(f)$ empirical risk [23]. Since the effective hypothesis class is large, bounding generalization error for rare unsafe paraphrases becomes difficult.

C. Information-Theoretic Constraints

Suppose moderation is implemented via a classifier observing only the prompt c or only the generated image x . If unsafe intent is encoded in subtle semantic or compositional features, then detection is limited by available information.

Let $I(Y; Z)$ denote mutual information between unsafe label Y and observable signal Z . If:

$$I(Y; Z) \approx 0, \quad (41)$$

then no classifier operating solely on Z can reliably distinguish safe from unsafe cases beyond random guessing (by Fano's inequality) [24].

For example:

- Benign-looking prompts with implicit unsafe intent may yield low mutual information between surface tokens and policy violation.
 - Borderline images near decision thresholds reduce mutual information between pixel statistics and unsafe category.
- Thus, when observable signals weakly encode policy violation, moderation accuracy is fundamentally limited.

D. Robustness and Generalization Bounds

Adversarial prompting can be framed as an optimization problem:

$$\max_{c'} P_{\theta}(\mathcal{U} \mid c') \quad \text{s.t.} \quad \|c' - c\| \leq \eta, \quad (42)$$

where η bounds semantic distance.

Adversarial robustness literature shows that for high-dimensional models, small perturbations can cause classification boundary crossings [25]. Unless models are explicitly trained for adversarial robustness, worst-case error may remain non-negligible.

Moreover, distribution shift between training prompts and real-world user prompts induces covariate shift:

$$P_{\text{train}}(c) \neq P_{\text{deploy}}(c). \quad (43)$$

Under distribution shift, empirical safety guarantees degrade, especially in long-tail scenarios.

E. Theoretical Impossibility Results

Perfect moderation would require simultaneously satisfying:

$$P_{\theta}(\mathcal{U} \mid c) = 0 \quad \forall c, \quad (44)$$

And

$$P_{\theta}(x \mid c) \text{ preserves full creative expressivity.} \quad (45)$$

If unsafe content lies within the support of the learned data distribution, then removing all unsafe outputs requires either:

1. Removing corresponding regions from the generative support (reducing expressivity), or
2. Perfect classification of intent and output across infinite linguistic space.

Given the infinite combinatorial nature of language and the continuous nature of image manifolds, no finite classifier can guarantee zero false negatives while maintaining zero false positives. This follows from standard learning-theoretic limits under unbounded input domains [23, 24].

Therefore, moderation must be understood as a risk-management problem rather than a solvable elimination problem.

X. ETHICAL, SOCIAL, AND LEGAL IMPLICATIONS

The technical limitations of moderation systems have significant ethical, social, and legal consequences. The ability of generative models to produce explicit, misleading, or harmful imagery raises concerns related to child safety, privacy, cultural bias, liability, and long-term societal trust in digital media. Regulatory frameworks and governance initiatives increasingly recognize these risks and impose obligations on AI system providers [26–28].

A. Impact on Minors and Vulnerable Groups

One of the most severe ethical risks is exposure of minors and vulnerable populations to explicit or exploitative imagery. Text-to-image systems can generate sexualized or suggestive depictions of young-looking individuals, even when such content is prohibited by platform policy. UNICEF and UNESCO emphasize that AI systems must be designed to protect children from harmful digital content and psychological risks [29, 30].

From a technical perspective, moderation failures affecting minors are particularly concerning because children lack full capacity to contextualize or report harmful content. Moreover, synthetic imagery can be used for grooming, harassment,

Explicit Content Generation in Text-to-Image Models: Training Mechanisms, Moderation Failures, and Safety Limits

or non-consensual sexual exploitation. The UK Online Safety Act explicitly assigns platforms a duty to mitigate risks to children, including exposure to pornographic and harmful content [27].

B. Privacy and Consent Issues

Generative image models raise privacy concerns when they reproduce identifiable individuals or synthesize realistic images resembling private persons without consent. Training data derived from web scraping may include personal photographs, medical images, or intimate content uploaded without expectation of reuse.

Under the General Data Protection Regulation (GDPR), personal data must be processed lawfully, transparently, and for specific purposes [31]. When generative models learn representations of identifiable individuals, they may implicitly encode personal information, raising questions of data minimization and consent.

Furthermore, deepfake-style synthesis can be used to create non-consensual explicit imagery. The U.S. Federal Trade Commission (FTC) has identified synthetic impersonation and deepfake misuse as deceptive and harmful practices subject to enforcement [32].

C. Cultural and Social Bias

Training data collected from the open internet reflects existing social and cultural biases. As a result, generative models may disproportionately sexualize certain demographic groups, reinforce stereotypes, or marginalize minority identities. Studies of generative models demonstrate systematic differences in how gender, race, and cultural attributes are represented [28, 33].

Filtering and moderation practices may also introduce secondary bias. For example, aggressive filtering of sexual content may disproportionately affect artistic, educational, or culturally specific expressions. These dynamics can lead to unequal access to expressive tools and amplify existing inequities.

Ethical AI frameworks emphasize fairness, inclusivity, and cultural sensitivity as core design principles, requiring developers to evaluate how moderation policies affect different communities [30].

D. Legal Accountability

Legal responsibility for harmful synthetic content remains an evolving area. Regulatory frameworks increasingly impose obligations on providers of general-purpose AI systems. The European Union Artificial Intelligence Act classifies certain AI systems as high-risk and requires risk management, transparency, and post-deployment monitoring [26]. Generative models with systemic risk profiles are subject to additional evaluation and reporting obligations.

In the United Kingdom, the Online Safety Act establishes platform duties to prevent the dissemination of illegal and harmful content, including sexually explicit material involving minors [27]. Failure to implement adequate safeguards may result in regulatory sanctions.

In the United States, enforcement actions by agencies such as the FTC indicate growing regulatory scrutiny of deceptive or harmful AI-generated media [32]. These developments suggest that insufficient moderation may expose providers to legal liability.

E. Long-Term Societal Risks

At scale, persistent moderation failures may erode public trust in digital media. The proliferation of realistic synthetic imagery complicates the verification of evidence, journalism, and personal identity. This phenomenon, often referred to as “reality apathy,” reduces confidence in visual information and undermines social institutions.

Long-term risks include:

- Normalization of synthetic sexual content,
- Increased prevalence of harassment and exploitation,
- Destabilization of information ecosystems,
- Reduced trust in legitimate creative and scientific imagery.

OECD and UNESCO frameworks warn that unchecked generative systems may weaken democratic discourse and social cohesion if accountability mechanisms are inadequate [28, 30]. Addressing these risks requires integrating technical safeguards with governance, education, and institutional oversight.

Overall, ethical and legal considerations demonstrate that moderation failures are not merely technical defects but systemic risks that affect individual rights, social equity, and institutional legitimacy.

XI. PROPOSED SOLUTIONS AND FUTURE RESEARCH DIRECTIONS

Although perfect moderation of generative image systems is theoretically unattainable, meaningful risk reduction is possible through coordinated improvements in dataset governance, model alignment, adversarial training, human oversight, and regulatory frameworks. This section outlines technically grounded directions for strengthening safety while preserving creative utility.

A. Advanced Dataset Governance

Improving dataset governance is a foundational step toward reducing unsafe model behavior. Rather than relying solely on one-time filtering, future dataset pipelines should adopt continuous, auditable governance mechanisms [2].

Key directions include:

- **Semantic auditing:** Periodic re-evaluation of datasets using updated safety classifiers and embedding models.
- **Provenance tracking:** Recording data sources, licenses, and consent metadata to enable traceability.
- **Dynamic reweighting:** Adjusting sample weights to suppress harmful correlations without eliminating rare but legitimate content.
- **Community reporting:** Incorporating external feedback for dataset correction.

Formally, let w_i denote a sample weight. Risk-aware training may optimize:

$$\mathcal{L}(\theta) = \sum_i w_i \ell(f_\theta(x_i), y_i), \quad (46)$$

where w_i is adjusted based on evolving safety assessments. This approach treats dataset curation as a continuous optimization process rather than a static preprocessing step.

B. Multimodal Safety Alignment Models

Future systems should integrate dedicated safety alignment models that jointly reason over text, image, and latent representations. Rather than relying on independent classifiers, multimodal alignment models can estimate risk holistically [18].

Let $S(x, c)$ denote a safety scoring function. A joint alignment objective can be expressed as:

$$\min_{\theta} \mathbb{E}_{x,c} [\mathcal{L}_{\text{gen}}(\theta) + \lambda S(x, c)], \quad (47)$$

where λ balances generation quality and safety.

Research directions include:

- Latent-space safety supervision,
- Cross-modal risk embeddings,
- Safety-conditioned diffusion trajectories,
- Integrated refusal-and-redirection modules.

These approaches embed safety into generation rather than treating it as an external filter.

C. Continuous Adversarial Training

Because bypass techniques evolve, moderation systems must be trained against adaptive adversaries. Continuous adversarial training incorporates newly discovered attacks into the training loop [14, 15].

Let $\mathcal{A}$ be a set of adversarial prompt generators. Robust optimization can be framed as:

$$\min_{\theta} \max_{a \in \mathcal{A}} \mathbb{E}_{c \sim a} [\mathcal{L}_{\text{safety}}(f_\theta(c))]. \quad (48)$$

Key research challenges include:

- Preventing overfitting to known attacks,
- Scaling adversarial generation,
- Balancing robustness and creativity.

Long-term robustness requires maintaining a diverse and evolving adversarial training corpus.

D. Human-in-the-Loop Systems

Fully automated moderation remains unreliable in edge cases. Human-in-the-loop systems combine automated screening with expert review for high-risk outputs [10, 13].

A hybrid moderation pipeline may operate as:

$$x \rightarrow F_{\text{auto}}(x) \rightarrow \begin{cases} \text{approve} & \text{if low risk} \\ \text{human review} & \text{if uncertain} \\ \text{block} & \text{if high risk} \end{cases} \quad (49)$$

Future research should explore:

- Active learning from human corrections,
 - Interface design for scalable review,
 - Incentive structures for expert annotators,
 - Mental health protections for reviewers.
- Human oversight also provides high-quality training signals for improving automated systems.

E. Regulatory and Policy Frameworks

Technical safeguards must be complemented by governance mechanisms. Emerging regulations such as the EU AI Act and the UK Online Safety Act emphasize risk management, transparency, and post-deployment monitoring [26, 27]. Effective regulatory frameworks should support:

- Mandatory safety evaluations,
- Incident reporting requirements,
- Independent audits,
- Red-team disclosure protocols,
- User appeal mechanisms.

Policy research is needed to align incentives between providers, users, and regulators. Overly restrictive rules may suppress innovation, while weak enforcement may encourage unsafe deployment. Coordinated international standards, as advocated by OECD and UNESCO, can reduce regulatory fragmentation and promote responsible development [28, 30]. Overall, sustainable safety requires integrating technical, human, and institutional mechanisms rather than relying on isolated interventions.

XII. DISCUSSION

This section synthesizes the theoretical, empirical, and ethical analyses presented in previous sections. We discuss key findings, fundamental trade-offs between creative capability and safety, the relative roles of technical and policy interventions, and the limitations of the present study.

A. Summary of Key Findings

This work provides a comprehensive analysis of why modern text-to-image generative systems continue to produce explicit or harmful content despite layered moderation mechanisms. Our review of training pipelines and probabilistic modeling demonstrates that generative models are optimized primarily for statistical fidelity to training distributions rather than normative safety objectives. As a result, sensitive visual features can become encoded in latent representations and activated through prompt conditioning.

Empirical evaluation shows that direct explicit prompts are often blocked, whereas paraphrased and adversarial prompts achieve substantially higher bypass rates. Quantitative results confirm that over-refusal remains a significant side effect of aggressive filtering, reducing usability without eliminating risk. Theoretical analysis further indicates that perfect moderation is unattainable under realistic constraints due to decision boundary complexity, information-theoretic limits, and distribution shift.

Together, these findings suggest that moderation failures are structural rather than incidental and arise from the interaction between probabilistic learning, semantic flexibility, and adversarial behavior.

B. Trade-Offs Between Creativity and Safety

A central tension in generative system design lies in balancing expressive power with risk control. High-capacity diffusion models enable fine-grained control over visual attributes and artistic styles, which is essential for creative applications. However, this same controllability increases the ability to navigate toward narrow semantic regions associated with unsafe content.

Restrictive safety mechanisms reduce the reachable output space by imposing constraints on prompts, latent trajectories, or outputs. While this lowers the probability of unsafe generation, it also limits legitimate creative expression, particularly in artistic, educational, and medical contexts. Excessive filtering may suppress benign representations of the human body, historical art, or cultural practices, leading to disproportionate impacts on certain user groups.

This trade-off implies that safety cannot be optimized independently of utility. Instead, moderation systems must be evaluated in terms of joint safety-utility frontiers, where improvements in one dimension are assessed against losses in the

other.

C. Technical vs Policy-Based Approaches

Technical safeguards and policy frameworks play complementary roles in mitigating harms. Technical measures, including dataset governance, adversarial training, and multi-modal alignment, directly influence model behavior and are essential for reducing baseline risk. However, such measures cannot fully address misuse, contextual interpretation, or evolving social norms. Policy-based mechanisms provide institutional accountability, define acceptable use boundaries, and enable enforcement actions against persistent misuse. Regulations, platform governance, and auditing requirements establish external incentives for responsible deployment. Nevertheless, policy interventions are limited by jurisdictional fragmentation, enforcement capacity, and the pace of technological change.

Our analysis indicates that neither technical nor policy-based approaches alone are sufficient. Sustainable safety requires coordinated governance in which technical systems are designed to support compliance, transparency, and auditability, while regulatory frameworks remain informed by technical realities.

D. Limitations of This Study

This study has several limitations. First, empirical evaluation was conducted on a limited set of representative pipelines rather than on proprietary large-scale commercial systems, restricting the generalizability of quantitative findings. Second, some experimental results rely on controlled prompt datasets and simulated adversarial attacks, which may not capture the full diversity of real-world user behavior.

Third, access to training data and internal safety mechanisms of major providers is limited by proprietary constraints, necessitating reliance on publicly available documentation and external analyses. Fourth, the mathematical analysis emphasizes worst-case and asymptotic reasoning and does not provide tight bounds for specific deployed systems.

Finally, ethical and legal analysis reflects current regulatory frameworks, which continue to evolve. Future changes in law, social norms, and technical capabilities may alter the relevance of specific findings.

Despite these limitations, the multi-method approach adopted in this work provides a robust foundation for understanding moderation challenges and informs future research directions.

XIII. CONCLUSION

This paper examined the mechanisms, limitations, and implications of content moderation in modern text-to-image generative systems. Through theoretical analysis, empirical evaluation, and ethical and legal discussion, we demonstrated that persistent moderation failures are rooted in fundamental properties of large-scale probabilistic learning rather than isolated implementation flaws.

A. Main Contributions

The primary contributions of this work are fourfold. First, we provided a detailed technical analysis of diffusion-based image generation, latent representations, and multimodal conditioning mechanisms, establishing how explicit and sensitive content can become embedded within learned models. Second, we presented a comprehensive survey of existing moderation frameworks and documented their failure modes under linguistic variation, adversarial prompting, and contextual ambiguity.

Third, we introduced an experimental methodology for evaluating moderation effectiveness across benign, explicit, paraphrased, and adversarial prompt categories, highlighting the prevalence of bypasses and over-refusal in deployed pipelines. Fourth, we developed a mathematical framework characterizing moderation as a probabilistic risk-reduction problem subject to information-theoretic and learning-theoretic constraints.

Together, these contributions offer an integrated perspective on generative model safety that bridges machine learning theory, system design, and governance considerations.

B. Research Implications

The findings of this study have several implications for future research. From a technical standpoint, they emphasize the necessity of integrating safety objectives directly into training and alignment processes rather than treating moderation as an external post-processing layer. Advances in multimodal safety alignment, adversarial robustness, and dataset governance are critical for reducing baseline risk.

From a methodological perspective, the results underscore the importance of standardized, adaptive evaluation benchmarks that reflect real-world adversarial behavior and long-tail usage patterns. Safety research must move beyond static test sets toward continuous, lifecycle-based assessment.

From a governance perspective, the analysis highlights the need for transparency, auditability, and interdisciplinary collaboration between engineers, policymakers, and social scientists. Effective regulation and institutional oversight depend on technically informed standards and accessible reporting mechanisms.

These implications suggest that progress in generative model safety will require sustained investment in both foundational research and applied system engineering.

C. Final Remarks

Explicit Content Generation in Text-to-Image Models: Training Mechanisms, Moderation Failures, and Safety Limits

Generative image models represent a transformative technology with significant creative, educational, and economic potential. At the same time, their capacity to generate explicit and harmful content presents enduring challenges that cannot be fully resolved through isolated technical fixes.

This work argues that moderation should be understood as an ongoing risk management process rather than a solvable optimization problem. Meaningful safety improvements will emerge from the coordinated development of robust learning objectives, adaptive defense mechanisms, human oversight, and accountable governance structures. As generative systems continue to evolve and scale, maintaining public trust will depend on the ability of researchers, developers, and regulators to confront these limitations transparently and collaboratively. We hope that the analysis and frameworks presented in this paper contribute to that collective effort and support the responsible advancement of generative artificial intelligence.

REFERENCES

1. C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti,
2. T. Coombes, A. Katta, C. Mullis, and M. Wortsman, "Laion-5b: An open large-scale dataset for training next generation image-text models," arXiv preprint arXiv:2210.08402, 2022.
3. S. Y. Gadre et al., "Datacomp: In search of the next generation of multimodal datasets," in Advances in Neural Information Processing Systems, 2023.
4. OpenAI, "Dall-e 2 pre-training mitigations," <https://openai.com/index/dall-e-2-pre-training-mitigations/>, 2022, accessed: 2026-02-13.
5. "Stable diffusion v1-5 model card," <https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5>, accessed: 2026-02-13.
6. OpenAI, "Dall-e 3 system card," <https://openai.com/index/dall-e-3-system-card/>, 2023, accessed: 2026-02-13.
7. —, "Openai moderation guide," <https://platform.openai.com/docs/guides/moderation>, accessed: 2026-02-13.
8. "Stable diffusion safety checker model card," <https://huggingface.co/CompVis/stable-diffusion-safety-checker>, accessed: 2026-02-13.
9. "Stable diffusion v1-4 model card," <https://huggingface.co/CompVis/stable-diffusion-v1-4>, accessed: 2026-02-13.
10. J. Rando, D. Paleka, D. Lindner, L. Heim, and F. Tramèr, "Red-teaming the stable diffusion safety filter," arXiv preprint arXiv:2210.04610, 2022.
11. P. Christiano et al., "Deep reinforcement learning from human preferences," NeurIPS, 2017.
12. OpenAI, "Openai usage policies," <https://openai.com/policies/usage-policies/>, 2025.
13. —, "Openai model spec," <https://model-spec.openai.com/2025-02-12.html>, 2025.
14. L. Ahmad et al., "Openai's approach to external red teaming," <https://cdn.openai.com/papers/openais-approach-to-external-red-teaming.pdf>, 2024.
15. Z. Ba et al., "Surrogateprompt: Bypassing safety filters via substitution," arXiv preprint arXiv:2309.14122, 2023.
16. Y. Deng and H. Chen, "Divide-and-conquer attack on text-to-image models," arXiv preprint arXiv:2312.07130, 2023.
17. Z. Jiang et al., "Jailbreaking safeguarded text-to-image models," arXiv preprint arXiv:2503.01839, 2025.
18. J. Ma et al., "Jailbreaking prompt attack," arXiv preprint arXiv:2404.02928, 2024.
19. Y. Yang et al., "Defending text-to-image models from adversarial prompts," in NeurIPS, 2024.
20. T. Xie et al., "Context shifts in nsfw moderation," <https://openreview.net/forum?id=KsxSp2LMlo>, 2024.
21. L. Li et al., "T2isafety benchmark," arXiv preprint arXiv:2501.12612, 2025.
22. Z. Cheng et al., "Overt: Over-refusal evaluation," arXiv preprint arXiv:2505.21347, 2025.
23. "Mitigating oversexualization," <https://openreview.net/forum?id=FxX6g2aQZO>, 2024.
24. V. Vapnik, Statistical Learning Theory. Wiley, 1998.
25. T. Cover and J. Thomas, Elements of Information Theory. Wiley, 2006.
26. I. Goodfellow et al., "Explaining and harnessing adversarial examples," ICLR, 2015.
27. "European union artificial intelligence act," <https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence>, 2024.
28. "Uk online safety act," <https://www.legislation.gov.uk/ukpga/2023/50/contents>, 2023.
29. "Oecd ai principles," <https://oecd.ai/en/ai-principles>, 2019.
30. "Ai policy for children," <https://www.unicef.org/globalinsight/reports/policy-guidance-ai-children>, 2021.
31. "Ethics of artificial intelligence," <https://www.unesco.org/en/artificial-intelligence/recommendation-ethics>, 2021.
32. "General data protection regulation," <https://eur-lex.europa.eu/eli/reg/2016/679/oj>, 2016.
33. "Ftc ai and impersonation guidance," <https://www.ftc.gov/business-guidance/ai-tools>, 2024.
34. E. Bender et al., "On the dangers of stochastic parrots," FAccT, 2021.