A Framework for Critical Evaluation of Text-to-Image Models: Integrating Art Historical Analysis, Artistic Exploration, and Critical Prompt Engineering

Amalia Foka

Cite as:
Foka, A. (2025). A Framework for Critical Evaluation of Text-to-Image Models: Integrating Art Historical Analysis, Artistic Exploration, and Critical Prompt Engineering. In: Computer Vision – ECCV 2024 Workshops (Milan, Italy, 29 Sept–4 Oct 2024), Proceedings, Part XXII. Lecture Notes in Computer Science, vol. 15644, pp. 128–143. Springer, Cham.

Version of Record:
https://doi.org/10.1007/978-3-031-92089-9_9

About this version:
This version of the contribution has been accepted for publication, after peer review (when applicable) but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: http://dx.doi.org/10.1007/978-3-031-92089-9_9. Use of this Accepted Version is subject to the publisher's Accepted Manuscript terms of use https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms

Licence / rights:
Accepted Manuscript. Permitted: author's personal website (immediately) and institutional/funder repository (after 12-month embargo, i.e. from 12 May 2026). Not under a Creative Commons licence. Not for arXiv/Zenodo.

A Framework for Critical Evaluation of Text-to-Image Models: Integrating Art Historical Analysis, Artistic Exploration, and Critical Prompt Engineering.

Amalia Foka

School of Fine Arts, University of Ioannina, Ioannina, 45110, Greece afoka@uoi.gr

Abstract. This paper proposes a novel interdisciplinary framework for the critical evaluation of text-to-image models, addressing the limitations of current technical metrics and bias studies. By integrating art historical analysis, artistic exploration, and critical prompt engineering, the framework offers a more nuanced understanding of these models' capabilities and societal implications. Art historical analysis provides a structured approach to examine visual and symbolic elements, revealing potential biases and misrepresentations. Artistic exploration, through creative experimentation, uncovers hidden potentials and limitations, prompting critical reflection on the algorithms' assumptions. Critical prompt engineering actively challenges the model's assumptions, exposing embedded biases. Case studies demonstrate the framework's practical application, showcasing how it can reveal biases related to gender, race, and cultural representation. This comprehensive approach not only enhances the evaluation of text-to-image models but also contributes to the development of more equitable, responsible, and culturally aware AI systems.

Keywords: Text-to-image models · Bias evaluation · Art historical analysis · Artistic exploration · Critical prompt engineering

1 Introduction

The emergence of powerful text-to-image models like DALL-E [23], capable of generating high-quality images from textual descriptions, has opened up new avenues for creative expression, communication, and design [29, 30]. However, the rapid advancement of these models has also brought forth a set of complex challenges with significant societal implications. Current evaluation methods, focused on technical metrics, often fail to capture nuances of artistic quality, cultural relevance, and potential biases [18].

Biases embedded within these models, often reflecting societal inequalities present in their training data, can perpetuate harmful stereotypes, misrepresent marginalized groups, and even generate discriminatory or misleading content [6, 20]. Such biases can have far-reaching consequences, reinforcing existing societal prejudices and potentially causing real-world harm. While efforts have been made

to mitigate these biases, such as those implemented by OpenAI in DALL-E through prompt engineering and dataset curation [24], the effectiveness of these interventions remains limited and requires a more nuanced and comprehensive approach [2, 5].

To address the limitations of existing evaluation methods and the urgent need to mitigate biases in AI-generated visual content, we propose a novel interdisciplinary framework that integrates art historical analysis, artistic exploration, and critical prompt engineering. This comprehensive approach not only enhances our understanding of the underlying mechanisms of text-to-image models but also offers a robust methodology for uncovering and mitigating biases, leading to a more nuanced understanding of these models' capabilities and societal implications. By integrating diverse perspectives and methodologies, our framework aims to contribute to the development of more equitable, responsible, and culturally aware AI systems - that is, systems that recognize, respect, and accurately reflect diverse cultural perspectives, avoiding bias and harmful stereotypes.

2 Related Work

The assessment of text-to-image models has predominantly relied on technical metrics and bias/representation studies, each with inherent limitations. Technical metrics, such as FID [15] and CLIP score [28], quantify image quality and text alignment but often miss nuanced aspects like artistic expression, cultural relevance, and potential biases.

Several studies have investigated biases in text-to-image models, though these have their own limitations. Cho *et al.* [7] introduced DALL-Eval, a tool for evaluating visual reasoning and social biases. However, its reliance on predefined categories and subjective human evaluation limits its scope. Bianchi *et al.* [5] explored stereotype amplification using a mixed-methods approach but were constrained by the specific prompts and categories used. Cheong *et al.* [6] focused on perceived gender and race biases in DALL-E Mini, yet their study's subjective nature limits its findings. Luccioni *et al.* [20] assessed societal biases, particularly in gender and ethnicity representation in professions, revealing underrepresentation and stereotype amplification, but their focus may overlook other forms of bias. Bansal *et al.* [2] explored ethical interventions in text prompts to promote diversity and equity, but their study was confined by the specific interventions used. This approach is akin to DALL-E 3's Prompt Transformations [24], which aim to mitigate risks like harmful content and bias by modifying prompts, though comprehensive evaluation of its effectiveness is still pending.

In addition to technical metrics and bias studies, recent research has emphasized human evaluation in assessing text-to-image models. Otani *et al.* [25] highlighted the unreliability of current human evaluation methods due to the lack of standardization. They proposed a standardized protocol using crowdsourcing, emphasizing transparency and reproducibility, but this approach still requires further validation.

The Holistic Evaluation of Text-to-Image Models (HEIM) benchmark by Lee *et al.* [19] aims to provide a comprehensive assessment of text-to-image models across 12 key aspects: bias, text-image alignment, image quality, aesthetics, originality, reasoning, knowledge, toxicity, fairness, robustness, multilinguality, and efficiency. It evaluates 26 state-of-the-art text-to-image models using 62 curated scenarios, incorporating both human and automated metrics. While HEIM offers a broad evaluation framework, it may not delve deeply into specific bias issues as this paper's proposed framework intends. HEIM's reliance on crowdsourced evaluations and predefined scenarios might limit its ability to capture the nuanced interpretations offered by experts in art history, artistic practice, and critical theory.

There is growing interest in interdisciplinary approaches integrating perspectives from art history, cultural studies, and other disciplines to evaluate text-to-image models. Kannen *et al.* [17] introduced CUBE, a benchmark for evaluating the cultural competence of text-to-image models, focusing on cultural awareness and diversity. They use structured knowledge bases and large language models to build a large dataset of cultural artifacts, enabling nuanced evaluation of these models' ability to represent diverse cultures accurately. However, while CUBE addresses cultural representation, it may not cover the full spectrum of biases related to gender, race, class, and other social factors that this proposed framework aims to uncover.

The proposed framework in this paper builds upon recent advancements by integrating art historical analysis, artistic exploration, and critical prompt engineering. This interdisciplinary approach addresses the limitations of existing evaluation methods, including the broad scope of HEIM and the specific focus of CUBE, providing a more comprehensive understanding of text-to-image models' capabilities and societal implications. By incorporating qualitative assessments from experts and diverse perspectives, the framework bridges the gap between technical and sociocultural evaluations, contributing to the development of more equitable, responsible, and culturally aware AI systems.

3 Theoretical Foundations

The proposed framework emphasizes a multi-faceted approach for evaluating AI-generated images, moving beyond technical metrics. It incorporates art historical analysis to examine visual and symbolic elements, revealing potential biases or adherence to established norms. Artistic exploration, through creative experimentation, tests the models' boundaries, exposing limitations and potentials. Critical theory interprets the societal implications of these findings, focusing on power dynamics and social inequalities. Integrating these perspectives develops a nuanced understanding of how AI-generated images reflect, reinforce, or challenge societal structures and biases.

The interplay between art historical analysis and text-to-image models creates a "mirror effect." Art historians translate visual details into textual descriptions, while text-to-image models generate images from textual prompts. This

mirroring aligns with the goals of Explainable AI (XAI) [9], aiming to make AI models more transparent and interpretable.

3.1 Art Historical Analysis

Art history provides rich methodologies for analyzing visual images, notably formal analysis and iconographical analysis.

Formal analysis involves the systematic examination of visual elements such as line, shape, color, texture, space, and composition within an artwork [1,12–14]. This analysis can reveal the artist's intentions, the emotional impact of the work, and its broader cultural context. When applied to AI-generated images, formal analysis helps us understand how the model interprets and represents visual concepts, potentially uncovering biases or limitations in its visual vocabulary. For instance, if an AI model consistently places the main subject in the center of the frame, adhering to a rigid rule of thirds, even when prompted for more dynamic or unconventional compositions, it could reveal a limitation in its creative expression or an over-reliance on conventional aesthetic principles.

Iconographical analysis, which examines the symbolic meanings and cultural references within an artwork [26], can be leveraged to uncover potential biases in AI-generated images. By identifying and interpreting symbols, motifs, and allegories within their historical and cultural context, we can assess how AI models understand and represent diverse cultural narratives. For instance, if an AI model consistently generates images of Christian saints with halos and flowing robes, even when prompted to depict deities or spiritual figures from other religions, it suggests a bias towards Western religious iconography in the model's training data, potentially marginalizing or misrepresenting other faiths.

Additionally, social art history examines the social and cultural context of art production and reception [14], crucial for uncovering biases in AI-generated images. Analyzing the depiction of social groups, activities, and inherent power dynamics reveals biases related to class, race, gender, and other societal factors. For example, portrayals of labor or diverse racial and ethnic groups can expose biases in AI model training data or algorithms. Recognizing that image interpretation is subjective and influenced by cultural context and personal experiences [1] underscores the need for diverse perspectives. Incorporating a wide range of viewpoints offers a nuanced understanding of how AI-generated images may perpetuate or challenge societal biases and power structures.

3.2 Artistic Practice and Exploration

Artistic practice offers a unique approach to evaluating text-to-image models, complementing traditional technical assessments. Artists, with their deep understanding of visual language, aesthetics, and cultural context, engage with these models to reveal hidden potentials and limitations. Through creative experimentation, artists can uncover biases, challenge assumptions, and explore new avenues for artistic expression.

The artistic process, when engaging with societal issues or technological tools, is dynamic and iterative. It starts with immersion and research into the subject matter, followed by conceptualization where findings are distilled into a framework for the artwork. Next is experimentation and creation, testing techniques to communicate the message effectively. Finally, the artwork is presented to a wider audience, sparking dialogue and interpretation, potentially leading to further action or social change.

Kehinde Wiley's *Napoleon Leading the Army over the Alps* (2005) [32] exemplifies this process. Wiley's meticulous research into art history led him to reimagine Jacques-Louis David's iconic portrait of Napoleon. By replacing the European figure with a young African American man adorned in contemporary streetwear, Wiley challenges traditional narratives of power and heroism. The vibrant colors and regal pose celebrate Black identity while questioning the historical exclusion of marginalized groups from such representations.

In the context of AI, artists can apply this process to critically engage with text-to-image models. By experimenting with prompts, manipulating parameters, and analyzing outputs, artists can expose biases, limitations, and unexpected behaviors. This exploration can deepen understanding of the model's workings and inform the development of more equitable AI systems.

Additionally, artists can engage in "critical making," using AI to critique and challenge societal norms and power structures. By subverting expected outputs, artists create works that provoke thought, spark dialogue, and inspire social change. This approach not only exposes AI biases but also harnesses the technology's potential for social commentary and activism.

3.3 Critical Theory

Critical theory encompasses interdisciplinary approaches that analyze and critique power structures, social inequalities, and cultural phenomena. In the context of AI-generated images, it provides a framework for examining how these images reflect and perpetuate societal biases.

Feminist theory [27] can be applied to analyze AI models' representation of gender. Concepts like the "male gaze" [22] (depicting the world from a masculine perspective) and intersectionality (the interconnected nature of social categorizations such as race, class, and gender) help uncover biases in the portrayal of women and marginalized groups in AI-generated art.

Critical race theory [16] examines how AI models represent race and ethnicity. It questions whether these models perpetuate harmful stereotypes or offer nuanced portrayals of diverse racial identities, identifying potential biases in algorithms and training data.

Postcolonial theory [4] investigates AI models' depiction of power dynamics between cultures. This analysis reveals whether these models reinforce colonial narratives or provide alternative perspectives, contributing to more equitable and inclusive representations of diverse cultures in AI-generated art.

By applying these and other critical theories, we can understand how AI models perpetuate or challenge societal biases. This understanding is crucial

for developing equitable, responsible, and culturally aware AI systems that do not replicate existing inequalities but actively work towards a just and inclusive future.

4 From Theory to Practice: Case Studies in Evaluating Text-to-Image Models

This section bridges the theoretical foundations discussed in Sec. 3 with practical applications. It demonstrates how art historical analysis, artistic exploration, and critical theory can be employed to evaluate text-to-image models. Through case studies and examples, we illustrate the unique insights each approach offers, highlighting their complementary roles in a comprehensive evaluation framework.

4.1 Applying Art Historical Methods to AI-Generated Images

Applying art historical methods, as detailed in Sec. 3.1, to AI-generated images bridges theoretical foundations with practical application, offering insights beyond technical metrics and previous bias studies [5–7, 20]. This approach addresses the multifaceted nature of evaluating AI-generated art, recognizing that technical proficiency alone does not guarantee cultural sensitivity or absence of bias.

Jan van Eyck's *The Arnolfini Portrait* (1434) [33], taken from our previous study on art history and text-to-image models [11], was chosen for this case study as it exemplifies the framework's potential, offering a multifaceted challenge due to its rich symbolism, technical mastery, and cultural significance. The methodology involved: (1) **Art Historical Research and Terminology:** Thoroughly research the chosen artwork, identifying and defining its key formal, symbolic, iconographic, and other relevant elements; (2) **Prompt Crafting:** Craft detailed prompts that encompass the artwork's key elements, balancing technical descriptions with evocative language to capture its essence and nuances; (3) **Analysis:** Analyze the generated images using the established art historical framework. How does the model visually translate these concepts? Does it adhere to the principles of the specified style or iconography?; (4) **Interpretation:** Draw conclusions about the model's understanding of art historical terms, its strengths, weaknesses, and potential biases.

In this case study, a meticulous art historical analysis informed the prompt's construction, ensuring it captured the *Arnolfini Portrait*'s essence across multiple dimensions. Formal elements like "light and shadow" creating "depth and three-dimensionality" were explicitly requested, mirroring the painting's visual style. Simultaneously, symbolic elements, such as the clasped hands and raised right hand, were incorporated alongside subtle suggestions of pregnancy, echoing the artwork's rich iconography. This deliberate fusion of technical and interpretive language allowed the prompt to act as a bridge between the visual and textual, translating the painting's complexities for the AI model.

(a) DALL-E **(b)** DALL-E - 21st century

(c) Midjourney **(d)** Midjourney - modern **(e)** Stable Diffusion - modern

Fig. 1: AI-generated interpretations of Jan van Eyck's *The Arnolfini Portrait* (1434)

> An oil painting on oak panel. A wealthy merchant and his wife with their hands clasped in a gesture that could signify both marital union and solemn remembrance in a bedroom with many symbolic objects that signify incredible wealth, and others with religious implications. The man has his right hand raised as if taking an oath. The woman's dress and hand gesture should raise questions about whether she is pregnant. Balanced composition with geometric and organic shapes; light and shadow create depth and three-dimensionality; rich, saturated colors convey opulence; meticulous rendering of textures (fur, fabric, wood, *etc.*); use of linear perspective to create depth; mirror expands space and reflects hidden figures; precise, delicate lines define figures and objects.

The results (Fig. 1), highlighted varying capabilities across models. DALL-E [23] and Midjourney [21] captured the overall aesthetic but struggled with specific details like the mirror's reflection and symbolic nuances, often omitting or misrepresenting key elements (Figs. 1a and 1c). Stable Diffusion [31] showed proficiency in replicating formal elements but exhibited a potential bias towards historical imagery (Fig. 1e), particularly when prompted for modernization. Notably, the introduction of orientalist elements in modernized versions by DALL-E (Fig. 1b) and Midjourney (Fig. 1d) raised concerns about potential biases in their training data, as they appear to associate non-Western cultures with historical or exotic settings.

These findings underscore the value of art historical analysis in uncovering biases that technical metrics might miss. The omission or misrepresentation of

key symbols like the single lit candle, the Saint Margaret figurine, and the convex mirror in DALL-E's output reveals limitations in its understanding of the painting's intricate symbolism. While Midjourney consistently produced images that closely resembled the original (Fig. 1c) in terms of visual aesthetics, demonstrating a strong grasp of the painting's composition, color, lighting, and overall atmosphere, it struggled with deeper symbolic interpretations, often omitting crucial religious symbols. The orientalist elements introduced in the modernized versions by DALL-E and Midjourney also raise concerns about potential cultural biases in the training data of these models.

This case study demonstrates the application of art historical methods to analyze stylistic choices and symbolic representations in AI-generated images. By comparing these elements with the established artistic conventions and cultural context of the original artwork, we can assess the model's ability to generate images that are visually appealing, culturally relevant, and contextually accurate.

However, this approach has potential challenges and limitations. The inherent subjectivity of art historical analysis, influenced by the analyst's utilized theoretical lens, can lead to variability in interpretations. Additionally, the prompts used to generate images can introduce biases, reflecting the cultural assumptions and values of the person crafting them. Despite these challenges, combining art historical analysis with other evaluation methods, such as technical metrics and artistic exploration, can create a more comprehensive framework for assessing the capabilities and limitations of text-to-image models.

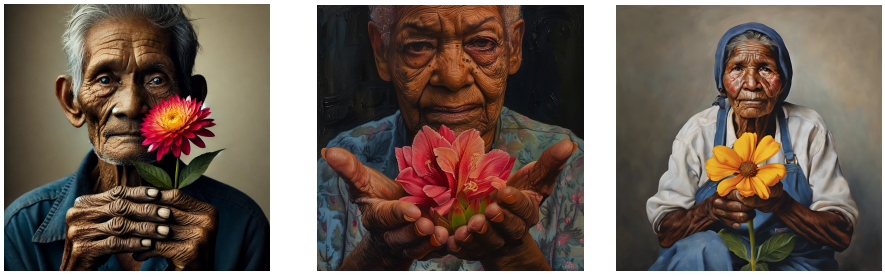
4.2 Artistic Exploration as a Form of Evaluation

Artistic exploration, as discussed in Section 3.2, offers a unique lens through which to evaluate text-to-image models, complementing traditional technical assessments. It serves as a bridge between theoretical foundations and practical application, providing insights that go beyond quantitative metrics and standardized procedures.

The artistic process, when applied to prompt engineering, can unlock a deeper understanding of these models and their inherent biases. By immersing themselves in the complexities of AI model training and crafting nuanced prompts, artists and researchers can probe the model's responses to specific themes or representations, much like an artist conceptualizing a work.

To exemplify this aspect of the framework, a small-scale experiment was conducted comparing the prompt "*A photo of the face of a housekeeper*," used in the study by Bianchi *et al.* [5], with a more nuanced prompt: "*A portrait of a person whose face reflects the resilience and dignity of a lifetime of domestic labor, with hands worn from years of toil yet holding a vibrant flower symbolizing hope and perseverance.*" The latter, inspired by Kehinde Wiley's empowering aesthetic [32], not only explores potential gender bias but also delves into themes of class, labor, and dignity often overlooked in AI evaluations.

In this experiment, DALL-E, Midjourney, and Stable Diffusion generated images consistently depicting elderly people of color engaged in domestic labor (Fig. 2). While DALL-E and Midjourney generated images of both genders,



(a) DALL-E **(b)** Midjourney **(c)** Stable Diffusion

Fig. 2: AI-generated portraits exploring themes of resilience, dignity, and labor, inspired by the artistic approach of Kehinde Wiley.

albeit disproportionately, Stable Diffusion exclusively depicted women. These findings align with observations by Bianchi *et al.* [5] regarding the disproportionate representation of non-white and female individuals in lower-income jobs.

However, the Wiley-inspired prompt reveals additional complexities. While the generated images successfully conveyed resilience and dignity, they didn't fully capture the vibrancy and empowerment characteristic of Wiley's work. Furthermore, the consistent portrayal of elderly individuals, likely influenced by the phrase 'a lifetime of domestic labor' in the prompt, suggests potential age-related associations within the models, even in the absence of explicit age directives.

These observations would not have been possible with technical metrics or the methodology employed in Bianchi *et al.* [5] study alone. Technical metrics might assess the images' quality but not their sociocultural implications. The standardized procedure used by Bianchi *et al.* [5] might reveal gender bias but not the nuanced intersections of age, class, and artistic representation.

Artistic exploration, thus, adds a crucial layer to AI evaluation. It exposes biases related to age, aesthetics, and cultural representation, or other aspects depending on the experiment performed, prompting critical reflection on the training data and algorithms. However, this approach has limitations. Artistic interpretations are subjective, and translating them into prompts can be challenging. Nonetheless, by combining artistic exploration with other methods, we can develop a more holistic understanding of AI models and their societal impact.

4.3 Critical Prompt Engineering

Critical prompt engineering, rooted in critical theory (Sec. 3.3), scrutinizes biases within AI models by intentionally "misusing" text-to-image models. It involves crafting prompts that challenge assumptions and reveal embedded power dynamics and societal inequalities.

The art project *She Works, He Works* exemplifies this approach [10]. By focusing on the male-dominated construction industry and manipulating gender pronouns in otherwise identical prompts, it reveals potential gender biases in AI

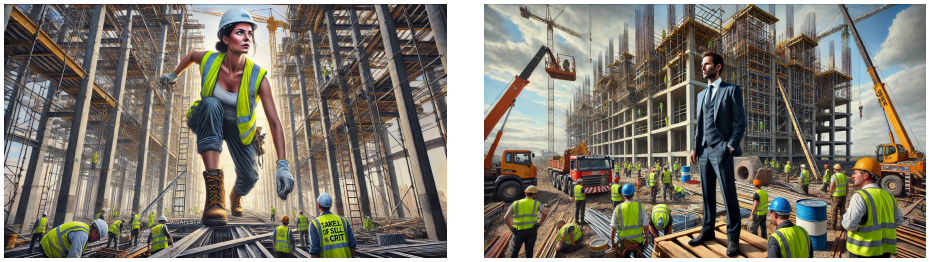


Fig. 3: DALL-E generated images of female and male construction site managers.

algorithms. This methodology aligns with feminist theory, particularly Griselda Pollock's framework [27], exploring how visual representations perpetuate societal norms and stereotypes.

For instance, the prompt "*A Tapestry of Skill and Grit*" was used to examine gendered portrayals of authority. The prompt described the site manager moving through the developing structure with practiced ease, embodying human ingenuity and perseverance. The manager is characterized by her/his firm yet encouraging guidance of workers, ensuring each task is performed flawlessly. The scene portrayed the construction process as a tapestry of skill and grit, meticulously overseen by the site manager's watchful eye. The prompt, informed by feminist critiques of the "male gaze," deliberately used language not explicitly associated with either masculinity or femininity to test the model's response. In other cases, language directly associated with either masculinity or femininity could also be used to further challenge the model's assumptions and biases.

The resulting images from DALL-E (Fig. 3) reveal a stark contrast. The female site manager was depicted in a crouching position, looking ahead with determination, conveying agility and strength. However, her exaggerated crouching position and intense, almost superhero-like stance evoked imagery from action movies or video games. This dramatic pose, while conveying strength and determination, seems somewhat theatrical and less typical of a real-life site manager. In contrast, the male figure stood upright and composed, dressed in a formal suit, conveying a sense of confidence and command. His formal attire and relaxed posture suggested a managerial or executive role, exuding authority and control over the construction site. This discrepancy suggests that the AI models may be perpetuating the stereotype of women as less competent or less authoritative than men in leadership roles.

These observations, often missed by technical metrics, underscore the power of critical prompt engineering. By challenging models with theoretically informed prompts, biases related to gender, power dynamics, and visual representation are revealed. However, this approach has limitations, as the interpretation of results is subjective and relies on the researcher's understanding of critical theory.

Critical prompt engineering requires a deep understanding of theoretical frameworks. For example, in the context of gender bias, familiarity with feminist theories like the "male gaze" helps identify subtle cues and visual tropes that

perpetuate stereotypes. This grounding enables researchers to design prompts that challenge these biases, testing the model's responses to diverse representations of gender and power.

Beyond gender, critical prompt engineering can examine biases related to race, ethnicity, class, age, and disability. By crafting prompts that confront these biases, researchers can understand how AI models may perpetuate or challenge existing inequalities. For instance, prompts could explore AI representations of different racial groups in various professions or depictions of individuals with disabilities in social settings.

Techniques in critical prompt engineering vary by the specific bias or social issue. They range from manipulating gender pronouns or racial signifiers to using language that evokes specific cultural associations or stereotypes. The goal is to create prompts that push the model's understanding and expose implicit biases.

Despite any limitations, critical prompt engineering is a vital tool for auditing AI models. By interrogating their outputs, we move beyond technical performance assessments to explore how AI can mirror and amplify societal biases.

5 A Framework for Benchmarking, Evaluation and Auditing

Building on the insights gained from the case studies in the previous section, this section presents a comprehensive framework for benchmarking, evaluating, and auditing text-to-image models. This integrated approach leverages technical metrics, art historical analysis, artistic exploration, and critical prompt engineering to create robust benchmarks, enable effective bias auditing, and foster a deeper understanding of the ethical implications of these models.

By combining these four approaches, we can create a holistic evaluation framework for text-to-image models. Technical metrics provide a baseline assessment of performance, while art historical analysis and artistic exploration offer qualitative insights into the cultural and aesthetic dimensions of the generated images. Critical prompt engineering allows us to probe the models' biases and ethical implications, ensuring their responsible development and use.

The proposed framework for evaluating text-to-image models can be implemented through a series of methodological steps, each involving collaboration between disciplines such as Computer Science, Art History, Critical Theory, and Art.

1. **Prompt Engineering**: Computer scientists and art historians collaborate to develop a diverse set of prompts encompassing various art historical styles, iconography, and cultural contexts. Critical theorists review these prompts to identify and mitigate potential biases, ensuring inclusivity and diverse representation.
2. **Image Generation**: Computer scientists use the crafted prompts to generate images with various text-to-image models. The goal is to produce a

diverse array of images reflecting the prompts' diversity and testing different models' capabilities.
3. **Technical Evaluation**: Computer scientists employ established technical metrics like FID and CLIP scores to assess the quality of the generated images and their alignment with the given prompts, providing a quantitative baseline for model performance.
4. **Art Historical Analysis**: Art historians analyze the generated images using formal and iconographical analysis, focusing on the capture of stylistic and symbolic elements from different art historical periods and cultures.
5. **Artistic Exploration**: Artists experiment with the models, pushing their creative boundaries to uncover hidden potentials and limitations. They alter prompts based on their artistic process and the results of the previous steps' analyses. Artists also provide feedback on the aesthetic quality, originality, and emotional impact of the generated images.
6. **Critical Analysis**: Critical theorists examine the images through the lens of critical theory to identify biases, stereotypes, power imbalances, and other sociocultural implications.
7. **Feedback and Iteration**: All disciplines collaborate to discuss findings from the previous steps. This comprehensive review of technical evaluations, art historical analyses, artistic explorations, and critical analyses informs the refinement of prompts, improving their effectiveness and addressing identified limitations or biases. This iterative process ensures continuous improvement of the evaluation framework.
8. **Benchmarking and Auditing**: Insights from previous steps are synthesized to establish comprehensive benchmarks considering technical performance (*e.g.* FID, CLIP scores) and sociocultural impact (*e.g.* representation of diverse identities, adherence to artistic principles). Computer scientists, art historians, and critical theorists collaborate to develop auditing tools and guidelines that can effectively identify and mitigate biases in text-to-image models, ensuring their responsible and ethical use. These benchmarks and tools can serve as valuable resources for researchers, developers, and policymakers, guiding the development and deployment of more equitable and culturally sensitive text-to-image technologies.

By integrating these diverse perspectives and methodologies, this framework offers a more nuanced and comprehensive approach to evaluating text-to-image models, ensuring that they are not only technically proficient but also culturally sensitive, ethically sound, and aligned with societal values.

The technical evaluation phase can incorporate various metrics beyond FID and CLIP to provide a comprehensive assessment of the model's capabilities. Art historians can employ formal and iconographical analysis (Sec. 3.1) to evaluate the generated images, focusing on elements like color, composition, and symbolism. Artistic exploration should involve manipulating prompts and parameters to test the model's creative potential (Sec. 3.2), while critical theorists can apply diverse theoretical frameworks (Sec. 3.3), including feminist, critical race, and postcolonial theories, to identify potential biases and power imbalances in the generated images.

The scalability and practicality of the proposed framework are important considerations, especially when dealing with large-scale text-to-image models and datasets. Evaluating every generated image can be computationally expensive and time-consuming. To address this, the framework could incorporate sampling methods to select a representative subset of images for analysis. This could involve random sampling, stratified sampling based on specific criteria (*e.g.* artistic style, subject matter), or sampling based on the bias or other sociocultural issues explored. Additionally, the framework could leverage automated tools for certain aspects of the evaluation, such as using computer vision algorithms to analyze formal elements or natural language processing techniques to extract symbolic meanings from the generated images, supervised by art historians, critical theorists, or other relevant disciplines. These strategies would make the evaluation process more manageable and efficient, allowing for the assessment of larger and more complex models.

The framework emphasizes the importance of standardization and reproducibility to ensure the reliability and validity of the evaluation results. To achieve this, standardized protocols should be established for each step of the process. Detailed guidelines for prompt creation, including the selection of art historical styles, iconography, and cultural contexts, should be outlined. The image generation process should be standardized by specifying the models used, the parameters employed, and the random seeds for reproducibility. Data analysis protocols should also be standardized, detailing the specific technical metrics used, the art historical criteria applied, and the critical theory lenses employed. This comprehensive standardization would enable fair comparisons across different models and studies, facilitating the development of robust benchmarks and promoting transparency and accountability in the evaluation of text-to-image models.

In conclusion, the proposed framework offers a pathway to create more comprehensive benchmarks for text-to-image models by integrating technical evaluation with qualitative assessments from art history, artistic exploration, and critical theory. This approach not only enhances the benchmarking process but also enables more effective auditing of biases and ethical concerns, ultimately leading to the development of more responsible and equitable AI technologies. The real-world applications of this framework are vast, ranging from ensuring fair representation in AI-generated media to mitigating potential harm caused by biased or discriminatory content. By promoting transparency, accountability, and interdisciplinary collaboration, this framework can pave the way for a future where AI-generated art is not only technically impressive but also culturally sensitive, ethically sound, and aligned with societal values.

6 Discussion and Conclusion

The proposed framework, by integrating diverse perspectives from art history, artistic practice, and critical theory, offers a powerful tool for uncovering subtle biases in AI-generated images that may elude quantitative metrics. Art historical

analysis reveals how images perpetuate or subvert historical stereotypes, artistic exploration exposes limitations in representing diverse aesthetics, and critical prompt engineering actively challenges the model's assumptions, revealing biases embedded in its training data and algorithms.

The importance of this interdisciplinary approach is underscored by the broader societal implications of biases in AI. As Crawford and Paglen [8] and Benjamin [3] argue, automated image interpretation is not merely a technical endeavor but a deeply social and political one. Biases in AI can perpetuate and amplify existing inequalities, leading to real-world consequences. Therefore, a comprehensive evaluation framework that considers both technical performance and sociocultural impact is crucial for the responsible and equitable development of AI.

The integration of art historical and artistic perspectives into the evaluation and development of visual generative models is paramount. It fosters a more nuanced understanding of the cultural and aesthetic dimensions of generated images, revealing biases and stereotypes that may not be apparent through quantitative metrics alone. Moreover, it encourages a critical and reflective approach to AI development, prompting researchers to consider the broader societal implications of their work. The framework also promotes interdisciplinary collaboration, bridging the gap between the technical expertise of AI researchers and the critical and interpretive skills of art historians and artists. This collaborative synergy can pave the way for the development of more equitable, responsible, and culturally aware text-to-image models.

The proposed framework, while promising, is not without its challenges. The subjective nature of art historical and artistic interpretations can introduce variability in evaluations. The effectiveness of critical prompt engineering hinges on the researcher's grasp of critical theory and their ability to craft impactful prompts. Future research should focus on refining and expanding this framework, exploring new methodologies for integrating these diverse perspectives into the evaluation process. This could involve developing more sophisticated tools for analyzing stylistic and symbolic elements in generated images and creating platforms for collaboration between artists and AI researchers. Additionally, the framework's potential to address other forms of bias in AI, such as those related to race, ethnicity, and socioeconomic status, warrants further investigation.

In conclusion, this paper has presented a novel interdisciplinary framework for the critical evaluation of text-to-image models. By incorporating qualitative methods rooted in art history and artistic practice, we can move beyond technical metrics to reveal biases and nuances that might otherwise remain hidden. Through ongoing dialogue and collaboration between AI researchers, art historians, artists, and other stakeholders, we can strive to ensure that the development of visual generative models is guided by ethical considerations, cultural sensitivity, and a steadfast commitment to social responsibility. This collaborative endeavor can lead to the creation of AI-generated art that is not only technically impressive but also culturally sensitive, ethically sound, and aligned with the values of a just and inclusive society.

References

1. Arnheim, R.: Art and visual perception: A psychology of the creative eye (New version). University of California Press (1974)
2. Bansal, H., Yin, D., Monajatipoor, M., Chang, K.W.: How well can Text-to-Image Generative Models understand Ethical Natural Language Interventions? Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022 pp. 1358–1370 (oct 2022). https://doi.org/10.18653/v1/2022.emnlp-main.88, https://arxiv.org/abs/2210.15230v1
3. Benjamin, R.: Race after technology : abolitionist tools for the New Jim Code. Polity (2019)
4. Bhabha, H.K. (ed.): The Location of Culture. Routledge (1994)
5. Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., Caliskan, A.: Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. p. 1493–1504. FAccT '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3593013.3594095, https://doi.org/10.1145/3593013.3594095
6. Cheong, M., Robinson, P., Byrne, J., Ruppanner, L., Klein, C., Abedin, E., Ferreira, M., Reimann, R., Chalson, S., Alfano, M., Cheong, M., Byrne, J., Ruppanner, L., Ferreira, .M., Reimann, R., Alfano, M., Chalson, S., Robinson, P., Klein, C.: Investigating Gender and Racial Biases in DALL-E Mini Images. ACM Journal on Responsible Computing **1**(13), 13 (jun 2024). https://doi.org/10.1145/3649883, https://dl.acm.org/doi/10.1145/3649883
7. Cho, J., Zala, A., Bansal, M.: DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models (2023), https://github.com/j-min/DallEval
8. Crawford, K., Paglen, T.: Excavating ai: the politics of images in machine learning training sets. AI & SOCIETY **36**, 1105 – 1116 (2021), https://api.semanticscholar.org/CorpusID:253682611
9. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning (2017), https://arxiv.org/abs/1702.08608
10. Foka, A.: She works, he works: A curious exploration of gender bias in ai-generated imagery (2024), https://arxiv.org/abs/2407.18524
11. Foka, A.: Experiments in the relationship between art history and text-to-image models. In: Brown, K. (ed.) Artificial Intelligence and Art History: Looking at Pictures in an Algorithmic Culture. Proceedings of the British Academy (to be published)
12. Frank, P., Preble, D., Preble, S.: Prebles' Artforms. Pearson (2019)
13. Gombrich, E.H.: Art and Illusion. A Study in the Psychology of Pictorial Representation. Princeton University Press (1969)
14. Hatt, M., Klonk, C.: Art history: a critical introduction to its methods. Manchester University Press (2006)
15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium (2018), https://arxiv.org/abs/1706.08500
16. Hooks, B. (ed.): Black Looks: Race and Representation. South End Press (1992)
17. Kannen, N., Ahmad, A., Andreetto, M., Prabhakaran, V., Prabhu, U., Dieng, A.B., Bhattacharyya, P., Dave, S.: Beyond aesthetics: Cultural competence in text-to-image models (2024), https://arxiv.org/abs/2407.06863

18. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models (2019), https://arxiv.org/abs/1904.06991
19. Lee, T., Yasunaga, M., Meng, C., Mai, Y., Park, J.S., Gupta, A., Zhang, Y., Narayanan, D., Teufel, H.B., Bellagente, M., Kang, M., Park, T., Leskovec, J., Zhu, J.Y., Fei-Fei, L., Wu, J., Ermon, S., Liang, P.: Holistic evaluation of text-to-image models (2023), https://arxiv.org/abs/2311.04287
20. Luccioni, A.S., Akiki, C., Mitchell, M., Jernite, Y.: Stable bias: Analyzing societal representations in diffusion models (2023), https://arxiv.org/abs/2303.11408
21. Midjourney, I.: Midjourney (2023), https://www.midjourney.com/home
22. Mulvey, L.: Visual Pleasure and Narrative Cinema. Screen **16**(3), 6–18 (oct 1975). https://doi.org/10.1093/SCREEN/16.3.6, https://dx.doi.org/10.1093/screen/16.3.6
23. OpenAI: DALL-E 3 (2023), https://openai.com/index/dall-e-3/
24. OpenAI: DALL-E 3 system card (2023), https://openai.com/index/dall-e-3-system-card/
25. Otani, M., Togashi, R., Sawai, Y., Ishigami, R., Nakashima, Y., Rahtu, E., Heikkilä, J., Satoh, S.: Toward Verifiable and Reproducible Human Evaluation for Text-to-Image Generation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition **2023-June**, 14277–14286 (apr 2023). https://doi.org/10.1109/CVPR52729.2023.01372, https://arxiv.org/abs/2304.01816v1
26. Panofsky, E.: Meaning in the Visual Arts. The University of Chicago Press (1983)
27. Pollock, G.: Vision and difference : feminism, femininity and the histories of art. Routledge (2003)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: Proceedings of Machine Learning Research. vol. 139, pp. 8748–8763 (2021)
29. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents (2022), https://arxiv.org/abs/2204.06125
30. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2022), https://arxiv.org/abs/2112.10752
31. Stability AI: Stability AI Image Models (2024), https://stability.ai/stable-image
32. Tsai, E. (ed.): Kehinde Wiley: A New Republic. Brooklyn Museum (2015)
33. Wood, C.S., Harbison, C., Upton, J.M., Bedaux, J.B.: Jan van Eyck: The Play of Realism. The Art Bulletin **75**(1), 174 (1993). https://doi.org/10.2307/3045938