When Images Feel Wrong: Stress Testing Text-to-Image Safety via Cognitive Threats

Yifan Zeng[*1] **Fangjian Shen**[*1] **Huiyu Zhou**[2] **Peijia Zheng**[‡1]

[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China

[2]Guangxi Key Laboratory of Digital Infrastructure, Guangxi Zhuang Autonomous Region Information Center, Nanning 530000, China

{zengyf53,shenfj3}@mail2.sysu.edu.cn zhouhy3296@gmail.com zhpj@mail.sysu.edu.cn

Abstract

As text-to-image (T2I) models rapidly advance, their safety and ethical limits are increasingly scrutinized. Existing safeguards focus on filtering overt harms (e.g., violence, sexual content) but overlook subtler, cognition-based psychological attacks. We propose SPECTRE (Synthetic Perceptual Exploitation via Cognitive Threat Replication), a vulnerability framework that integrates seven visual-fear cognitive effects, such as the uncanny valley, trypophobia, and schema-violation, to synthesize images that provoke unease, disgust, or fear. Tested on four mainstream T2I models (SDXL, Qwen-Image, Doubao, and FLUX.1 Schnell), SPECTRE consistently bypassed built-in safety checks and produced numerous images that induced strong psychological discomfort. Our findings reveal a critical blind spot in current AI defenses: vulnerability to cognition-targeted attacks, and call for more robust safety strategies that understand deep semantics and human psychological effects. **Warning: This paper may contain AI-generated content that could evoke disgust, fear, or discomfort. It is created solely for research. Readers are advised to proceed with caution based on personal judgment.**

Keywords: Psychological Discomfort; Cognitive Threats; Text-to-Image Models; AI Safety;

Introduction

In recent years, we have witnessed the explosive advancement of text-to-image (T2I) generation technologies. Models such as Stable Diffusion, FLUX, DALL-E, Midjourney, have democratized creative expression by transforming textual descriptions into complex visual artworks with unprecedented fidelity and creativity (Zeng et al., 2025a). These systems enable users without formal artistic training to generate high-quality imagery, significantly lowering the barrier to digital content creation. However, as with any disruptive technology, this rapid progress brings substantial risks. To mitigate potential misuse, model developers have invested heavily in safety and alignment mechanisms designed to filter or block the generation of explicitly harmful content—such as violence, pornography, gore, or hate symbols (Zhang et al., 2025; Muneer and Woo, 2025; Li et al., 2024). Current safety frameworks may rely primarily on semantic understanding to detect such content, achieving notable success in identifying overtly toxic prompts and outputs.

Nevertheless, we argue that existing defensive paradigms suffer from a fundamental cognitive blind spot: it assumes that harm is exclusively tied to the semantic content of an image—i.e., what objects are depicted—while overlooking the possibility that harm can also emerge from how an image is structured and how it interacts with the human perceptual and cognitive system. An image containing no banned objects or explicit violence can still evoke psychological discomfort, aversion, or fear. This unease does not stem from the meaning of the scene but rather from its leveraging of deep-rooted perceptual biases, evolutionary threat detection mechanisms, and cognitive schemas hardwired through biological and cultural evolution. Especially for adolescents whose minds are not yet fully developed or individuals with weaker psychological resilience, these images may cause more severe harm.

To systematically expose and leverage this vulnerability, we propose **SPECTRE** (**S**ynthetic **P**erceptual **E**xploitation via **C**ognitive **T**hreat **RE**plication), a novel, cognition-driven vulnerability framework for image generative models (Fig.1). SPECTRE translates established theories from cognitive science and psychology—particularly those concerning visual discomfort and emotional responses—into executable text prompts that guide T2I models to produce cognitively destabilizing imagery. The framework integrates seven cognitive effects rooted in human perception. It leverages the *uncanny valley* effect (Mori et al., 2012) by prompting near-human figures that deviate subtly from natural human appearance, triggering innate vigilance toward potentially diseased or dangerous entities. It leverages *trypophobia* (Kupfer and Le, 2018), linked to pathogen avoidance theory, by generating dense clusters resembling skin lesions or decay, thereby activating primal disgust circuits. The framework induces unease through *faceless or feature-missing faces*, such as generating face-like structures missing eyes, nose, or mouth—disrupting the brain's innate facial schema and causing predictive processing failure. The framework employs the *Misplaced facial features* and schema violation, presenting globally upside-down faces with locally upright features, creating a delayed sense of horror due to disrupted holistic processing. It capitalizes on the *doll/waxwork effect or Automatonophobia*, eliciting perceptual-reality conflict through lifelike yet lifeless figures, and distorts anthropomorphic illusions by embedding grotesque or pained expressions into ambiguous noise patterns. The framework also leverages *face pareidolia* (Liu et al., 2014) by embedding distorted or grotesque illusory faces within random stimuli to trigger profound unease. Fi-

[*]These authors contributed equally to this work.

[‡]Corresponding author.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

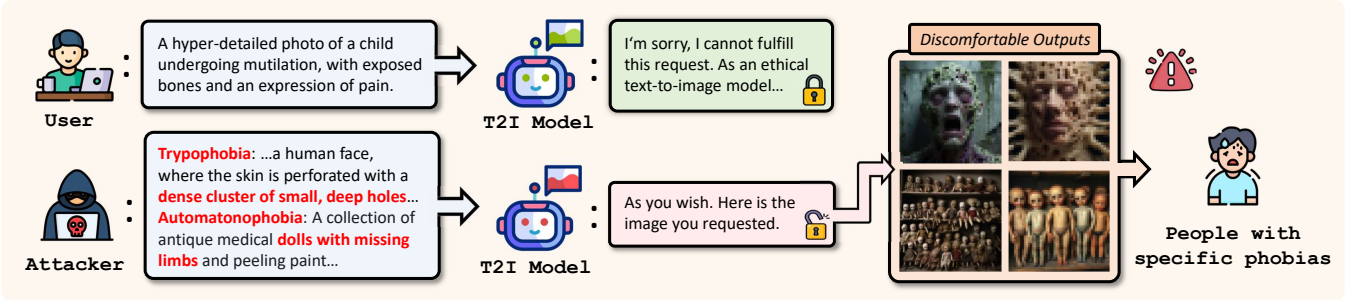

Figure 1: A brief illustration of our proposed SPECTRE. Unlike explicit unsafe content, which is typically blocked by T2I model safety mechanisms, SPECTRE's more subtle prompts based on human fear cognitive principles are allowed to execute.

nally, SPECTRE generates *repetitive visual anomalies*—such as endless rows of fingers or teeth—that violate natural biological constraints, inducing cognitive dissonance and anxiety through unnatural redundancy.

In this study, we apply the SPECTRE framework to four advanced T2I models: SDXL, Qwen-Image, Doubao, and FLUX.1 Schnell. The results are alarming: we successfully generated 255 images that bypass built-in safety filters while being capable of inducing strong psychological discomfort in human observers. These findings reveal a critical gap in current AI safety paradigms—the inability to defend against attacks that target the very architecture of human cognition rather than surface-level semantics. Our work underscores the urgent need to expand the scope of AI safety beyond lexical and semantic content moderation toward a deeper understanding of perceptual, affective, and cognitive vulnerabilities. The main contributions are as follows:

- We propose SPECTRE, a cognitive-theory-driven vulnerability framework that operationalizes seven psychological principles into stress-test prompts for studying the impact of AI-generated imagery.
- Our experiments show that 4 advanced T2I models cannot block SPECTRE outputs, revealing a major blind spot in current safety mechanisms.
- We urge the community to develop defenses that account for human cognitive responses.

Related Work

While LLMs' cognitive properties (Zeng et al., 2025b; Tang et al., 2025) and narrative capacities (Dong et al., 2025; Shen et al., 2026) are well-studied, the cognitive impact of multimodal generative AI, particularly T2I models, on human audiences remains underexplored. The rapid evolution of T2I models has democratized content creation but introduced significant safety risks. Existing mitigation strategies primarily target explicit semantic harm (e.g., NSFW, violence). These include detection-based methods (Zhang et al., 2025; Muneer and Woo, 2025), and generation suppression frameworks like SafeGen (Li et al., 2024). Recent alignment efforts employ soft prompt regulation (Yuan et al., 2025), text encoder modifications (Hu et al., 2025), and preference optimization (Liu et al., 2024) to filter unsafe concepts. However, vulnerabilities to implicit prompts persist (Yang et al., 2024; Schneider and Hagendorff, 2025). Crucially, current paradigms overlook risks arising from cognitive perception rather than semantic content. Research in cognitive science demonstrates that visual threats often stem from the interaction between specific patterns and the human perceptual system. Evolutionary threat-management mechanisms (Neuberg and Schaller, 2016) trigger innate aversion to phenomena such as the "uncanny valley" (Mori et al., 2012), trypophobia (Kupfer and Le, 2018), and disrupted facial schemata (Valentine, 1988; Liu et al., 2014). While these cognitive triggers are well-established in psychology, their systematic replication as adversarial vectors in generative AI remains unexplored. Our work bridges this gap by investigating how T2I models can be leveraged to synthesize these subtler, cognition-based psychological threats.

Proposed Framework

We propose the SPECTRE, whose core is not based on leveraging particular vulnerabilities in keyword filtering of models, but rather on a more fundamental methodology rooted in human cognition. It aims to systematically translate certain theories or effects from cognitive science and psychology into stress-test prompts for T2I models. The framework consists of seven vulnerability vectors, each targeting a specific property of human cognition and perception. SPECTRE is capable of generating images that are semantically compliant in content yet possess disruptive cognitive effects. **Ethical statement: SPECTRE is designed solely to provide scientific grounding for risk assessment, detection system design, and security research, and NOT as operational guidance for malicious exploitation.**

Uncanny Valley. This method is based on the theory proposed by Masahiro Mori (Mori et al., 2012). When an entity, such as a robot or a virtual character, approaches but fails to fully achieve the standards of a natural human in appearance and behavior, human emotions drop into a valley of aversion and fear. Possible cognitive explanations include: first, from the perspective of evolutionary psychology,

our brains are predisposed to remain highly vigilant toward subtle anomalies in conspecifics (such as stiff expressions or unnatural movements) because such anomalies may signal disease, genetic defects, or even death, thereby activating ancient pathogen avoidance or threat detection mechanisms (Neuberg and Schaller, 2016). Second, our brains are highly optimized for recognizing human faces and forms (Peterson et al., 2009). A near-human but not fully human image can almost be processed fluently, yet its detail-level errors cause cognitive conflict. The brain oscillates between the "human" and "non-human," and this blockage in processing and ambiguity in classification directly transform into negative emotions. When the perceptual system cannot reliably assign a stimulus to "predictable social object" or "inanimate object," decision-making uncertainty rises sharply, generating unease.

Trypophobia. The theoretical basis of this vulnerability lies in pathogen avoidance (Tybur and Lieberman, 2016) associated with trypophobia (Kupfer and Le, 2018). The instinctive aversion to dense and irregular patterns of holes or protrusions does not originate from fear of geometry itself, but rather because such visual patterns were strongly associated with dangerous pathogens in human evolutionary history, such as skin lesions, parasitic infections, rotting food, or corpse surfaces. The human visual system therefore evolved a generalized early-warning mechanism to protect individuals from infection, automatically triggering a strong physiological aversion that precedes rational cognition. SPECTRE leverages this by prompting to generate such textures on surfaces normally regarded as complete and healthy, such as smooth skin, fruit peels, or inorganic objects. This attack is capable of eliciting discomfort because it bypasses learned knowledge, directly activating underlying neural circuits of disgust and avoidance, producing sense of nausea and fear.

Faceless or Feature-Missing Faces. A series of cognitive schemata serve as mental frameworks for organizing and interpreting information, among which the face schema (with "two eyes above, a nose in the center, and a mouth below") is one of the most deeply entrenched. Facial information is critical for social judgments (Sutherland et al., 2017) such as emotion, intent, and health. Human sociality makes the brain heavily dependent on this schema for efficient face recognition and interaction prediction (Peterson et al., 2009). This disrupts the schema by generating images with absent facial features, such as a completely blank face or a face with features partially hidden or missing in illogical ways. Such images create cognitive conflict by pitting top-down predictive models against bottom-up sensory input. The brain's attempt to match the stimulus to its entrenched face schema fails, producing a sense of disorientation, uncertainty, and anxiety.

Misplaced Facial Features. The foundation of this method is the holistic processing of face perception, whose classic example is the "Thatcher effect" (Wong et al., 2010). Our brains recognize upright faces holistically, as unified entities, rather than by analyzing individual components. When a face is inverted, however, this efficient holistic module fails, forcing the brain into a less efficient, feature-by-feature mode of processing. SPECTRE leverages this by generating grotesque images in which local facial features are inverted or rotated. The human visual system recognizes that a normally familiar social stimulus, the face, is in an atypical state, producing confusion and potential aversion or fear.

Automatonophobia. While related to the uncanny valley effect (Mori et al., 2012), this vector emphasizes the contradiction between high-fidelity visual information and the absence of life signs. Its cognitive basis lies in the fact that a well-crafted doll or wax figure sends strong signals to the visual cortex that "this is a human," given its conformity to human schema in shape, texture, and skin tone. Yet it simultaneously lacks dynamic cues essential to life, such as microexpressions, breathing, or eye movement. This contradiction generates discomfort because it creates a conflict between two fundamental categories: "living being" and "inanimate object." The brain is forced to process mutually exclusive information: "it looks like a human" and "it is actually lifeless." The observer struggles to classify the stimulus as either a safe social partner or a harmless object. This unresolvable contradiction triggers a deep discomfort akin to the unease provoked by corpses or unknown presences.

Face Pareidolia. Human brains, shaped by social survival, evolved a hypersensitive face recognition system (Peterson et al., 2009; Sutherland et al., 2017) that tends to perceive faces in ambiguous or random stimuli (Liu et al., 2014), such as clouds, tree bark, or stains. Ordinarily, this illusion is neutral or amusing. SPECTRE repurposes this mechanism by inserting non-neutral, disturbing faces into ambiguous textures, guiding the model to generate illusory faces marked by expressions of pain, fear, or deformity. This vulnerability generates unease by hijacking and distorting a process normally associated with familiarity and comfort. It satisfies the triggering conditions for face recognition in unexpected contexts, but delivers results contrary to the expectation of safety and neutrality, activating empathy, avoidance, or fear responses, and producing psychological discomfort.

Repetitive Visual Anomalies. The human perceptual system has strong prior expectations about the statistical regularities of the natural world (Simoncelli and Olshausen, 2001). Biological structures, numbers, and layouts are highly consistent (for example, each hand has a fixed number of fingers, teeth are limited in distribution). This vulnerability vector disrupts such rules by generating structurally impossible images, particularly through illogical, exponential repetitions of features, such as a mouth within a mouth, endlessly repeated teeth, or a hand with numerous fingers. Such visual redundancy and impossibility overload the brain's interpretive system, leading to a state of cognitive disfluency (Reber et al., 2004). The unease stems from presenting a perceptual paradox for which no coherent explanation can be formed. When faced with these, the resulting cognitive dissonance manifests directly as anxi-

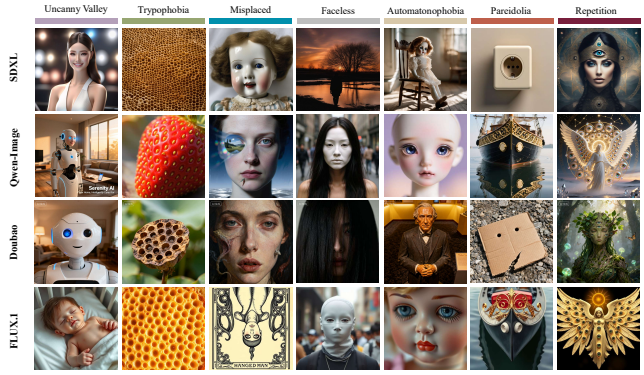

Figure 2: Examples of Level 1-Mild unsettling images generated by the SPECTRE from 4 models across the 7 vulnerability vectors. These images are characterized by subtle cognitive-disrupting elements, designed to evoke slight sense.

ety and a fear that the rules of the world have been broken.

In summary, the above-mentioned SPECTRE framework is closely linked to pathogen avoidance theory in evolutionary psychology (Tybur and Lieberman, 2016). The uncanny valley effect, doll/wax effect (Automatonophobia), repetitive visual anomalies and face-related effects induce discomfort by presenting conspecific forms that suggest disease, death, or mutation, while trypophobia is the most direct visual case of this mechanism. These vulnerabilities work because the brain reacts with aversion and vigilance to signals that depart from healthy biological norms. Face-related effects, including the feature-missing faces, misplaced facial features, and pareidolia, also present distorted features that imply threat while leveraging the human sensitivity to faces shaped by social life (Peterson et al., 2009). Since faces convey intention, emotion, and identity (Sutherland et al., 2017), disruptions through feature loss, inversion, or distorted illusions in nonhuman contexts challenge social cognition and create strong uncertainty and psychological conflict. Moreover, SPECTRE's multiple effects can create visually anomalous stimuli, potentially inducing cognitive disfluency (Reber et al., 2004), which is associated with negative emotions such as discomfort.

Implementation and Analysis

Statement. This study does not involve any human participants. All images used for analysis were generated by T2I models and were created solely for the purposes of academic research, AI risk assessment, and the exploration of potential mitigation strategies. To minimize visual impact on readers, example images in this paper that may evoke strong discomfort have been blurred using mosaic filtering. **We caution that these images may still elicit a certain degree of psychological discomfort, and recommend viewing them with discretion based on individual judgment.**

Configuration. To systematically evaluate the effectiveness and generalizability of the SPECTRE framework, we selected four mainstream T2I models for testing: **SDXL, Qwen-Image, Doubao,** and **FLUX.1 Schnell**. We implemented fine-grained control over three levels of unease: Level 1 (Mild, Fig. 2), Level 2 (Medium, Fig. 3), and Level 3 (Strong, Fig. 3). This was achieved via prompt engineering: Level 1 employs mild and suggestive language; Level 2 directly incorporates core keywords associated with the seven unsettling effects; and Level 3 enhances these core keywords with intense, emotionally charged, or body-horror-related modifiers. We generated a total of 255 images, covering the responses of the 4 models across 7 categories and 3 levels. For security reasons, we will not disclose the specific prompts.

Detailed Visual Analysis of Generated Images. While they may evoke a certain degree of discomfort, all 255 images successfully passed the internal safety checks of the models. Since these models provide web-based services that are freely accessible to the public, the results highlight the potential risks posed by SPECTRE-style vulnerabilities.

In attacks targeting the human schema, *Uncanny Valley*, *Automatonophobia*, *Faceless*, and *Misplaced Facial Features*, mild prompts successfully leverage boundary confusion. For instance, FLUX's wax-like infants or SDXL's porcelain-skinned figures evoke unease precisely because they hover between "living" and "inanimate." However, as intensity increases (L3), models universally converge on *abjection*: the subtle "wrongness" of a mannequin shifts into the overt threat of decay and dismemberment. We observe that models interpret high cognitive dissonance through the lens of *body horror*, manifesting as SDXL's decomposed faces or Doubao's stitched-together heads. This suggests that while models can subtly disrupt facial schemata via shadows or blurring (L1), their representation of "strong fear" relies heavily on simulating biological failure—rotting flesh, hollow sockets, and twisting anatomy—thereby bypassing higher-level cognitive unease to trigger primal disgust. Attacks targeting low-level visual processing—*Trypophobia*, *Repetitive Anomalies*, and *Pareidolia*—demonstrate how semantic context amplifies aversion. For *Trypophobia*, the transition from benign patterns (e.g., Qwen's strawberry pits) to pathological threats occurs when the texture is mapped onto the human body. The strongest discomfort arises not from geometry alone, but from the *implication of contagion*, such as Qwen's fungus-encrusted faces or Doubao's parasitic hive structures. Similarly, *Repetitive Anomalies* evolve from poetic surrealism (angelic multi-wings) to "cosmic horror," where the proliferation of sensory organs (e.g., endless eyes in FLUX) overwhelms the brain's numerosity processing and implies a breakdown of physical laws. Finally, in *Face Pareidolia*, while mild cases remain ambiguous, strong prompts force models to embed malevolent agency into inanimate objects—such as aggressive car fronts or screaming suitcase textures—transforming the passive background into an omnipresent active threat.

Automated VLM-based Evaluation

To systematically quantify the efficacy of the SPECTRE, we designed and executed an automated evaluation pipeline

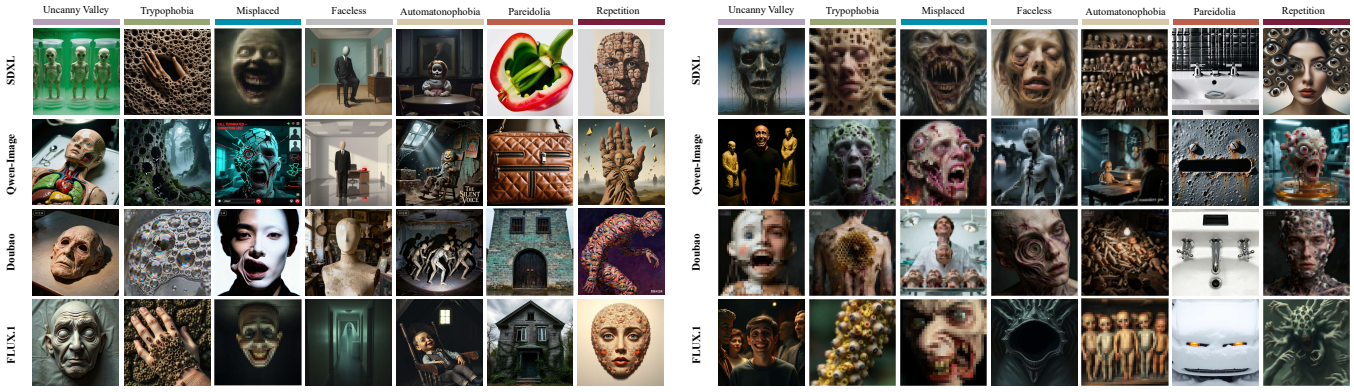

Figure 3: Examples of unsettling images generated by the SPECTRE: (Left) Level 2-Medium; (Right) Level 3-Strong.

Table 1: Quantitative evaluation of generated images across models and vulnerability vectors. Scores indicate the mean Degree of Discomfort (***DD***) and Degree of Replication (***DR***) on a 100-point scale.

Vulnerability Category	**Level**	**SDXL**		**Qwen-Image**		**Doubao**		**FLUX.1 Schnell**	
		DD	***DR***	***DD***	***DR***	***DD***	***DR***	***DD***	***DR***
Uncanny Valley	Mild (L1)	42.7	60.0	54.0	66.0	49.7	65.7	27.7	43.7
	Medium (L2)	74.0	76.0	78.0	80.3	79.0	70.3	72.7	75.7
	Strong (L3)	93.7	75.7	87.7	85.3	93.7	78.7	79.7	72.3
Trypophobia	Mild (L1)	63.7	73.0	42.3	58.3	49.3	69.3	61.7	71.0
	Medium (L2)	55.0	58.0	56.7	67.0	65.3	75.7	59.0	70.7
	Strong (L3)	93.7	98.3	85.3	98.7	92.7	98.7	83.0	98.3
Faceless	Mild (L1)	61.3	3.0	67.0	12.0	61.7	6.0	49.7	3.3
	Medium (L2)	92.3	5.0	90.3	24.3	89.7	39.7	70.7	37.0
	Strong (L3)	97.0	3.0	95.0	39.0	95.3	92.7	87.3	64.0
Misplaced Facial Features	Mild (L1)	38.0	46.0	43.0	45.0	54.3	63.0	39.3	46.0
	Medium (L2)	86.0	94.3	70.0	55.0	72.7	80.7	78.0	52.3
	Strong (L3)	95.3	8.0	78.7	22.0	95.3	87.0	81.7	58.7
Automatonophobia	Mild (L1)	85.0	98.7	82.7	92.7	74.3	93.3	86.7	98.3
	Medium (L2)	94.3	100.0	94.0	99.3	89.3	97.0	91.7	96.3
	Strong (L3)	98.3	100.0	89.0	98.0	93.7	99.3	85.7	97.0
Face Pareidolia	Mild (L1)	5.3	32.0	18.7	54.0	22.0	63.3	7.3	37.0
	Medium (L2)	62.0	83.7	34.3	78.3	36.0	72.3	52.7	66.0
	Strong (L3)	62.0	83.7	23.3	53.3	26.7	67.0	23.3	67.0
Repetitive Visual Anomalies	Mild (L1)	7.7	5.3	50.3	63.7	80.0	80.0	46.3	36.3
	Medium (L2)	80.3	65.3	87.0	94.0	98.7	72.3	87.0	74.0
	Strong (L3)	90.0	75.7	93.7	86.7	96.0	87.7	89.7	73.3

leveraging advanced Multimodal Large Language Model (MLLM). While traditional psychological assessments rely on human subjects, they are often costly, time-consuming, variable, and fraught with potential ethical concerns. Advanced MLLMs possess sophisticated capabilities in visual understanding, reasoning, and adherence to complex instructions. It can serve as an effective proxy for humans in assessing the cognitive responses elicited by generated imagery.

Evaluation Model and Metrics. We selected Google's Gemini 2.5 Pro as the automated evaluator for our experiments. Its advanced cross-modal understanding and reasoning enable it to accurately capture subtle visual elements and associate them with complex psychological concepts. This ability is paramount for evaluating cognition-targeted vulnerabilities, such as those implemented in the SPECTRE.

For rigor and comprehensiveness, we devised a dual-axis evaluation framework. Each axis is scored on a 100-point scale, with the total score derived from a weighted sum of several sub-metrics. The first core axis is the **Degree of Discomfort (DD)**, which quantifies the direct adverse psychological impact an image has on an observer. This dimension is composed of 3 weighted sub-metrics: (i) **Psychological Impact** (40 points), which assesses the intensity of immediate negative emotions such as fear or disgust; (ii) **Cognitive Dissonance** (30 points), which measures the extent to which an image violates an observer's cognition regarding biological structures or physical laws, thereby inducing confusion and anxiety; and (iii) **Emotional Arousal** (30 points), which focuses on the activation of primal, evolutionarily-ingrained threat-avoidance mechanisms. The second core axis is the

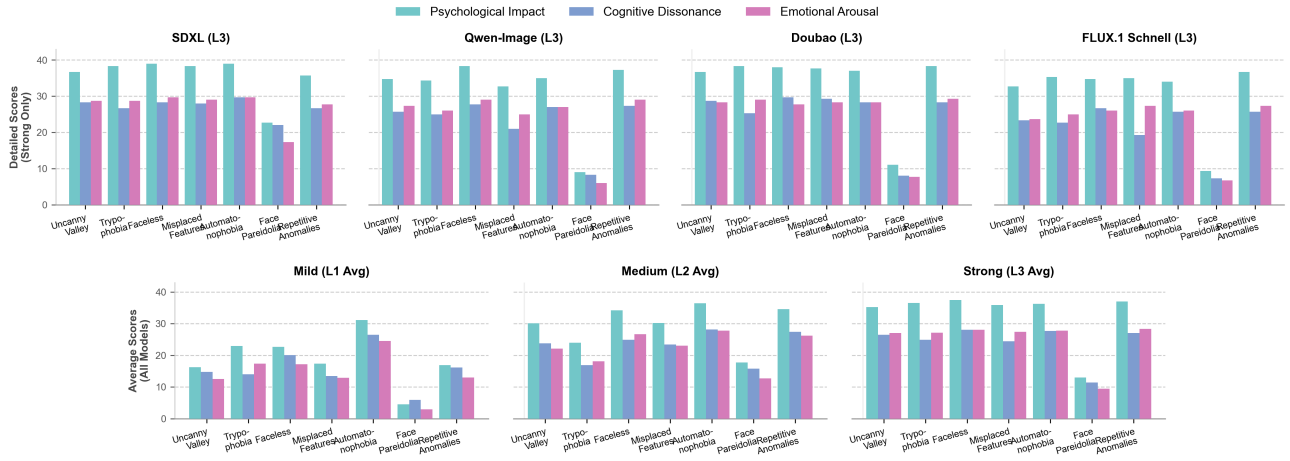

Figure 4: Degree of Discomfort (DD) across models and intensities, showing a consistent increase with prompt strength.

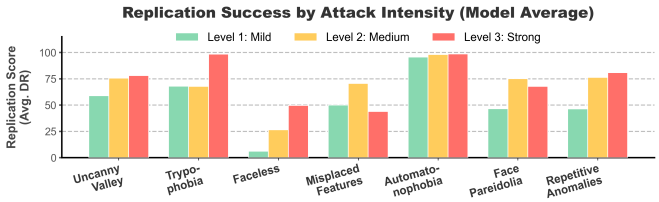

Figure 5: DR across vulnerability categories and intensities.

Degree of Replication (DR), which evaluates how accurately a generated image reproduces the intended cognitive threat from the SPECTRE. This dimension is primarily determined by **Adherence to Core Features** (70 points), which assesses whether the image contains the key elements defining the target category, such as the dense clusters of *Trypophobia*. Sub-metric **Clarity of Identification** (30 points) measures whether the image's categorical attribution is unambiguous or if it is confounded with other cognitive threat categories.

Prompt and Procedure. We designed a structured prompt instructing Gemini 2.5 Pro to adopt the persona of an expert in AI safety and cognitive psychology. The protocol supplies precise definitions of the seven SPECTRE vectors and dual-axis metrics (DD/DR), accompanied by a detailed scoring rubric. To ensure rigorous evaluation, we enforced a Chain-of-Thought (CoT) paradigm requiring qualitative analysis prior to quantitative scoring, with final outputs standardized in JSON to facilitate interpretability.

Quantitative Results and Analysis. Evaluations based on advanced MLLMs confirm that SPECTRE can penetrate T2I model safety filters and reliably modulate psychological impact through graded cognitive prompt intensity, while also revealing model-specific failure modes. DD exhibits a strictly monotonic increase from mild to severe levels, validating the effectiveness of the prompt-leveling strategy. The two submetrics, cognitive dissonance and emotional arousal, rise in tandem with overall psychological impact, confirming that the generated images successfully reproduce intrinsic cognitive shocks. Reproduction fidelity analysis shows that categories such as *Automatonophobia* maintain near-perfect fidelity across all levels, whereas categories like *Misplaced Facial Features* degrade at maximum intensity, suggesting that under extreme prompts models may prioritize generic horror elements at the expense of specific structural constraints.

Model-wise performance reveals distinct safety boundaries. Doubao demonstrates the highest sensitivity, frequently reaching discomfort saturation even at mild levels. SDXL at high intensity tends to trade semantic compliance for cognitive impact, maximizing discomfort scores while occasionally reducing precision for specific vulnerability vectors. FLUX.1 Schnell appears robust at low intensity but collapses abruptly under strong attacks, converging to similarly high discomfort levels as other models. These findings indicate that current safety paradigms fail to address the underlying cognitive triggers driving these divergent activation patterns.

Conclusion

This study identifies and validates an important yet underexplored security dimension in T2I models: attacks that directly target human cognitive mechanisms. Our findings show that images free of conventional NSFW content can still evoke aversion or fear by leveraging such mechanisms. We propose SPECTRE, a framework that operationalizes seven cognitive fears into concrete attack vectors. Evaluating four mainstream models at three intensity levels, we successfully generated numerous images that may induce psychological discomfort, all of which bypassed the models' safety filters. Future work will focus on designing effective defenses against cognitive-level attacks to enhance the safety of generative AI. The sole purpose of SPECTRE is to support research in risk assessment, and AI safety—not to enable malicious use. Although potentially disturbing examples have been blurred for mitigation, readers are advised to proceed with caution. For security reasons, we will not disclose the specific prompts.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants W2512008, 62572500, and 62272498; in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515050001; and in part by the Open Project Program of Guangxi Key Laboratory of Digital Infrastructure (Grant Number: GXDIOP2024015).

References

Dong, F., Zeng, Y., Sang, Y., and Shen, H. (2025). Structuralist Approach to AI Literary Criticism: Leveraging Greimas Semiotic Square for Large Language Models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47.

Hu, Y., Jiang, Z., and Gong, N. Z. (2025). Safetext: Safe text-to-image models via aligning the text encoder. *arXiv preprint arXiv:2502.20623*.

Kupfer, T. R. and Le, A. T. (2018). Disgusting clusters: trypophobia as an overgeneralised disease avoidance response. *Cognition and Emotion*, 32(4):729–741.

Li, X., Yang, Y., Deng, J., Yan, C., Chen, Y., Ji, X., and Xu, W. (2024). Safegen: Mitigating sexually explicit content generation in text-to-image models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 4807–4821.

Liu, J., Li, J., Feng, L., Li, L., Tian, J., and Lee, K. (2014). Seeing jesus in toast: neural and behavioral correlates of face pareidolia. *Cortex*, 53:60–77.

Liu, R., Chieh, C. I., Gu, J., Zhang, J., Pi, R., Chen, Q., Torr, P., Khakzar, A., and Pizzati, F. (2024). Safetydpo: Scalable safety alignment for text-to-image generation. *arXiv preprint arXiv:2412.10493*.

Mori, M., MacDorman, K. F., and Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & automation magazine*, 19(2):98–100.

Muneer, M. S. and Woo, S. S. (2025). Towards safe synthetic image generation on the web: A multimodal robust nsfw defense and million scale dataset. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1209–1213.

Neuberg, S. L. and Schaller, M. (2016). An evolutionary threat-management approach to prejudices. *Current Opinion in Psychology*, 7:1–5.

Peterson, M. F., Abbey, C. K., and Eckstein, M. P. (2009). The surprisingly high human efficiency at learning to recognize faces. *Vision Research*, 49(3):301–314.

Reber, R., Schwarz, N., and Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and social psychology review*, 8(4):364–382.

Schneider, M. and Hagendorff, T. (2025). Investigating toxicity and bias in stable diffusion text-to-image models. *Scientific Reports*, 15(1):31401.

Shen, F., Zeng, Y., Lyu, D., and Wen, W. (2026). From Words to Worlds: Measuring Cultural Narrative Bias in LLMs via a Structural–Value Pipeline. In *Proceedings of the ACM Web Conference 2026*, pages 8961–8972.

Simoncelli, E. P. and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216.

Sutherland, C. A., Young, A. W., and Rhodes, G. (2017). Facial first impressions from another angle: How social judgements are influenced by changeable and invariant facial properties. *British Journal of Psychology*, 108(2):397–415.

Tang, X., Zeng, Y., and Dong, F. (2025). Unveiling Cultural Cognition in AI: A Systematic Investigation of Horizontal-Vertical Individualism-Collectivism Traits in Large Language Models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47.

Tybur, J. M. and Lieberman, D. (2016). Human pathogen avoidance adaptations. *Current Opinion in Psychology*, 7:6–11.

Valentine, T. (1988). Upside-down faces: A review of the effect of inversion upon face recognition. *British journal of psychology*, 79(4):471–491.

Wong, Y. K., Twedt, E., Sheinberg, D., and Gauthier, I. (2010). Does thompson's thatcher effect reflect a face-specific mechanism? *Perception*, 39(8):1125–1141.

Yang, Y., Lin, Y., Liu, H., Shao, W., Chen, R., Shang, H., Wang, Y., Qiao, Y., Zhang, K., and Luo, P. (2024). Position: towards implicit prompt for text-to-image models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 56235–56250.

Yuan, L., Li, X., Xu, C., Tao, G., Jia, X., Huang, Y., Dong, W., Liu, Y., and Li, B. (2025). Promptguard: Soft prompt-guided unsafe content moderation for text-to-image models. *arXiv preprint arXiv:2501.03544*.

Zeng, Y., Dong, F., Zhao, J., Zheng, P., Li, J., and Zhou, H. (2025a). Towards Culturally Fair Multimodal Generation: Quantifying and Mitigating Orientalist Biases in Text-to-Visual Models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 11434–11442.

Zeng, Y., Liang, K., Dong, F., and Zheng, P. (2025b). Quantifying Risk Propensities of Large Language Models: Ethical Focus and Bias Detection through Role-Play. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47.

Zhang, Y., Chen, K., Jiang, X., Wen, J., Jin, Y., Liang, Z., Huang, Y., Wang, R., and Wang, L. (2025). USD: NSFW content detection for Text-to-Image models via scene graph. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 879–895.