A Survey of Text-to-Image Generation Models and AI Security: Adversarial Attacks, Vulnerability Assessment, and Model Robustness

Abstract

Text-to-image generation models have emerged as transformative tools in artificial intelligence, enabling the synthesis of high-quality images from textual descriptions. This survey examines the intersection of these models with AI security, focusing on adversarial attacks, vulnerability assessment, and model robustness. It highlights the significance of integrating large-scale models with language models to enhance image synthesis capabilities while addressing the security and ethical implications of their deployment. The survey explores the vulnerabilities of these models to adversarial attacks, which exploit weaknesses in neural networks, and underscores the necessity for robust defense mechanisms. Additionally, it delves into the integration of multimodal systems, emphasizing the need for effective alignment of text and image embeddings to prevent incorrect associations. The survey also covers neural network security, discussing techniques for enhancing model robustness and innovative defense mechanisms. By providing a comprehensive examination of these aspects, the survey aims to bridge the gap between the development of text-to-image models and the security challenges they face, offering insights into future research directions for ensuring the safe and ethical advancement of AI technologies.

1 Introduction

1.1 Significance of Text-to-Image Generation Models

Text-to-image generation models are revolutionizing artificial intelligence by enabling the creation of high-quality images from textual descriptions with remarkable precision [1]. Their potential spans various domains, such as art, design, and content creation, significantly enhancing the creative landscape for visual artists. The ability of these models to generate photorealistic images not only highlights their technical prowess but also positions them as transformative tools in applications requiring robust image synthesis, including face recognition and autonomous vehicles [2].

The integration of large-scale models with language models has further refined the synthesis process, allowing for more nuanced outputs. However, these advancements raise critical security and ethical concerns. The risk of generating inappropriate or harmful images through text-to-image (T2I) models necessitates comprehensive safety measures and ethical guidelines, especially as these models can produce content that contravenes usage policies. Recent improvements in T2I generation have enhanced image quality but have also exposed vulnerabilities that can be exploited to create unsafe outputs. Strategies like the Embedding Sanitizer framework aim to sanitize textual prompts, preventing harmful content generation while preserving semantic integrity. Ongoing research into the detection and attribution of fake images underscores the importance of accountability in model misuse, emphasizing the need for robust safeguards and ethical frameworks to mitigate risks associated with T2I technologies [3, 4, 5, 6]. Additionally, biases inherent in training datasets can lead to unequal

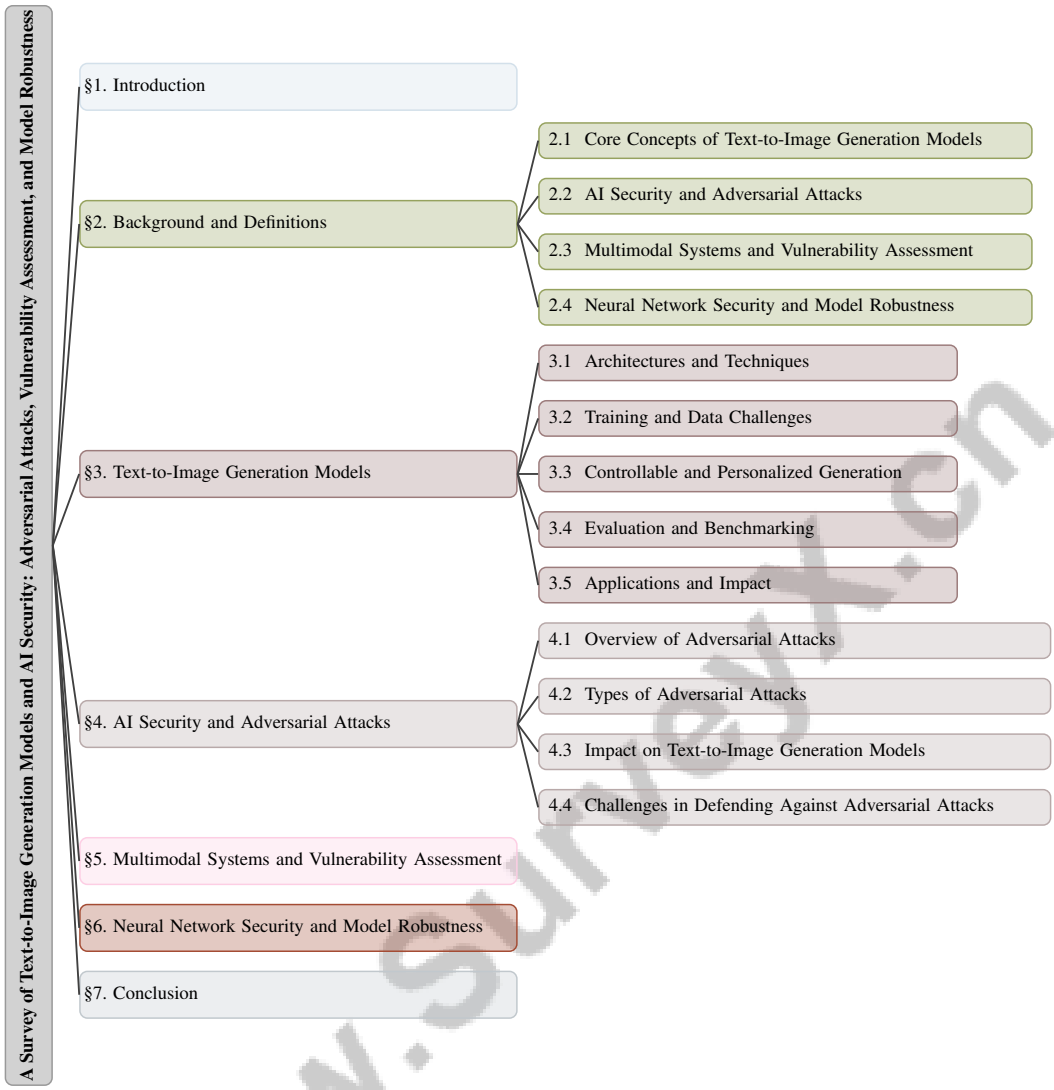


Figure 1: chapter structure

representations of marginalized groups, highlighting the necessity for careful dataset curation and bias mitigation strategies.

The practical applications of text-to-image models, such as person text-image matching, further underscore their significance in tasks involving image retrieval based on textual descriptions [7]. As these models gain traction, addressing their unique security challenges is vital for their safe and effective deployment across industries.

1.2 Security and Privacy Concerns

The deployment of text-to-image generation models introduces numerous security and privacy challenges due to their vulnerabilities to adversarial attacks and the potential for misuse in generating misleading or harmful content [8]. Adversarial attacks exploit neural network weaknesses, posing significant threats by introducing perturbations that can lead to erroneous predictions and classifications, particularly in high-stakes applications [9]. This scenario necessitates the development of robust defense mechanisms to counteract such vulnerabilities [10].

The incorporation of Large Language Models (LLMs) exacerbates these security concerns, particularly regarding data poisoning attacks and machine hallucinations, which can compromise system security and user privacy [8]. The discrete nature of text data complicates these issues, as minor

perturbations can significantly alter text semantics, challenging traditional defense strategies designed for continuous data like images [7].

Privacy risks are equally pressing, as the extensive datasets utilized for training these models often contain sensitive information susceptible to exposure through membership inference attacks. This highlights the need for comprehensive privacy assessments and benchmarks to safeguard user data [10]. The potential for adversarial attacks to manipulate generated content without the model's awareness underscores the necessity of implementing protective measures against such vulnerabilities [9].

Moreover, while denoising diffusion probabilistic models (DDPM) have made significant strides in image generation, they also introduce new security and privacy risks, particularly in facilitating transferable adversarial attacks [9]. The challenge of generating diverse, high-quality images from a single reference image, especially when incorporating novel concepts, raises additional security and privacy concerns [7].

The malicious use of AI technologies, including text-to-image models, threatens digital, physical, and political security by enabling the creation of deceptive or harmful content [8]. In sensitive domains like healthcare, the integration of machine learning techniques raises further concerns about system vulnerabilities to adversarial attacks, emphasizing the need for enhanced security measures [10]. Addressing these security and privacy issues demands rigorous research and innovative solutions to ensure the safe and ethical deployment of text-to-image models, protecting against vulnerabilities in both vision and language models [9].

1.3 Purpose and Scope of the Survey

This survey aims to provide a comprehensive examination of text-to-image generation models within the broader context of AI security, focusing on adversarial attacks, vulnerability assessment, and model robustness. It addresses the intersection of advancements in text-to-image generation technologies and the associated security challenges by synthesizing recent research findings. The impressive capabilities of contemporary models, such as diffusion models and GANs, are highlighted, alongside the vulnerabilities these models face from adversarial attacks. The survey also explores innovative approaches to model customization that enhance detail preservation without regularization, as well as trends in adversarial machine learning [11, 8, 12].

The survey covers AI-specific threats, including traditional attack surfaces and unique security pitfalls encountered in AI product development [13]. It delves into the intricacies of text-to-image diffusion models, exploring methods for text-conditioned image synthesis, text-guided creative generation, and text-guided image editing [14]. Furthermore, it emphasizes theoretical and practical advancements in controllable generation, focusing on techniques for generating images under specific conditions, multiple conditions, and universal controllable generation [15].

In the realm of AI security, five critical aspects are addressed: adversarial attacks, backdoor attacks, federated learning, uncertainty, and explainability [16]. By examining these elements, the survey seeks to provide a holistic view of the vulnerabilities and defense mechanisms pertinent to text-to-image generation models. It also explores the evolution of text-to-image generation methods driven by advancements in large models and their implications for AI security and robustness [11].

Additionally, the survey extends its scope to include practical threat models in artificial intelligence, with applications in predictive modeling and data preprocessing, particularly in sensitive domains like healthcare [17]. By integrating insights from diverse areas, the survey aims to offer a robust framework for understanding the challenges and opportunities in developing secure and resilient text-to-image generation systems. This comprehensive analysis aspires to inform future research directions and contribute to the safe and ethical advancement of AI technologies.

1.4 Structure of the Survey

This survey paper is meticulously structured to provide a comprehensive exploration of text-to-image generation models and their associated security challenges, adhering to a logical progression from foundational concepts to advanced topics. The survey begins with an **Introduction**, which sets the stage by highlighting the significance of text-to-image generation models in AI research and the

imperative of ensuring their security against adversarial attacks. This section also outlines the purpose and scope of the survey, providing a roadmap for the subsequent discussions.

The second section, **Background and Definitions**, delves into the core concepts underlying text-to-image generation models, AI security, adversarial attacks, multimodal systems, vulnerability assessment, neural network security, and model robustness. This section establishes a foundational understanding necessary for comprehending the complexities discussed in later sections.

Following this, the survey transitions into a detailed examination of **Text-to-Image Generation Models**, where various architectures and techniques are discussed. This section also addresses challenges related to training and data, methods for achieving controllable and personalized generation, and the evaluation and benchmarking processes for these models. It concludes with an exploration of the applications and societal impact of text-to-image generation models.

The fourth section, **AI Security and Adversarial Attacks**, focuses on the security challenges faced by AI systems, with a particular emphasis on adversarial attacks. This section provides an overview of adversarial attacks, describes their various types, analyzes their impact on text-to-image generation models, and discusses the challenges involved in defending against them.

In **Multimodal Systems and Vulnerability Assessment**, the survey explores the integration of multimodal systems in AI and the specific security challenges they pose. This section details the methods used for assessing vulnerabilities in these systems, offering insights into the complexities of securing multimodal AI applications.

The penultimate section, **Neural Network Security and Model Robustness**, analyzes security measures and techniques to enhance neural network robustness. It discusses various techniques for enhancing model resilience, explores strategies for improving model robustness, and describes frameworks for benchmarking and evaluating robustness. Additionally, it highlights innovative defense mechanisms developed to counter adversarial attacks.

The **Conclusion** synthesizes the principal findings of the survey on detecting multimedia generated by Large AI Models (LAIMs), highlighting the urgent need for ongoing research in AI security and robustness. It underscores the critical challenges posed by AI-generated content, including potential misuse and ethical dilemmas, and outlines future research directions that aim to address gaps in detection methodologies, enhance model resilience against adversarial attacks, and improve the overall integrity of AI systems. This comprehensive overview aims to serve as a foundational resource for researchers and practitioners committed to advancing AI security in a rapidly evolving digital landscape [18, 8, 19, 20, 21]. This structured approach ensures a comprehensive understanding of the multifaceted challenges and opportunities in the development of secure and resilient text-to-image generation systems.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Core Concepts of Text-to-Image Generation Models

Text-to-image generation models exemplify a significant advancement in AI, synthesizing images from textual inputs by integrating natural language processing and computer vision [22]. A key challenge is producing diverse and complex imagery from text prompts, requiring comprehensive world knowledge and nuanced detail incorporation [7]. Diffusion models often struggle with rendering intricate details and specific conditions, such as artistic styles or novel subjects, limiting their effectiveness [22]. Moreover, these models face challenges in creating inclusive representations across diverse attributes, including skin tones and accessories, raising ethical concerns about generating harmful or inappropriate content due to unsafe prompt embeddings.

Innovative methodologies, such as prompt engineering, enhance the performance of large language models (LLMs) and vision-language models (VLMs) by refining input prompts for more accurate outputs [22]. Developing benchmarks to differentiate between fake images generated by text-to-image models and real images is essential for evaluating authenticity and provenance. These models also support visual artists by offering frameworks for automation, exploration, and mediation, enhancing creative processes and expanding artistic expression possibilities [22]. Additionally, synthetic data generated by these models can improve classification models in data-scarce environments, illustrating their applicability beyond mere image synthesis.

2.2 AI Security and Adversarial Attacks

AI security is essential for safeguarding deep neural networks (DNNs) from adversarial attacks that exploit vulnerabilities to induce incorrect outputs. These attacks pose significant threats across domains such as natural language processing (NLP) and computer vision, where minor perturbations can lead to substantial misclassifications [23]. In text-to-image generation, adversarial attacks exploit inherent biases, severely degrading model accuracy and reliability [8]. The transferability of adversarial attacks on Vision-Language Pre-training (VLP) models complicates defense strategies, as attackers can manipulate inputs across modalities, leading to incorrect predictions [7]. Evaluating machine learning models against adversarial examples involves benchmarks designed to assess robustness, essential for understanding vulnerabilities and developing effective defenses [5].

Adversarial attacks in NLP, particularly with pre-trained language models, pose significant challenges due to subtle alterations that can lead to major misclassifications, raising ethical concerns regarding misinformation and content moderation [8]. These models' vulnerability to adversarial attacks generating non-natural samples further compromises their robustness [23]. Addressing these challenges requires a comprehensive understanding of adversarial techniques and the development of robust defense mechanisms to ensure the safe deployment of AI technologies.

2.3 Multimodal Systems and Vulnerability Assessment

The integration of multimodal systems in AI enables the synthesis and processing of information across various data types, such as text, images, and audio, leveraging individual modalities to enhance overall performance. A critical challenge is the alignment of text and image embeddings, as misalignments can lead to incorrect associations, compromising functionality and reliability [24]. Innovations like LaVi-Bridge, which use Low-Rank Adaptation (LoRA) and adapters, improve text alignment and image quality [25]. Assessing vulnerabilities in multimodal systems is crucial for identifying and mitigating risks. Research categorizes adversarial attacks and defenses in NLP based on levels such as character, word, sentence, and multi-level, as well as the adversary's knowledge, distinguishing between white-box and black-box attacks [8].

The discrete nature of graph structures complicates generating imperceptible adversarial examples and highlights the absence of a unified framework for comparing different adversarial methods. Understanding attacks like jailbreaking and prompt injection, along with vulnerabilities such as data leakage and insecure code generation, is vital. Generating adversarial images through maliciously crafted text prompts exemplifies a novel strategy for deceiving classification models, raising concerns about intellectual property theft and the integrity of prompt marketplaces [26, 8].

In large language models (LLMs), vulnerabilities are classified into model-based vulnerabilities from architecture and design; training-time vulnerabilities arising during data preparation and training; and inference-time vulnerabilities occurring during real-world deployment. This classification elucidates the multifaceted security challenges associated with LLMs, underscoring the need for targeted mitigation strategies [3, 27]. The prevalence of adversarial attacks across modalities, including image, text, and audio, necessitates a comprehensive understanding of multimodal integration challenges and robust assessment methods to enhance AI security and reliability.

2.4 Neural Network Security and Model Robustness

Neural network security is vital for protecting systems from adversarial attacks that exploit vulnerabilities, leading to incorrect outputs or degraded performance. The complexity and opacity of modern AI systems challenge the detection and mitigation of these security issues [10]. This complexity is compounded by the trade-off between model accuracy and robustness, as defenses that generalize across various attack methods are difficult to develop [9].

Model robustness ensures that neural networks maintain performance against adversarial perturbations and diverse noise forms, especially in sensitive domains where reliability is essential [28]. The architecture and capacity of neural networks influence their susceptibility to adversarial examples, prompting efforts to enhance robustness, such as leveraging attention mechanisms in image classification models [29].

The vulnerabilities of current NLP models to visual adversarial attacks highlight the necessity for improved shielding techniques to bolster robustness [5]. Innovative approaches like Defense-GAN,

which serves as a pre-processing step for classifiers without prior attack knowledge, exemplify diverse strategies to counter adversarial threats [30]. Anomaly detectors that distinguish between natural and non-natural adversarial samples provide realistic assessments of model vulnerabilities [31].

Systematic evaluation frameworks, such as the Adversarial Robustness Toolbox, are vital for assessing model robustness against various adversarial attacks and input perturbations [32]. These frameworks enhance understanding of model vulnerabilities by focusing on both deterministic models and the probabilistic nature of Bayesian networks. The use of adaptive genetic algorithms to optimize prompts without requiring gradients further illustrates the evolution of adversarial techniques [33].

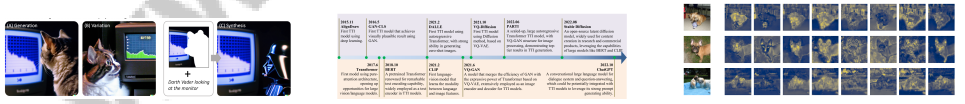
3 Text-to-Image Generation Models

The exploration of text-to-image generation models reveals significant advancements that have reshaped the artificial intelligence landscape. Understanding the foundational architectures and techniques that underpin these models is crucial for appreciating their contributions to image fidelity, control, and robustness. Table 2 provides a comprehensive comparison of the architectures and techniques employed by the DreamArtist, Parti, and LaVi-Bridge models in the realm of text-to-image generation. This section delves into these elements, setting the stage for a deeper exploration of the specific architectures and techniques emerging in this rapidly evolving field.

3.1 Architectures and Techniques

The evolution of text-to-image generation models is marked by diverse architectures and techniques aimed at enhancing image fidelity, control, and robustness. The DreamArtist model, for instance, employs a dual-embedding approach to capture salient features and rectify deficiencies, improving image controllability and diversity [1]. The Parti model, utilizing a Transformer-based image tokenizer and scaling up to 20 billion parameters, achieves state-of-the-art performance, underscoring the importance of scaling and advanced tokenization [34]. Challenges in integrating advanced language and vision models without altering original weights are addressed by LaVi-Bridge through Low-Rank Adaptation (LoRA) and adapters, ensuring seamless interoperability [12].

The ControlGAN model introduces a word-level spatial and channel-wise attention-driven generator and discriminator, enhancing adherence to input text [35]. Additionally, the ProFusion framework offers customization through an encoder network (PromptNet) and a novel sampling method (Fusion Sampling), facilitating personalized outputs [12]. Models like GLIDE and Imagen operate in pixel space, while Stable Diffusion and DALL-E 2 function in latent space, each with unique strengths and limitations [22]. Cao's taxonomy of controllable generation methods provides a structured understanding of advancements in generation with specific, multiple, and universal conditions [36].



(a) Synthesis of a Darth Vader Looking at a Monitor[37]

(b) The Evolution of Text-to-Image Generation Techniques[11]

(c) DeepLabv3+: A Comparative Study on Image Segmentation[38]

Figure 2: Examples of Architectures and Techniques

As depicted in Figure 3, advancements in text-to-image generation, driven by innovative architectures and techniques, have redefined creativity and machine learning. This figure illustrates the key innovations, frameworks, and model spaces in text-to-image generation, highlighting advancements such as DreamArtist, Parti, and LaVi-Bridge for model innovations, as well as ControlGAN and ProFusion for frameworks, while also distinguishing between pixel and latent spaces. The "Synthesis of a Darth Vader Looking at a Monitor" illustrates the stages of image synthesis, while "The Evolution of Text-to-Image Generation Techniques" chronicles pivotal milestones from 2015 to 2022. "DeepLabv3+: A Comparative Study on Image Segmentation" provides insights into algorithm efficacy in processing diverse images [37, 11, 38].

AI-generated, for reference only.

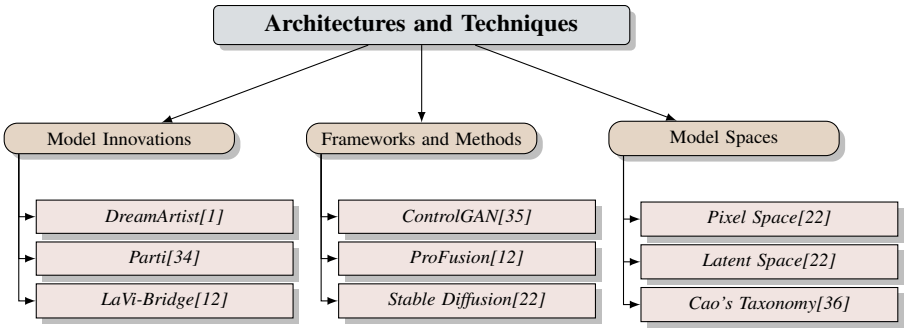


Figure 3: This figure illustrates the key innovations, frameworks, and model spaces in text-to-image generation, highlighting advancements such as DreamArtist, Parti, and LaVi-Bridge for model innovations, ControlGAN and ProFusion for frameworks, and the distinction between pixel and latent spaces.

3.2 Training and Data Challenges

Training text-to-image generation models involves challenges due to extensive data requirements and the complexity of aligning textual and visual modalities. Large datasets of high-quality image-text pairs are costly and labor-intensive [39]. Existing methods often alter multiple image attributes with single-word text changes, complicating accurate image generation [35]. Integrating new concepts can lead to catastrophic forgetting, highlighting the need for balanced training techniques [40]. Regularization techniques may cause information loss, degrading model performance [12]. Current assessment methods inadequately evaluate visual generation models, failing to adapt to evolving complexities [41].

Experiments with the Stable Diffusion model show the P+ method's effectiveness in addressing challenges, improving over traditional techniques [42]. The interaction between image and text modalities remains a hurdle, with methods failing to utilize sufficient data diversity for generating transferable adversarial examples [43]. Evaluating generative models like T5-small and GPT-2 highlights the impact of encoder-decoder and decoder-only architectures on adaptability and robustness [44]. Adversarial training methods on datasets like MNIST and CIFAR-10 illustrate efforts to enhance model resilience against adversarial attacks [45].

3.3 Controllable and Personalized Generation

Controllable and personalized generation in text-to-image models is crucial for enhancing user interaction and tailoring outputs to specific preferences. Intuitive user interfaces empower users to articulate creative intentions, enabling models to generate outputs aligned with expectations [37]. A significant challenge is accurately fine-tuning outputs in response to diverse inputs while maintaining image integrity, as current methods struggle to balance customization with fidelity [15, 35, 12]. Advanced conditioning techniques in diffusion models play a pivotal role in this process.

Customization necessitates enhancing model adaptability to diverse user needs, involving algorithms that learn from interactions and incorporate personalized features. Modular architectures facilitate targeted activation of specific components, a strategy effective in prompt engineering and advanced model architectures [38, 19, 18, 12]. The social impact of large-scale text-to-image generation models on the art community is critical, raising questions about authorship, originality, and AI's role in artistic creation. Future research should address these implications, ensuring alignment with creative community values [37].

3.4 Evaluation and Benchmarking

Evaluation and benchmarking of text-to-image generation models ensure quality and reliability. Metrics reflecting human judgment and automated assessments provide a comprehensive evaluation framework [46]. DyEval, an interactive visual assessment tool, identifies more failure cases than traditional methods, offering insights into model limitations [41]. Benchmarking involves comparing models against standardized datasets to establish baselines and identify best practices. Holistic

Benchmark	Size	Domain	Task Format	Metric
HEIM[46]	500,000	Image Generation	Text-to-Image Generation	Overall alignment, Photorealism
VLA[47]	1,000	Visual Question Answering	Adversarial Example Generation	Attack Success Rate, Adversarial Probability
ASR-Bench[48]	960,000	Speech Recognition	Adversarial Attack Evaluation	Word-Error Rate
ASR-LD[49]	162,000	Speech Recognition	Adversarial Attack	Hit Rate, WER
OARN[50]	70,000	Image Classification	Adversarial Robustness Evaluation	Error Rate, SEC
DE-FAKE[51]	328,000	Image Generation	Fake Image Detection	Accuracy, F1-score
MIA-TIG[52]	60,000	Text-to-Image Generation	Membership Inference	Accuracy
MIA-TIG[53]	60,000	Text-to-image Generation	Membership Inference	Accuracy

Table 1: This table presents a comprehensive summary of various benchmarks used in the evaluation of image generation, visual question answering, speech recognition, and image classification models. Each benchmark is characterized by its size, domain, task format, and the metrics employed to assess model performance. The table highlights the diversity of tasks and evaluation criteria pivotal for advancing model robustness and accuracy across different domains.

Evaluation of Text-to-Image Models (HEIM) assesses models across 12 dimensions, informing development strategies for robustness and accuracy [8, 46, 54, 55, 56]. Table 1 provides an overview of key benchmarks utilized in the evaluation and benchmarking of models across several domains, illustrating the diverse metrics and tasks involved in assessing model performance.

Evaluation and benchmarking are critical for advancing the field, providing understanding of model capabilities and limitations in image quality, aesthetic appeal, originality, and ethical considerations. Benchmarks like HEIM ensure models produce high-quality images and address safety, bias, and fairness issues, fostering reliable applications in diverse domains [40, 46, 12].

3.5 Applications and Impact

Text-to-image (T2I) generation models impact various domains, revolutionizing creative industries by enabling high-quality visual content generation from textual descriptions [12]. The ProFusion framework exemplifies detail preservation enhancement, critical for applications requiring high fidelity [12]. T2I models have implications in education, entertainment, and e-commerce, facilitating educational tools, enhancing storytelling, and improving product visualization [11]. As illustrated in Figure 4, the applications, challenges, and innovative approaches in the field of T2I models are highlighted, showcasing their impact across various domains, associated risks, and advancements in model development. Despite advancements, privacy vulnerabilities like membership inference attacks pose risks, emphasizing the need for privacy-preserving mechanisms [52]. Detecting multimedia-generated content addresses challenges in distinguishing authentic from synthetic media [19].

Adversarial learning in real-world applications like fraud detection reveals domain-specific challenges, necessitating continued research for system robustness [57]. Understanding and mitigating vulnerabilities in T2I models impact AI-generated content reliability [58]. Language-free training approaches, as demonstrated by L AFITE, reduce resource requirements while achieving state-of-the-art results in zero-shot settings [39]. Synthetic datasets with diverse object categories and scenes facilitate zero-shot and few-shot learning tasks, expanding T2I models' utility in data-scarce environments [59].

Feature	DreamArtist	Parti	LaVi-Bridge
Architecture	Dual-embedding	Transformer-based	Adapter-based
Optimization Technique	Feature Rectification	Scaling	Low-Rank Adaptation
Space Operation	Latent Space	Pixel Space	Latent Space

Table 2: This table presents a comparative analysis of three prominent text-to-image generation models: DreamArtist, Parti, and LaVi-Bridge. The comparison focuses on their architectural frameworks, optimization techniques, and operational spaces, highlighting the distinct approaches each model employs to enhance image generation capabilities.

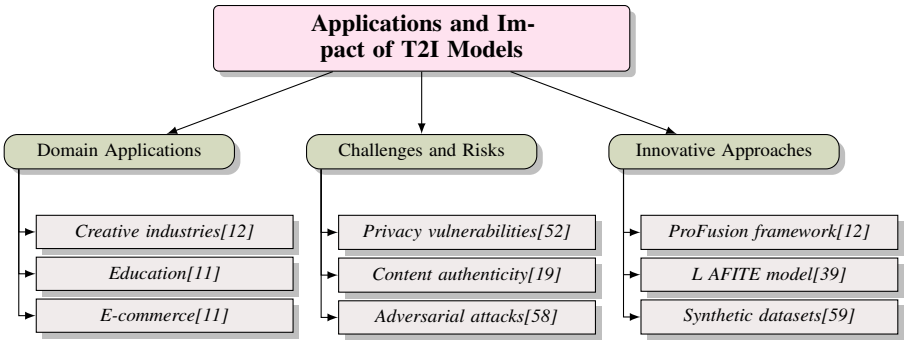


Figure 4: This figure illustrates the applications, challenges, and innovative approaches in the field of text-to-image (T2I) models, highlighting their impact across various domains, associated risks, and advancements in model development.

4 AI Security and Adversarial Attacks

The examination of adversarial attacks in AI security reveals a complex interaction between model vulnerabilities and attacker tactics. Understanding these attacks, especially in natural language processing (NLP), where they introduce subtle perturbations that mislead models, is crucial. This foundational knowledge aids in addressing NLP model vulnerabilities and utilizing tools like TextAttack for improved performance through adversarial training and data augmentation [55, 54, 8]. This section outlines the mechanisms, characteristics, and challenges adversarial attacks pose to AI systems.

4.1 Overview of Adversarial Attacks

Adversarial attacks pose significant threats by exploiting AI model vulnerabilities to produce incorrect outputs, even in high-accuracy systems. These attacks subtly modify inputs, leading to major misclassifications while maintaining data semantic integrity [60]. Deep neural networks (DNNs) in facial recognition, sentiment analysis, and neural machine translation (NMT) are particularly vulnerable to these perturbations, which can drastically affect predictions with minimal changes.

The diversity of adversarial attacks involves various strategies targeting AI models. In NLP, adversarial examples challenge models due to the discrete nature of text, where minor changes can significantly alter meaning. This complexity is heightened by NLP models' susceptibility to backdoor attacks, which manipulate predictions by inserting triggers into training data [61]. Generating effective adversarial examples that incorporate semantic information remains a critical limitation, restricting understanding of model failures in real-world conditions [23].

In vision-based systems, adversarial attacks exploit input sensitivity, complicating accountability and investigative efforts due to the difficulty in identifying responsible models in large language models (LLMs). The intricate interactions in multimodal systems combining text and vision present additional complexities inadequately addressed by current studies [36].

Defense strategies against adversarial attacks are essential, yet traditional methods often fall short. Innovative approaches are needed to develop robust defenses capable of withstanding various attack strategies [62]. Enhancing resilience against adversarial challenges requires new techniques incorporating semantic understanding and addressing existing limitations.

4.2 Types of Adversarial Attacks

Adversarial attacks on machine learning models vary based on the attacker's knowledge and perturbation nature. These attacks exploit model vulnerabilities through subtle input modifications leading to significant misclassifications without altering semantic content [60]. Key types include white-box, black-box, and adaptive attacks, each presenting unique challenges and necessitating distinct defenses.

White-box attacks assume full knowledge of the target model, including architecture and parameters, allowing precise gradient-based perturbations, making them particularly effective and challenging to defend against [63]. Black-box attacks, operating with limited information, rely on model outputs to infer gradients and generate adversarial examples, often requiring numerous queries, which can be inefficient for large-scale applications [64].

Adaptive attacks, such as Backward Pass Differentiable Approximation (BPDA), adapt to existing defenses, often bypassing them and compromising model performance in real-world scenarios [65]. These attacks highlight the limitations of current defenses, which struggle to maintain robustness against evolving adversarial techniques.

In NLP, adversarial attacks exploit language-specific dependencies, as shown in comparative analyses of Automatic Speech Recognition (ASR) systems in different languages, such as German and English [49]. The discrete nature of text data poses additional challenges, as small perturbations can lead to significant semantic changes. TextHacker exemplifies a hybrid approach that systematically estimates word importance to enhance hard-label attack effectiveness, illustrating the complexity of adversarial strategies in NLP [66].

Generating adversarial examples for mixed-type data samples requires minimal perturbations to mislead well-trained models, complicating the defense landscape [67]. Existing detection methods often fall short against concealed and hard-to-detect attacks, highlighting the need for more robust defense mechanisms [68].

The landscape of adversarial attacks is diverse and continually evolving, necessitating innovative approaches to enhance model resilience and safeguard AI systems against these sophisticated threats [69]. The reliance on manual observation methods for evaluating these attacks remains a significant challenge, emphasizing the need for automated and scalable solutions [70].

4.3 Impact on Text-to-Image Generation Models

Adversarial attacks compromise the integrity and reliability of text-to-image (T2I) generation models, which rely on accurately synthesizing images from textual descriptions. These attacks exploit vulnerabilities by introducing subtle perturbations, leading to severe misclassifications and degrading model performance, undermining effectiveness in applications requiring high precision [71]. The susceptibility of T2I models to adversarial perturbations necessitates robust security measures to prevent unsafe prompt embeddings, which can generate inappropriate or harmful content [7].

The impact of adversarial attacks on T2I models is multifaceted, affecting both visual fidelity and semantic integrity. Experiments indicate that altering all pixels in an image by small amounts can significantly reduce classifier confidence and increase misclassification rates, demonstrating model vulnerability to comprehensive perturbations [71]. Visual perturbations can lead to significant performance drops, with some models experiencing up to an 82

The development of universal perturbations for Vision-Language Pre-training (VLP) models underscores cross-modal vulnerabilities in T2I systems. These perturbations exhibit strong transferability across various datasets and tasks, enhancing understanding of model vulnerabilities and highlighting the need for advanced defense strategies [72]. Additionally, regional attributions indicate that adversarially-trained DNNs focus more on foreground perturbations, decreasing the score of the true category more than normally-trained DNNs, revealing nuanced impacts of adversarial conditions on model decision-making [62].

Despite advancements in model robustness, such as those seen in Kolmogorov-Arnold Networks (KANs), significant vulnerabilities to adversarial attacks persist, underscoring the ongoing challenge of ensuring model resilience [28]. The necessity for robust defense mechanisms is further emphasized by the potential for adversarial attacks to compromise model performance and reliability in practical applications, such as insurance risk assessments and fraud detection [7].

4.4 Challenges in Defending Against Adversarial Attacks

Defending against adversarial attacks remains challenging due to the evolving complexity and sophistication of these threats. A significant obstacle is the reliance of existing adversarial attack methods on excessive model information or their ineffectiveness against robust defenses, complicating

the evaluation of model robustness in practical applications [60]. The computational complexity involved in accurately calculating metrics like the Shapley value, which is NP-hard, exacerbates the difficulty of decomposing perturbation components, hindering comprehensive defense strategy development [62].

The specificity of datasets used in studies also limits generalizability across diverse systems, particularly in domains like facial recognition, where models are trained on varied datasets [71]. This highlights the need for defense mechanisms that are robust across different models and adaptable to various data environments. Current methods may still struggle to generate adversarial examples that are robust against all forms of defenses, necessitating further refinements to enhance effectiveness [23].

Another core challenge is the lack of benchmarks that adequately address the unique vulnerabilities of specific architectures, such as Kolmogorov-Arnold Networks (KANs), under adversarial conditions. This gap underscores the necessity for developing new benchmarks tailored to the distinct characteristics of various model architectures [28].

In recent years, the advancement of artificial intelligence (AI) has been significantly influenced by the integration of multimodal systems. These systems combine various forms of data—such as text, images, and audio—to enhance AI capabilities and improve decision-making processes. Figure 6 illustrates the primary challenges in defending against adversarial attacks, categorized into complexity and evaluation, dataset and generalizability, and benchmarks and architectures. This figure highlights the reliance on model information, computational complexity, dataset specificity, robustness across models, and the need for new benchmarks tailored to specific architectures. By examining these elements, we can better understand the complexities and implications of multimodal systems in the context of AI development.

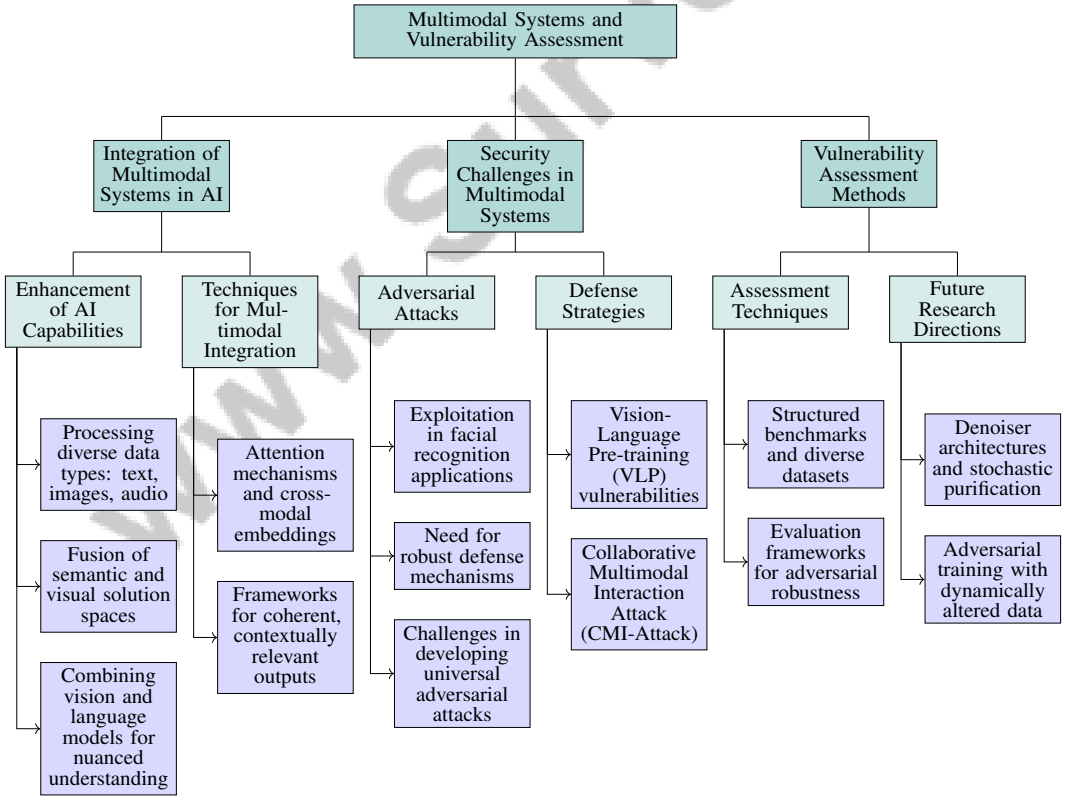


Figure 5: This figure illustrates the hierarchical structure of multimodal systems in AI, focusing on integration, security challenges, and vulnerability assessment methods. It highlights key enhancements in AI capabilities through multimodal integration, identifies security challenges posed by adversarial attacks, and outlines methods for assessing vulnerabilities and potential future research directions.

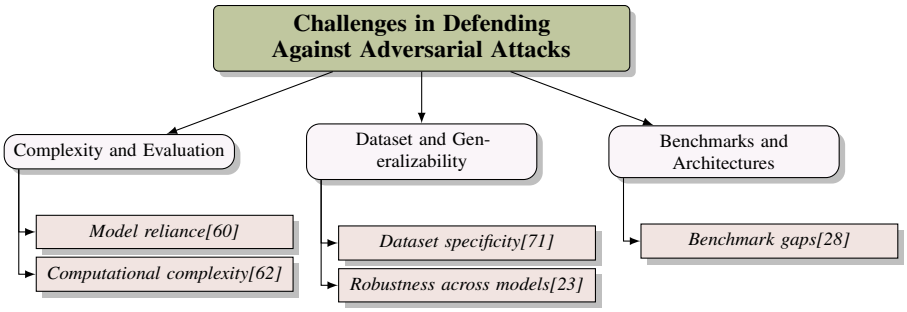


Figure 6: This figure illustrates the primary challenges in defending against adversarial attacks, categorized into complexity and evaluation, dataset and generalizability, and benchmarks and architectures. It highlights the reliance on model information, computational complexity, dataset specificity, robustness across models, and the need for new benchmarks tailored to specific architectures.

5 Multimodal Systems and Vulnerability Assessment

5.1 Integration of Multimodal Systems in AI

The integration of multimodal systems into AI represents a significant advancement, enhancing AI's ability to process diverse data types such as text, images, and audio. This integration facilitates a comprehensive understanding of information, improving AI's performance in complex tasks. A critical aspect is the fusion of semantic and visual solution spaces, essential for generating adversarial text and multimodal outputs [73].

By leveraging the complementary strengths of different modalities, multimodal systems enhance the robustness and adaptability of AI applications. For instance, combining vision and language models enables a nuanced understanding and generation of content, crucial for tasks like image captioning where visual content must be interpreted to produce relevant descriptions [40, 24, 74, 12].

Advanced techniques such as attention mechanisms and cross-modal embeddings are employed to align and interact data from distinct modalities. These techniques significantly improve AI's performance in tasks requiring a nuanced understanding, such as synthesizing text, images, and audio [24, 11, 5, 19, 72].

Frameworks supporting multimodal integration are crucial for advancing AI capabilities. By merging semantic and visual solution spaces, these frameworks enable coherent and contextually relevant outputs, addressing the limitations of unimodal approaches [73]. This integration not only enhances AI functionality but also facilitates high-fidelity multimedia content generation, improving the accuracy and robustness of AI-generated outputs [24, 11, 19].

5.2 Security Challenges in Multimodal Systems

Multimodal systems in AI face complex security challenges due to interactions among different data modalities. Adversarial attacks exploiting these interactions pose significant threats, particularly in applications like facial recognition, where minor alterations can mislead systems [71]. The ability to align any text with any image through subtle modifications highlights the need for robust defense mechanisms.

Existing benchmarks often fail to address various adversarial attack scenarios, impacting the generalizability of findings and leaving questions about robustness unanswered. This underscores the need for comprehensive evaluation frameworks that address the diverse vulnerabilities of multimodal systems. Despite shielding techniques, model performance remains lower than in clean data scenarios, indicating persistent vulnerabilities [5].

In vision-based tasks, the strengths of vision transformers contribute to safer AI applications, especially in sensitive fields like healthcare. However, variability in object detection models complicates the development of universal adversarial attacks effective across architectures. The unique vulnerabilities exposed by adversarial attacks and ensuring semantic consistency across inputs further complicate standardized security measures [3, 24, 4, 29, 55].

The rise of multimodal models necessitates innovative defense strategies against adversarial attacks, as traditional mechanisms often fall short. Vision-Language Pre-training (VLP) models are particularly vulnerable to subtle perturbations, necessitating a deeper understanding of modality interactions for effective defenses. The Collaborative Multimodal Interaction Attack (CMI-Attack) illustrates how leveraging interactions between text and image embeddings can enhance attack transferability, emphasizing the critical role of these interactions in bolstering multimodal systems' robustness [72, 75, 76]. Robust defenses are essential to maintaining the reliability and security of multimodal systems as they become more prevalent in AI applications.

5.3 Vulnerability Assessment Methods

Assessing vulnerabilities in multimodal systems is crucial for identifying risks and developing effective defense strategies against adversarial attacks. Comprehensive vulnerability assessment involves creating structured benchmarks and using diverse datasets that cover various scenarios and attack types. For instance, the Vision-Fused Attack (VFA) method demonstrates the importance of understanding visual characteristics in human text perception for effective adversarial text generation [73], highlighting the need for cross-modal interactions in vulnerability assessments.

Structured benchmarks, like those for NeuroEvolution models, offer detailed assessments of adversarial robustness, identifying strengths and limitations [10]. These benchmarks are essential for pinpointing vulnerabilities and guiding robust defense development. Evaluation frameworks assessing adversarial robustness in language models, particularly in out-of-domain contexts, provide insights into model vulnerabilities and inform comprehensive defense strategies.

Datasets covering various topics and lengths, such as those used in benchmarks for detecting word-level adversarial examples, are instrumental in evaluating model vulnerabilities across different contexts. These datasets facilitate the exploration of how adversarial attacks target specific vulnerabilities in multimodal systems, supporting precise defense strategies and enhancing model robustness in natural language processing and vision-language tasks [72, 8, 76, 55].

Domain-specific evaluations, particularly focusing on adversarial vulnerabilities in facial recognition systems, underscore the importance of tailored assessments in promoting further research and methods for evaluating vulnerabilities. Integrating diverse datasets into benchmarks for evaluating adversarial attacks in NLP emphasizes the need for thorough assessments across domains and data types, addressing challenges in comparing different adversarial attack methodologies and developing robust defenses. This approach enhances understanding of model vulnerabilities across contexts and facilitates effective mitigation strategies, contributing to the robustness and reliability of NLP systems [8, 5, 77, 29, 55].

Future research should aim to bridge gaps in defense strategies against adversarial attacks in NLP by advancing denoiser architectures, such as stochastic purification methods like MaskPure, which enhance model robustness without requiring adversarial classifier training. Exploring alternative techniques, including adversarial training with dynamically altered data and inherent stochasticity in neural networks, may provide new avenues for mitigating adversarial noise's impact on language models. These approaches seek to improve resilience while balancing performance and robustness against complex adversarial threats [76, 8, 78, 45, 29]. Integrating multimodal data in prompting, as identified in prompt engineering research, presents opportunities for enhancing multimodal systems' robustness through optimized prompt design across diverse tasks.

6 Neural Network Security and Model Robustness

6.1 Techniques for Enhancing Model Robustness

Enhancing neural network robustness against adversarial attacks is crucial for their reliability in real-world applications. Adversarial (re)training, involving exposure to adversarial examples, equips models to mitigate perturbations during inference, complemented by strategies enhancing resilience to text and image feature variations [7]. The complexity of defense is underscored by innovative attack methods like SA-Attack, which uses self-augmentation for Vision-Language Pre-training (VLP) models, generating adversarial examples across modalities [43]. Understanding decision boundaries, as exploited by Boundary Attacks, is vital for developing robust defenses [60].

Generative models, particularly GANs, contribute to robustness by producing high-quality text outputs [8]. The MAELS method, utilizing supervised generative models, creates adversarial examples preserving semantic integrity, offering a novel robustness enhancement [23]. Universal adversarial perturbations challenge VLP models across datasets, emphasizing comprehensive defense strategies [72].

6.2 Robustness Enhancement Strategies

Ensuring the security and reliability of AI systems requires robust strategies against adversarial attacks. Generative adversarial networks (GANs), like Rob-GAN, enhance robustness and training efficiency compared to traditional methods [79]. Attention mechanisms, though variable in impact across datasets, offer potential robustness improvements [62]. Data augmentation, exemplified by the SAT method, provides effective randomization defenses with minimal computational cost.

Maintaining accuracy amidst adversarial attacks without extensive retraining is critical, especially in resource-constrained applications [28]. Research on over-parameterized models reveals trade-offs between complexity and vulnerability, while techniques like NADAR optimize architecture for enhanced robustness and accuracy. Architectural shifts, such as from ResNet to ConvNeXt, improve robustness significantly under specific attacks [80, 81, 82, 38].

Pravin et al. explore data augmentation for authorship verification systems, though results vary across datasets [83, 8, 12]. Enhancing vulnerable neurons can improve adversarial context performance, maintaining stable predictions despite input perturbations. Global adversarial methods explore broader input spaces, revealing vulnerabilities local methods may miss, enhancing model robustness and generalization [84, 8].

6.3 Benchmarking and Evaluation Frameworks

Robust benchmarking and evaluation frameworks are vital for assessing neural network resilience against adversarial attacks, guiding robustness enhancements in NLP systems [85, 8, 54, 55]. Using pre-trained models, like Inception V3 on ImageNet, to evaluate defense methods, underscores the importance of established models for reliable assessments [65]. Experiments with ResNet architectures reveal varying robustness, informing future resilient architecture development [86].

Comparing pre-trained models' performance, such as VGG-16, before and after defense mechanisms, highlights the need for real-world scenario evaluations [69]. Benchmarking frameworks advance AI security by identifying NLP model vulnerabilities and supporting robust defense strategies, crucial for applications like rumor detection and sentiment analysis [87, 55]. Future research should refine these frameworks, incorporating new metrics to address evolving adversarial threats.

6.4 Innovative Defense Mechanisms

Innovative defense mechanisms are essential for enhancing AI system resilience against adversarial attacks. The Synthetic Adversarial Data Generation Pipeline produces diverse examples, preparing models for real-world adversarial scenarios [36]. Random perturbations during testing introduce variability, countering adversarial predictions effectively [88]. Anomaly detectors enhance detection capabilities against white-box attacks by distinguishing non-natural adversarial samples [79].

In multi-modal systems, frameworks providing provable robustness guarantees are crucial, utilizing sub-sampling strategies tailored to modality characteristics [61]. Defense-GAN exemplifies generative adversarial networks' utility in protecting classifiers without structural modifications, generalizing across attack types [89]. Adversarial training techniques for reinforcement learning indicate promising robustness enhancements [43].

Hybrid approaches leveraging unsupervised and supervised learning improve performance over traditional methods, fortifying AI systems against adversarial threats [60]. Introducing adversarial examples challenges classifiers to learn robust distinguishing features, refining adversarial training methodologies [71].

7 Conclusion

7.1 Future Directions in AI Security

Advancements in AI security and robustness are poised to evolve significantly through several promising research trajectories. Key areas of focus include refining robustness metrics to address a broader spectrum of adversarial attacks and enhancing model training to mitigate issues like neighborhood under-fitting. Future research should prioritize the development of sophisticated models that integrate visual understanding to bolster robustness against adversarial perturbations.

In the realm of adversarial strategy development, improving the efficiency of perturbation generation and evaluating the robustness of methods against diverse defense mechanisms are essential. This endeavor encompasses refining gradient estimators to strengthen defenses against adversarial attacks and exploring their applicability to a range of transformations. Techniques such as Bayesian modeling of weights also offer potential for enhancing neural network robustness against adversarial examples.

Expanding the scalability of Model Tuning with Prompts (MVP) to larger models and its applicability across a wider array of NLP tasks is vital. Developing advanced evaluation strategies for robustness in varied contexts will further fortify AI security. Additionally, enhancing model efficiency and exploring hybrid architectures that capitalize on the strengths of different generative models can significantly advance text-to-image (T2I) models within broader AI-generated content applications. Establishing standardized frameworks for prompt engineering and integrating prompts across multiple modalities should be prioritized in future research efforts.

Refining prompting techniques and incorporating additional languages are crucial for robust content moderation strategies, particularly in image generation systems. The integration of diverse attack methods and datasets, alongside sophisticated detection techniques, will be imperative to adapt to evolving adversarial strategies. Strengthening the robustness of detection methods against defenses and extending their applicability to other text-to-image generation models could significantly inform future research in AI security.

Further investigations may focus on reducing optimization times and improving the accuracy of generated images, especially for tasks requiring precise shapes. Developing phrase-based adversarial examples and exploring human strategies in generating effective adversarial inputs can provide valuable insights into enhancing AI robustness. Additionally, improving the robustness of data attribution methods and exploring further adversarial strategies remain critical for advancing AI security.

Collectively, these research directions aim to fortify AI systems against evolving adversarial threats, ensuring their safe and reliable deployment across diverse applications. Exploring architectural modifications and integrating additional adversarial attack types could further enhance robustness. Addressing the dynamics of multi-agent debates over extended interactions and among models from different providers could also significantly contribute to enhancing adversarial robustness.

References

[1] Ziyi Dong, Pengxu Wei, and Liang Lin. Dreamartist++: Controllable one-shot text-to-image generation via positive-negative adapter, 2025.

[2] Anurag Arnab, Ondrej Miksik, and Philip H. S. Torr. On the robustness of semantic segmentation models to adversarial attacks, 2018.

[3] Sara Abdali, Richard Anarfi, CJ Barberan, and Jia He. Securing large language models: Threats, vulnerabilities and responsible practices, 2024.

[4] Huming Qiu, Guanxu Chen, Mi Zhang, and Min Yang. Safe text-to-image generation: Simply sanitize the prompt embedding, 2024.

[5] Steffen Eger, Gözde Gül Şahin, Andreas Rücklé, Ji-Ung Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, Edwin Simpson, and Iryna Gurevych. Text processing like humans do: Visually attacking and shielding nlp systems, 2020.

[6] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models, 2023.

[7] Fan Li, Hang Zhou, Huafeng Li, Yafei Zhang, and Zhengtao Yu. Person text-image matching via text-feature interpretability embedding and external attack node implantation, 2022.

[8] Izzat Alsmadi. Adversarial machine learning in text analysis and generation, 2021.

[9] Samuel Henrique Silva and Peyman Najafirad. Opportunities and challenges in deep learning adversarial robustness: A survey, 2020.

[10] B Kereopa-Yorke. Quantifying ai vulnerabilities: A synthesis of complexity, dynamical systems, and game theory, 2024.

[11] Fengxiang Bie, Yibo Yang, Zhongzhu Zhou, Adam Ghanem, Minjia Zhang, Zhewei Yao, Xiaoxia Wu, Connor Holmes, Pareesa Golnari, David A Clifton, et al. Renaissance: A survey into ai text-to-image generation in the era of large model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[12] Yufan Zhou, Ruiyi Zhang, Tong Sun, and Jinhui Xu. Enhancing detail preservation for customized text-to-image generation: A regularization-free approach, 2023.

[13] Ebenezer R. H. P. Isaac and Jim Reno. Ai product security: A primer for developers, 2023.

[14] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion models in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023.

[15] Pu Cao, Feng Zhou, Qing Song, and Lu Yang. Controllable generation with text-to-image diffusion models: A survey. *arXiv preprint arXiv:2403.04279*, 2024.

[16] Yonghao Xu, Tao Bai, Weikang Yu, Shizhen Chang, Peter M. Atkinson, and Pedram Ghamisi. Ai security for geoscience and remote sensing: Challenges and future trends, 2023.

[17] Kathrin Grosse, Lukas Bieringer, Tarek Richard Besold, and Alexandre Alahi. Towards more practical threat models in artificial intelligence security, 2024.

[18] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. Unleashing the potential of prompt engineering in large language models: a comprehensive review, 2024.

[19] Li Lin, Neeraj Gupta, Yue Zhang, Hainan Ren, Chun-Hao Liu, Feng Ding, Xin Wang, Xin Li, Luisa Verdoliva, and Shu Hu. Detecting multimedia generated by large ai models: A survey, 2024.

[20] Yitong Li, Trevor Cohn, and Timothy Baldwin. Learning robust representations of text, 2016.

[21] Banghua Zhu, Norman Mu, Jiantao Jiao, and David Wagner. Generative ai security: Challenges and countermeasures, 2024.

[22] Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. Large-scale text-to-image generation models for visual artists' creative works. In *Proceedings of the 28th international conference on intelligent user interfaces*, pages 919–933, 2023.

[23] Shuai Li, Xiaoyu Jiang, and Xiaoguang Ma. Transcending adversarial perturbations: Manifold-aided adversarial examples with legitimate semantics, 2024.

[24] Shaeke Salman, Md Montasir Bin Shams, and Xiuwen Liu. Unaligning everything: Or aligning any text to any image in multimodal models, 2024.

[25] Shihao Zhao, Shaozhe Hao, Bojia Zi, Huaizhe Xu, and Kwan-Yee K Wong. Bridging different language models and generative vision models for text-to-image generation. In *European Conference on Computer Vision*, pages 70–86. Springer, 2024.

[26] Xinyue Shen, Yiting Qu, Michael Backes, and Yang Zhang. Prompt stealing attacks against text-to-image generation models, 2024.

[27] Atmane Ayoub Mansour Bahar and Ahmad Samer Wazan. On the validity of traditional vulnerability scoring systems for adversarial attacks against llms, 2024.

[28] Tal Alter, Raz Lapid, and Moshe Sipper. On the robustness of kolmogorov-arnold networks: An adversarial perspective, 2025.

[29] Álvaro Huertas-García, Alejandro Martín, Javier Huertas-Tato, and David Camacho. Camouflage is all you need: Evaluating and enhancing language model robustness against camouflage adversarial attacks, 2024.

[30] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models, 2018.

[31] Jiayi Wang, Rongzhou Bao, Zhuosheng Zhang, and Hai Zhao. Distinguishing non-natural from natural adversarial samples for more robust pre-trained language model, 2022.

[32] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian M. Molloy, and Ben Edwards. Adversarial robustness toolbox v1.0.0, 2019.

[33] Xiaopei Zhu, Peiyang Xu, Guanning Zeng, Yingpeng Dong, and Xiaolin Hu. Natural language induced adversarial images, 2024.

[34] Jiliang Zhang and Chen Li. Adversarial examples: Opportunities and challenges, 2019.

[35] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. *Advances in neural information processing systems*, 32, 2019.

[36] Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Yu Cheng, and Wei Hu. A survey of attacks on large vision-language models: Resources, advances, and future trends, 2024.

[37] Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. Large-scale text-to-image generation models for visual artists' creative works, 2023.

[38] Francesco Croce and Matthias Hein. On the interplay of adversarial robustness and architecture components: patches, convolution and attention, 2022.

[39] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17907–17917, 2022.

[40] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[41] Xiaoyue Mi, Fan Tang, Juan Cao, Qiang Sheng, Ziyao Huang, Peng Li, Yang Liu, and Tong-Yee Lee. Interactive visual assessment for text-to-image generation models, 2024.

[42] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. p+: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023.

[43] Bangyan He, Xiaojun Jia, Siyuan Liang, Tianrui Lou, Yang Liu, and Xiaochun Cao. Sa-attack: Improving adversarial transferability of vision-language pre-training models via self-augmentation, 2023.

[44] Shuli Jiang, Swanand Ravindra Kadhe, Yi Zhou, Ling Cai, and Nathalie Baracaldo. Forcing generative models to degenerate ones: The power of data poisoning attacks, 2023.

[45] Ayoub Arous, Andres F Lopez-Lopera, Nael Abu-Ghazaleh, and Ihsen Alouani. May the noise be with you: Adversarial training without adversarial examples, 2023.

[46] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36:69981–70011, 2023.

[47] Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrell, and Dawn Song. Fooling vision and language models despite localization and attention mechanism, 2018.

[48] Raphael Olivier and Bhiksha Raj. Recent improvements of asr models in the face of adversarial attacks, 2022.

[49] Karla Markert, Donika Mirdita, and Konstantin Böttinger. Language dependencies in adversarial attacks on speech recognition systems, 2022.

[50] Zhang Chen, Luca Demetrio, Srishti Gupta, Xiaoyi Feng, Zhaoqiang Xia, Antonio Emanuele Cinà, Maura Pintor, Luca Oneto, Ambra Demontis, Battista Biggio, and Fabio Roli. Over-parameterization and adversarial robustness in neural networks: An overview and empirical analysis, 2024.

[51] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security*, pages 3418–3432, 2023.

[52] Yixin Wu, Ning Yu, Zheng Li, Michael Backes, and Yang Zhang. Membership inference attacks against text-to-image generation models. *arXiv preprint arXiv:2210.00968*, 2022.

[53] Yixin Wu, Ning Yu, Zheng Li, Michael Backes, and Yang Zhang. Membership inference attacks against text-to-image generation models, 2022.

[54] Ninghao Liu, Mengnan Du, Ruocheng Guo, Huan Liu, and Xia Hu. Adversarial attacks and defenses: An interpretation perspective, 2020.

[55] John X Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. *arXiv preprint arXiv:2005.05909*, 2020.

[56] Zi Xiong, Lizhi Qing, Yangyang Kang, Jiawei Liu, Hongsong Li, Changlong Sun, Xiaozhong Liu, and Wei Lu. Enhance robustness of language models against variation attack through graph integration, 2024.

[57] Danele Lunghi, Alkis Simitsis, Olivier Caelen, and Gianluca Bontempi. Adversarial learning in real-world fraud detection: Challenges and perspectives, 2023.

[58] G M Shahariar, Jia Chen, Jiachen Li, and Yue Dong. Adversarial attacks on parts of speech: An empirical study in text-to-image generation, 2024.

[59] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition?, 2023.

[60] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017.

[61] Lichao Sun. Natural backdoor attack on text data, 2021.

[62] Xin Wang, Shuyun Lin, Hao Zhang, Yufei Zhu, and Quanshi Zhang. Interpreting attributions and interactions of adversarial attacks, 2021.

[63] Kaleel Mahmood, Rigel Mahmood, and Marten van Dijk. On the robustness of vision transformers to adversarial examples, 2021.

[64] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition, 2019.

[65] Yi Zeng, Han Qiu, Gerard Memmi, and Meikang Qiu. A data augmentation-based defense method against adversarial attacks in neural networks, 2020.

[66] Zhen Yu, Xiaosen Wang, Wanxiang Che, and Kun He. Texthacker: Learning based hybrid local search algorithm for text hard-label adversarial attack, 2022.

[67] Han Xu, Menghai Pan, Zhimeng Jiang, Huiyuan Chen, Xiaoting Li, Mahashweta Das, and Hao Yang. Towards generating adversarial examples on mixed-type data, 2022.

[68] Abigail Swenor and Jugal Kalita. Using random perturbations to mitigate adversarial attacks on sentiment analysis models, 2022.

[69] Shreyasi Mandal. Defense against adversarial attacks using convolutional auto-encoders, 2023.

[70] Yingdi Wang, Wenjia Niu, Tong Chen, Yingxiao Xiang, Jingjing Liu, Gang Li, and Jiqiang Liu. A training-based identification approach to vin adversarial examples, 2018.

[71] Yigit Alparslan, Ken Alparslan, Jeremy Keim-Shenk, Shweta Khade, and Rachel Greenstadt. Adversarial attacks on convolutional neural networks in facial recognition domain, 2021.

[72] Haonan Zheng, Wen Jiang, Xinyang Deng, and Wenrui Li. Sample-agnostic adversarial perturbation for vision-language pre-training models, 2024.

[73] Yanni Xue, Haojie Hao, Jiakai Wang, Qiang Sheng, Renshuai Tao, Yu Liang, Pu Feng, and Xianglong Liu. Vision-fused attack: Advancing aggressive and stealthy adversarial text against neural machine translation, 2024.

[74] Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. Semantics disentangling for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2327–2336, 2019.

[75] Jiyuan Fu, Zhaoyu Chen, Kaixun Jiang, Haijing Guo, Jiafeng Wang, Shuyong Gao, and Wenqiang Zhang. Improving adversarial transferability of vision-language pre-training models through collaborative multimodal interaction, 2024.

[76] Aminul Huq and Mst. Tasnim Pervin. Adversarial attacks and defense on texts: A survey, 2020.

[77] KiYoon Yoo, Jangho Kim, Jiho Jang, and Nojun Kwak. Detection of word adversarial examples in text classification: Benchmark and baseline via robust density estimation, 2022.

[78] Harrison Gietz and Jugal Kalita. Maskpure: Improving defense against text adversaries with stochastic purification, 2024.

[79] Rajeev Sahay, Rehana Mahfuz, and Aly El Gamal. Combatting adversarial attacks through denoising and dimensionality reduction: A cascaded autoencoder approach, 2018.

[80] Tao Yu, Shengyuan Hu, Chuan Guo, Wei-Lun Chao, and Kilian Q. Weinberger. A new defense against adversarial images: Turning a weakness into a strength, 2019.

[81] Prachi Agrawal, Narinder Singh Punn, Sanjay Kumar Sonbhadra, and Sonali Agarwal. Impact of attention on adversarial robustness of image classification models, 2021.

[82] Georgii Mikriukov, Gesina Schwalbe, Franz Motzkus, and Korinna Bade. The anatomy of adversarial attacks: Concept-based xai dissection, 2024.

[83] Silvia Corbara and Alejandro Moreo. Forging the forger: An attempt to improve authorship verification via data augmentation, 2024.

[84] Inci M. Baytas and Debayan Deb. Robustness-via-synthesis: Robust training with generative adversarial perturbations, 2021.

[85] Yuguang Yao, Jiancheng Liu, Yifan Gong, Xiaoming Liu, Yanzhi Wang, Xue Lin, and Sijia Liu. Can adversarial examples be parsed to reveal victim model information?, 2024.

[86] Chandresh Pravin, Ivan Martino, Giuseppe Nicosia, and Varun Ojha. Adversarial robustness in deep learning: Attacks on fragile neurons, 2022.

[87] Izzat Alsmadi, Kashif Ahmad, Mahmoud Nazzal, Firoj Alam, Ala Al-Fuqaha, Abdallah Khreishah, and Abdulelah Algosaibi. Adversarial attacks and defenses for social network text processing applications: Techniques, challenges and future research directions, 2021.

[88] Prathyusha Devabhakthini, Sasmita Parida, Raj Mani Shukla, and Suvendu Chandan Nayak. Analyzing the impact of adversarial examples on explainable machine learning, 2023.

[89] Mohammed Alkhowaiter, Hisham Kholidy, Mnassar Alyami, Abdulmajeed Alghamdi, and Cliff Zou. Adversarial-aware deep learning system based on a secondary classical machine learning verification approach, 2023.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.