Text-to-image Generation: A Structured Review of Adaptation Strategies, Efficiency, and Environmental Sustainability

Subuhi K. Ansari *, Fatma A. Alqahtani, Chamandeep Kaur, Wafaa Abu Shamlah, Najla Mohammed Galam, and Nedaa Abduaziz Hadi

Department of Computer Science, College of Engineering & Computer Science, Jazan University, Jazan, Saudi Arabia
Email: subuhiwasim_786@yahoo.co.in (S.K.A.); falqahtani@jazanu.edu.sa (F.A.A.); Kaur.Chaman83@gmail.com (C.K.); wahakami@jazanu.edu.sa (W.A.S.); Najlagalam@gmail.com (N.M.G.); nahadi@jazanu.edu.sa (N.A.H.)
*Corresponding author

***Abstract*—Text-to-Image (T2I) generation has advanced rapidly with the emergence of diffusion, transformer-based, and vision–language models. These approaches enable high-quality, semantically aligned image synthesis from textual prompts. In addition, zero-shot and few-shot adaptation strategies and parameter-efficient fine-tuning methods have improved the flexibility and scalability of T2I systems. However, these advancements have introduced significant computational and environmental challenges. In this paper, we present a structured review of T2I generation, focusing on the relationships between model architectures, adaptation strategies, computational efficiency and environmental sustainability. The analysis synthesizes 85 studies and introduces a unified mapping that links techniques to their efficiency and environmental implications. The findings show that although parameter-efficient methods, such as Low-Rank Adaptation (LoRA), adapters, and prompt tuning, effectively reduce adaptation costs during the adaptation stage, the overall computational burden remains high because of the reliance on large pre-trained models. Furthermore, although approximately 61% of studies consider computational efficiency, only approximately one-third provide an empirical evaluation of environmental impact. This highlights a clear imbalance between performance optimization and sustainability considerations. Our study emphasizes the need for integrated evaluation frameworks and energy-aware model designs. This contributes to a more structured qualitative understanding of the efficiency–sustainability trade-offs in T2I systems.**

***Keywords*—text-to-image generation, diffusion models, zero-shot learning, parameter-efficient fine-tuning, computational efficiency, green Artificial Intelligence (AI)**

I. Introduction

A. Background of T2I Generation

Text-to-Image (T2I) generation is a combination of Natural Language Processing (NLP) and Computer Vision (CV) in artificial intelligence. There has been a transition in the field from fundamental Generative Adversarial Networks (GANs), which generate realistic images through adversarial training with generator and discriminator models, to more advanced frameworks such as diffusion and transformers. GANs achieve high-resolution outputs [1], whereas diffusion models achieve modern fidelity and controllability based on input text [2, 3]. Diffusion models represent a move using probabilistic and sampling approaches to transform noise into clean images with diverse attributes [2]. Simultaneously, vision-language foundation models and transformer-based models came into being and gained popularity. These models work particularly on subject-driven and scene-level details of the input text through prior learning and adapters [3–5]. Architectures such as Contrastive Language-Image Pre-training (CLIP)-based systems have been successful in proving the effectiveness of shared embedding spaces for guiding image synthesis and assisting zero-shot manipulation [4]. In addition, hybrid approaches that combine diffusion models with vision-language frameworks successfully increase the diversity and controllability of the generated outputs. Diffusion models, transformer architectures, and multimodal learning have come together to provide a new definition for the current state of T2I research. This move prioritizes generation quality and flexibility through large-scale pre-trained systems [2, 4, 5]; however, it also introduces new challenges related to computational cost and efficiency.

Zero-shot and few-shot learning are important in modern Text-to-Image (T2I) systems. They enable flexible image generation with minimal, or no task-specific training required. In the field of T2I generation, zero-shot learning refers to the generation of images for unseen prompts without task-specific examples, enabling immediate mapping from text to perceptual outputs in shared embedding spaces [2, 5, 6]. Few-shot learning extends this by including a small number of examples to

Manuscript received December 29, 2025; revised April 22, 2026; accepted May 15, 2026; published September 4, 2026.

adjust the model to new concepts and styles [7–9]. These approaches provide fine outputs when generating novel objects, scenes, and artistic styles [2, 5, 9, 10]. However, while they reduce the cost of task-specific adaptation, they remain dependent on large pre-trained models, which continue to impose substantial computational and environmental overheads. Fig. 1 shows the general process of generating an image from a given text input prompt.

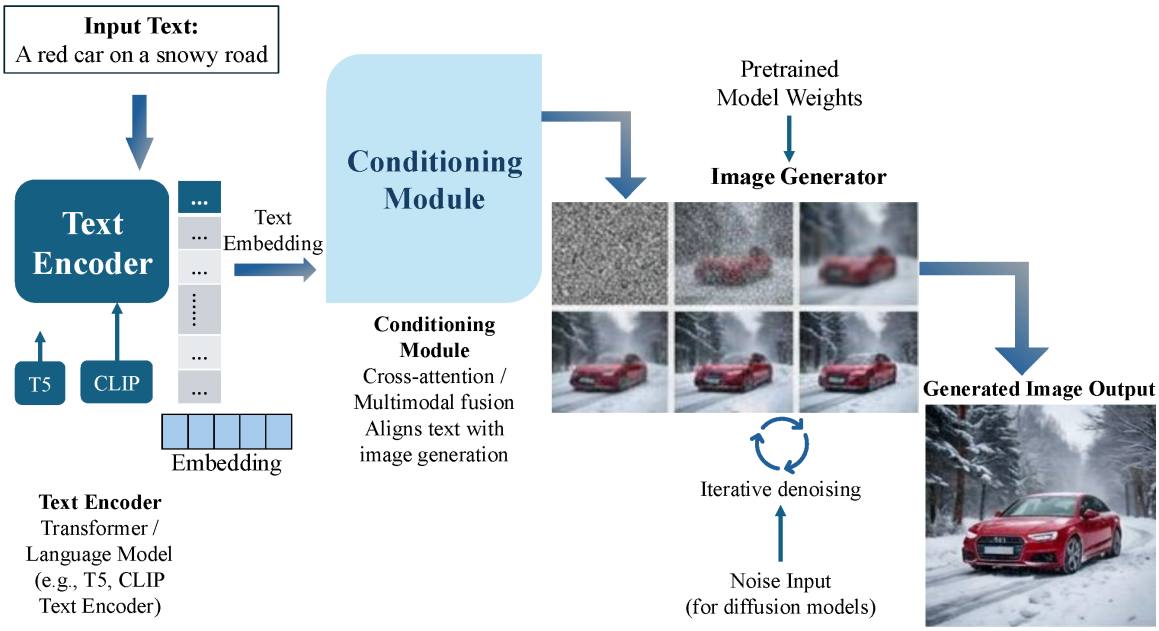

Fig. 1. General pipeline of text-to-image generation systems, showing text encoding, conditioning, and image synthesis stages.

B. *Challenges in Computational Efficiency and Environmental Sustainability in T2I Systems*

Modern Text-to-Image (T2I) models impose significant computational and environmental challenges, primarily because of their scaling and training complexities. Current diffusion and transformer-based systems require extensive training on large datasets. They also require high-performance computing infrastructure, resulting in significant energy consumption and associated carbon emissions in large amounts. Empirical studies have highlighted the growing computational cost of such models, particularly in terms of prolonged training times, increased Floating-Point Operations (FLOPs), and reliance on multi-Graphics Processing Unit (GPU) environments [6, 9, 11]. To address these concerns, the concept of Green AI has gained attention. This emphasizes the need to evaluate machine learning systems in terms of performance, energy efficiency, and environmental impact. Recent studies have promoted a lifecycle perspective, considering the cumulative effects of data collection, model training, deployment, and maintenance. Within this context, monitoring tools and carbon accounting frameworks have been proposed to quantify emissions and support more sustainable AI practices [7, 9, 12]. Parameter-efficient methods, such as adapters, Low-Rank Adaptation (LoRA), and prompt-based techniques, offer promising approaches to reduce computational overhead. These methods enable model adaptation with fewer, trainable parameters. This lowers the training and inference costs while maintaining performance [4, 10, 13, 14]. However, even with these advancements, most T2I systems still rely on large pre-trained models, and efficiency improvements are often limited to the adaptation stage rather than the full model lifecycle [7, 9, 12]. The literature shows a growing awareness of sustainability challenges in T2I systems. The literature also reveals a persistent gap between performance-driven development and environmentally responsible designs. Therefore, routine accounting of emissions and the development of tools to track and minimize CO_2 across AI pipelines are needed [11, 12].

C. *Research Gap and Motivation*

Several critical gaps have been identified in the existing literature on text-to-image generation. These gaps limit a detailed understanding of the efficiency and sustainability of modern generative systems. First, the integration of efficiency-focused techniques and environmental sustainability analyses is clearly lacking. Parameter-efficient methods, such as adapters, Low-Rank Adaptation (LoRA), and prompt-based tuning, have been widely studied to reduce computational overhead; however, their implications for energy consumption and carbon emissions remain underexplored. This results in a distorted view of how technical efficiency translates into environmental benefits within T2I systems. Second, there is limited empirical evidence quantifying the environmental impact of T2I systems. Many studies have highlighted the energy demands of large-scale models, but very few have provided systematic measurements of carbon emissions or adopted a lifecycle perspective covering training, deployment, and inference. Therefore, the environmental footprint of diffusion- and

transformer-based T2I systems is still not well characterized. Third, standardized and structured comparisons of adaptation strategies are scarce. Techniques such as LoRA, adapters, and fine-tuning methods are often evaluated using different datasets, metrics, and experimental conditions. This makes it difficult to assess their relative performance in terms of controllability, image quality, and computational efficiency. These gaps highlight the need for a structured understanding of the efficiency-sustainability trade-offs in T2I systems. Addressing these gaps requires the development of standardized efficiency and sustainability metrics, shared evaluation protocols, and controlled experimental settings. This would enable a direct comparison across methods. Such efforts would also support a more integrated and evidence-based understanding of sustainable innovation in T2I.

D. Objectives of the Study

This study aims to provide a structured and integrative analysis of recent advancements in Text-to-Image (T2I) generation, with a particular focus on the relationships between model architectures, adaptation strategies, computational efficiency, and environmental sustainability. The objectives of this study are as follows:

(1) To analyze the evolution of Text-to-image (T2I) architectures, including diffusion models, transformer-based approaches, and vision-language models, and examine key trends in the model design.
(2) To examine adaptation strategies, including zero-shot and few-shot learning, as well as parameter-efficient fine-tuning methods, and evaluate their role in enabling scalable and flexible image generation.
(3) To assess the computational efficiency across different stages of the T2I pipeline, including pre-training, adaptation, and inference, with an emphasis on where efficiency improvements are most evident.
(4) To investigate the extent to which environmental sustainability is addressed in current T2I research, including the availability of empirical evidence and measurement frameworks.
(5) To develop a structured mapping between techniques, efficiency characteristics, and environmental implications, enabling a structured qualitative understanding of efficiency-sustainability trade-offs in T2I systems.

E. Contributions of the Study

This study synthesizes evidence from a dataset of 85 studies, providing a comprehensive overview of current trends and gaps in T2I research. Our study makes several key contributions to the literature on Text-to-Image (T2I) generation. First, it provides a structured synthesis of zero- and few-shot adaptation strategies. We examine how techniques such as prompt-based conditioning, adapter-based tuning, and subject-driven exemplars enable flexible image generation without requiring extensive task-specific training. By linking these approaches to cross-modal representation learning, our study clarifies how modern T2I systems achieve rapid generalization across diverse prompts and visual domains. Second, we offer a comparative analysis of the computational efficiency across different adaptation techniques. We consolidate evidence on parameter-efficient methods, including Low-Rank Adaptation (LoRA), adapters, and prompt tuning, and evaluate their impact on training cost, memory usage, and inference efficiency within diffusion- and transformer-based frameworks. Third, our review examines environmental sustainability in T2I systems. To this end, we synthesized existing research on energy consumption and carbon emissions. This highlights the importance of adopting a lifecycle perspective that considers not only model training but also deployment and inference. We also discuss the role of monitoring tools and Green AI principles in supporting responsible AI development. Finally, a key contribution of this study is the development of a structured mapping that links adaptation techniques to computational efficiency and environmental implications, enabling a systematic understanding of the trade-offs in T2I systems. Thus, this review supports a more informed and evidence-based design of scalable, resource-efficient T2I systems. Existing surveys primarily focus on model architectures and performance optimization. In contrast, this study integrates computational efficiency and environmental sustainability into a unified analytical framework, providing a perspective that has not been systematically addressed in prior reviews.

F. Positioning Against Existing Reviews

Existing review studies on Text-to-Image (T2I) generation primarily focus on the model architectures, image quality, and generative performance. While several surveys discuss diffusion models, transformer-based systems, or multimodal learning, limited attention has been paid to the combined analysis of computational efficiency, adaptation strategies, and environmental sustainability. In contrast, this study provides an integrated qualitative review that connects these dimensions using structured comparative mapping. This review further distinguishes itself by analyzing efficiency considerations across different stages of the T2I pipeline, including pre-training, adaptation, inference, and sustainability assessment, thereby offering a broader perspective on responsible and scalable generative AI development.

G. Organization of the Paper

In Section II, a comprehensive review of the foundational concepts and related work on Text-to-Image (T2I) generation is presented. This includes generative architectures, vision–language models, adaptation strategies, parameter-efficient techniques, and sustainability considerations. The materials and methods, outlining the structured review approach, data sources, selection criteria, and data extraction process, are described in Section III. Section IV presents the main results and discussion, including an analysis of the model architectures, adaptation strategies, computational efficiency, and environmental sustainability. It also provides a proposed mapping of techniques, efficiency, and environmental impact. Finally, Section V concludes

the paper by summarizing the key findings, contributions and future research directions.

II. Literature Review

A. Generative Architectures

The evolution of text-to-image generation has seen multiple architectural paradigms, with diffusion models, transformer-based approaches, and Generative Adversarial Networks (GANs) representing the primary frameworks.

Diffusion-based models are extensively used for current text-to-image generation tasks. These models work on the principle of iteratively denoising a latent representation until it becomes a clearly sharp image that matches the input prompt [2, 6, 15, 16]. The forward process gradually adds noise, and the backward process learns to recover data by denoising using a conditional U-Net backbone with cross-attention mechanisms or embedding-based conditioning (CLIP). Thus, high-fidelity images can be obtained using this iterative process. These images have a strong semantic alignment with prompts and flexible editing capabilities. Therefore, diffusion models have emerged as a leading approach for text-to-image generation, particularly in terms of photorealism and robustness [2, 6, 15]. However, this performance comes at a computational cost, as multi-step sampling and large model sizes significantly increase training and inference requirements [2, 5, 10, 11], although recent work explores efficiency improvements through guidance optimization and parameter-efficient tuning.

Another approach, called the transformer-based approach, is an alternative paradigm that frames image generation as an autoregressive sequence modeling problem. Here, images are represented as a sequence of tokens generated step-by-step in accordance with the textual prompt input [17, 18]. This enables end-to-end multimodal learning, and the text-image alignment is very strong, especially when combined with large-scale pre-training. This paradigm supports flexible and controllable image generation and allows fine-grained manipulation of attributes and scene composition [4, 17, 18]. Transformer-based models incur high computational costs owing to their reliance on sequential token generation. Therefore, research in this area has focused on improving scalability through efficient decoding strategies and parameter-efficient fine-tuning techniques [4, 17–19].

Before the advent of diffusion models, GAN-based architectures played a vital role in T2I generation. The learning process of GANs is adversarial. There are 2 neural networks called the generator that produces the images and the discriminator that attempts to distinguish those images from real samples [20]. When GANs must generate images aligned with textual inputs [21], once trained, they can produce visually sharp and realistic images with relatively fast inference [3, 13, 22]. However, GAN-based T2I models suffer from challenges such as training instability, sensitivity to hyperparameters, and difficulties in scaling to complex prompts and large datasets [3, 13]. Because of these limitations, they have gradually been replaced by diffusion-based methods in modern T2I systems.

Therefore, the evolution from GANs to diffusion and transformer-based architectures reflects a broader shift towards models that prioritize generation quality, controllability, and multimodal alignment. In addition, this progression has introduced increasing computational demands and reinforced the need to consider efficiency and sustainability in the design of next-generation T2I systems. The evolution of the above architectures in text-to-image systems from GANs to hybrid models is illustrated in Fig. 2.

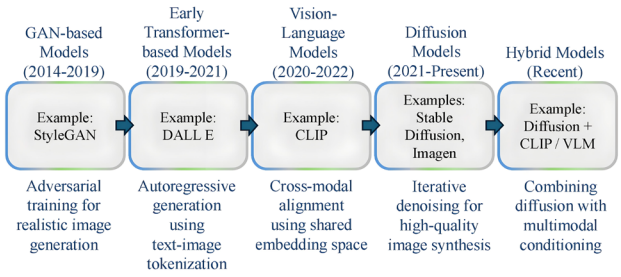

Fig. 2. Evolution of generative architectures in text-to-image systems from GANs to diffusion and hybrid models.

B. Vision-Language Models

Vision-Language Models (VLMs) are central to modern Text-to-Image (T2I) systems. They enable the alignment of textual and visual outputs through shared representation spaces. They learn joint embeddings from large-scale image-text datasets and allow semantic correspondence between language and visual content to be captured within a unified latent space. This alignment forms the foundation for tasks such as cross-modal retrieval, text-guided image generation and multimodal reasoning. Well-known VLMs, such as CLIP, Bootstrapping Language-Image Pre-training (BLIP), and Flamingo, illustrate different multimodal learning approaches. CLIP provides a powerful cross-modal embedding space that can be used to guide image generation and evaluate text-image alignment. It has become a widely adopted technique in T2I studies. BLIP adds to this paradigm by supporting both image understanding and generation tasks. This enables more flexible interactions between vision and language through unified pre-training. Flamingo further advances multimodal modeling by incorporating long context conditioning. It allows few-shot learning across diverse vision–language tasks without extensive retraining [2, 4, 6, 23]. These models contribute significantly to T2I generation by enabling flexible conditioning mechanisms, where textual prompts can directly influence the visual output through learned representations. Therefore, VLMs support zero-shot generation, style transfer, and scene-level control. They improve both the fidelity and semantic consistency of the generated images [2, 4, 6, 23]. However, they rely on large-scale pre-training and introduce substantial computational and memory requirements. In addition, challenges remain in achieving robust alignment for rare or complex concepts [2, 4, 6, 23]. Overall, VLMs serve as a critical bridge between language and image generation tasks. They provide the multimodal grounding necessary for coherent and controllable T2I systems but also raise

important considerations regarding their scalability and efficiency. The T2I architectures and their characteristics are summarized in Table I.

TABLE I. OVERVIEW OF T2I ARCHITECTURES AND CHARACTERISTICS

Architecture	Core Idea	Strength	Limitation
Diffusion	Iterative denoising	High image quality	Slow inference
Transformer	Autoregressive generation	Structured control	High compute
GAN	Adversarial training	Fast inference	Instability
VLM	Multimodal alignment	Zero-shot ability	Large models

C. Zero-Shot and Few-Shot Adaptations

Zero-shot and few-shot learning are essential mechanisms for T2I generation. These enable flexible and scalable image generation that is well aligned with textual prompts. These approaches reduce the need for extensive task-specific training by leveraging large, pre-trained multimodal models. They allow systems to generalize across diverse prompts and domains.

Zero-shot T2I generation refers to the process of generating images directly from textual prompts without any additional task-specific training. This capability is enabled by pre-trained vision-language models that learn shared embedding spaces in which textual semantics are aligned with visual representations. Therefore, zero-shot approaches support immediate image synthesis, style transfer, and prompt-driven editing for previously unseen concepts [4, 13, 24]. The main advantage of this approach is its ability to generalize broadly across domains without requiring new data or retraining [4, 24]. However, zero-shot methods also face limitations, such as sensitivity to prompt formulation, weaker alignment for rare or complex concepts, and reduced controllability compared to adapted models [4, 13, 24]. In addition, although zero-shot inference is computationally efficient, it relies on large pre-trained models, which incur substantial costs during the pre-training stage [4, 10, 11, 24].

Few-shot adaptation adds to this paradigm by introducing a small number of examples to guide the generation towards new concepts, styles, or subjects. Instead of retraining the entire model, few-shot methods typically employ parameter-efficient techniques, such as adapters, Low-Rank Adaptation (LoRA), or lightweight fine-tuning, to incorporate task-specific information [3, 13, 25]. This enables more precise control and improved fidelity in the generated outputs, particularly in personalization and domain adaptation scenarios [13, 25]. Compared to zero-shot approaches, few-shot methods achieve better alignment with user-specific or context-specific requirements while maintaining a relatively low computational overhead [3, 13]. Despite these advantages, few-shot methods introduce additional complexity, such as the need for carefully selected examples and the risk of overfitting when training data are limited. Nevertheless, by updating only a small subset of parameters, they remain significantly more efficient than full-model retraining [10, 24]. Overall, zero-shot and few-shot approaches represent complementary strategies within T2I systems: zero-shot learning enables broad generalization with minimal overhead, whereas few-shot adaptation provides targeted refinement and improved control [3, 13, 25]. Together, they highlight a shift towards efficient, flexible, and user-driven generation, while underscoring the continued dependence on large pre-trained models and their associated computational and environmental costs.

D. Parameter-Efficient Techniques

Parameter-Efficient Fine-Tuning (PEFT) has emerged as a key strategy for improving the computational efficiency of large generative models. These models work on the principle of modifying only a small subset of components, rather than updating all model parameters during adaptation. They significantly reduce the training cost while preserving performance. The most widely adopted PEFT techniques are Low-Rank Adaptation (LoRA), adapters, and prompt tuning.

LoRA introduces low-rank decomposition matrices into the existing transformer layers. This allows the model updates to be represented with a significantly reduced number of parameters. This approach enables substantial reductions in trainable weights while maintaining generation quality, making it particularly effective for adapting large-scale models without full retraining [10]. Adapters follow a similar principle by inserting lightweight bottleneck modules into the network architecture. These modules capture task-specific information and keep the original model parameters frozen, allowing for efficient reuse across multiple tasks and domains [19].

Prompt tuning offers an alternative approach by conditioning a frozen model using learnable input representations, often referred to as soft or continuous prompt. Instead of modifying internal parameters, prompt tuning adjusts the manner in which the model interprets input signals, thereby enabling flexible task adaptation with minimal computational overhead [24]. This method is particularly useful in scenarios that require rapid switching between tasks or domains.

Collectively, these techniques reduce memory usage, computational requirements, and training time, making them well-suited for large diffusion- and transformer-based systems. In many cases, PEFT methods achieve significant parameter reductions, often by several orders of magnitude, while maintaining a performance comparable to full fine-tuning [10, 19, 24]. However, these efficiency gains are not without trade-offs. PEFT methods may be sensitive to data quality, and their effectiveness can vary depending on the task complexity and domain diversity. Despite these limitations, they represent one of the most practical and scalable approaches for improving the efficiency of modern T2I pipelines [10, 19, 24].

Overall, parameter-efficient techniques play a central role in bridging the gap between the large-scale model performance and computational feasibility. They are a critical component of efficient and sustainable generative AI systems.

E. Efficiency & Sustainability in AI

Environmental sustainability in Artificial Intelligence (AI) has emerged as a critical concern. This is because of the growing computational demands of modern machine learning systems. Large-scale models require a significant amount of energy during both training and inference. This is because of the increasing parameter sizes and computational complexity. This leads to higher electricity consumption and associated carbon emissions [6, 12, 26]. Therefore, the environmental footprint of AI systems is an important area of investigation. The concept of Green AI has been introduced to address these challenges by promoting efficiency, transparency, and accountability in the development of models. Green AI frameworks emphasize the need for lifecycle-aware evaluation, considering energy consumption across the data collection, training, deployment, and inference stages [7, 12, 26]. In this context, researchers have proposed the use of standardized metrics and carbon accounting tools to quantify emissions and enable more informed decision making. Empirical studies have shown that energy usage varies significantly depending on hardware configurations, data center efficiency, and geographic factors such as electricity sources [11, 12, 27]. This variability highlights the importance of context-aware evaluation and the use of monitoring tools to track and optimize energy consumption. In addition, efficiency-oriented techniques such as parameter-efficient fine-tuning, adapters, and prompt tuning offer practical approaches to reduce computational costs [6, 12, 26]. These methods can reduce energy consumption during model adaptation and inference. However, their effectiveness depends on how they are integrated into the broader model lifecycle. Overall, the literature emphasizes the need to balance performance and environmental responsibility. This can be achieved by measuring, reporting, and minimizing the energy and carbon costs of AI systems. This perspective is particularly relevant for large-scale T2I models, where efficiency and sustainability remain challenges.

III. Materials and Methods

A. Review Approach

This study adopts a structured review approach to synthesize research on Text-to-Image (T2I) generation. The focus is on zero-shot and few-shot adaptation, computational efficiency, and environmental sustainability. Our objective is to provide a thematic and comparative analysis of key developments in generative architectures, adaptation strategies, and efficiency-oriented techniques. This review emphasizes conceptual integration and cross-study comparisons. This enables the identification of patterns, trade-offs, and gaps in the literature. Attention is paid to how different techniques, such as diffusion models, transformer-based systems, and parameter-efficient methods, relate to computational costs and environmental impact. This approach supports the development of a unified perspective linking adaptation strategies with efficiency and sustainability.

B. Data Sources

Relevant studies were identified from major academic databases and digital libraries that indexed high-quality research on artificial intelligence and machine learning. These include Scopus, Web of Science, IEEE Xplore, ACM Digital Library, ScienceDirect, and SpringerLink.

The search targeted peer-reviewed journal articles and conference papers published between 2018 and 2025, reflecting the rapid evolution of modern T2I systems, particularly diffusion-based and multimodal systems.

Search queries were constructed using combinations of keywords related to 3 main themes: text-to-image generation (e.g., diffusion models, image synthesis), adaptation strategies (e.g., zero-shot, few-shot, prompt tuning, LoRA), and efficiency and sustainability (e.g., energy consumption, Green AI, carbon footprint). A representative query used in Scopus and Web of Science was:

("text-to-image" OR "text to image" OR "image generation" OR "image synthesis" OR "diffusion model*" OR "generative model*" OR "multimodal generation")

AND

("zero-shot" OR "few-shot" OR "prompt tuning" OR "LoRA" OR "adapter*" OR "parameter-efficient" OR "fine-tuning")

AND

("computational efficiency" OR "training cost" OR "inference cost" OR "FLOPs" OR "resource utilization" OR "energy consumption" OR "carbon emission*" OR "Green AI" OR "sustainability")

C. Selection Criteria

To ensure relevance and quality, studies were selected based on predefined inclusion and exclusion criteria.

1) Inclusion criteria

- Peer-reviewed journal articles or conference papers.
- Studies published between 2018 and 2025.
- Research related to T2I generation, multimodal models, or generative AI systems.
- Studies addressing adaptation strategies (zero-shot, few-shot, or parameter-efficient methods).
- Papers discussing computational efficiency, resource usage, or environmental impact.
- While the review prioritizes peer-reviewed literature, selected high-impact preprints are also included, as they represent foundational or widely adopted methods in text-to-image generation (e.g., diffusion models and parameter-efficient techniques). These works are frequently cited and have significantly influenced subsequent peer-reviewed studies. Their inclusion ensures comprehensive coverage of the rapidly evolving developments in the field.

2) Exclusion criteria

- Non-peer-reviewed sources (e.g., preprints, blogs, editorials).
- Studies unrelated to generative or multimodal AI.
- Works focused solely on domain-specific applications without methodological relevance.

- Papers lacking sufficient detail on model design, adaptation, or evaluation.

The selection process prioritized studies that contributed to understanding the relationship between model design, adaptation strategies, and computational or environmental implications.

The screening process identified 312 initial records. After removing duplicates and screening the titles and abstracts, 146 studies remained for full-text assessment. After applying the inclusion and exclusion criteria, a final set of 85 studies were included in the review.

D. Data Extraction

The extracted data were analyzed using a 3-stage qualitative synthesis framework. First, a thematic coding process was applied to identify recurring patterns in model architectures, adaptation strategies, efficiency metrics, and environmental considerations. Second, the studies were grouped into comparative clusters based on technique categories (e.g., PEFT, full training, prompting), enabling cross-study comparison. Third, a structured mapping framework was developed to link each technique with its efficiency characteristics and environmental implications. This mapping was constructed using explicitly defined criteria, including the type of efficiency (training/inference), presence of quantitative evidence, and reported sustainability indicators. The extracted information included the following:

- Model type (e.g., diffusion, transformer, vision-language model).
- Task type (e.g., text-to-image, multimodal generation).
- Learning regime (e.g., pre-training, zero-shot, few-shot, fine-tuning).
- Adaptation technique (e.g., full training, LoRA, adapters, prompt tuning).
- Datasets and evaluation metrics.
- Key contributions and findings.
- Efficiency considerations (training, inference, or both).
- Energy or carbon impact (if reported).
- Evidence type (empirical or conceptual).
- Responsible AI aspects (e.g., sustainability, bias, ethics).

In addition, an evidence-strength classification was introduced to assess the robustness of efficiency-related claims. Studies were categorized as providing strong, moderate, weak, or no empirical evidence based on a structured assessment of the type and depth of the efficiency and environmental evaluation reported. Strong evidence refers to studies that provide quantitative empirical measurements of energy consumption (e.g., kWh), carbon emissions (e.g., CO_2e), or computational costs, supported by comparative analyses across models or techniques. However, although 30 studies provide strong empirical evidence, the reported metrics are heterogeneous and lack standardization across studies, limiting the feasibility of consistent quantitative comparisons. Moderate evidence includes studies that report quantitative proxy metrics, such as Floating Point Operations (FLOPs), parameter count, training time, or GPU usage with or without comparative analysis. Weak evidence refers to studies that include only qualitative discussions of computational costs or environmental impacts without supporting empirical data. Studies that do not address efficiency or environmental considerations are not assigned empirical evidence. This classification ensures the transparency and reproducibility of the evidence assessment.

The extracted data were analyzed through qualitative synthesis and comparative mapping, focusing on identifying the relationships between adaptation strategies, computational efficiency, and environmental implications. The overall process is illustrated in Fig. 3. This process enabled the development of a structured framework linking technical approaches to their resource and sustainability characteristics, which forms the central contribution of this paper.

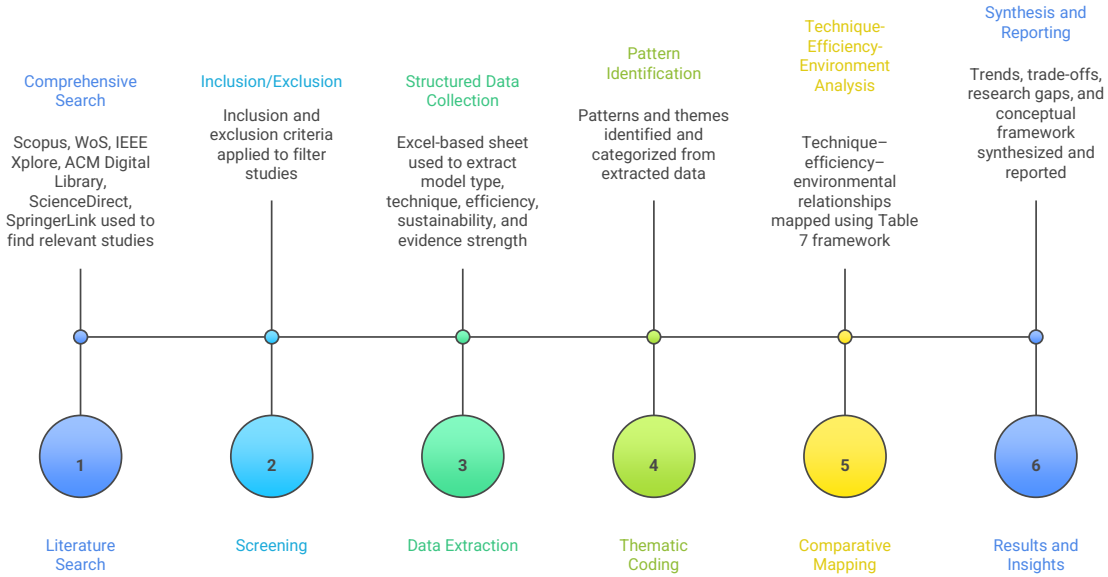

Fig. 3. Methodological workflow of the structured review.

E. Limitations of the Review Methodology

This review adopts a qualitative synthesis approach to analyze the relationships between adaptation strategies, computational efficiency, and environmental sustainability in Text-to-Image (T2I) systems. A quantitative meta-analysis was not feasible because the reviewed studies reported efficiency and environmental metrics in highly heterogeneous ways. Metrics such as energy consumption, carbon emissions, FLOPs, GPU usage, and training time were measured under different experimental settings, hardware configurations, datasets, and evaluation protocols. In many cases, studies provide only partial or qualitative discussions of efficiency, without standardized empirical reporting. Consequently, the comparative mapping developed in this study is conceptual and qualitative rather than based on directly comparable numerical benchmarks. Despite this limitation, the structured synthesis enables the identification of recurring trends, trade-offs, and research gaps across the literature while also highlighting the need for standardized sustainability evaluation frameworks in future T2I research.

IV. Results and Discussion

A. Model Architectures and Trends

The analysis of 85 selected studies highlights the diverse landscape of model architectures in Text-to-Image (T2I) systems. This shows that transformer-based and hybrid vision–language models constitute a substantial proportion of the reviewed literature. This also reflects their flexibility in handling multimodal data and enabling controllable generation. Diffusion models, although fewer in number, play a crucial role in advancing the field in terms of performance, adoption, and influence in modern Text-to-Image (T2I) systems. These models produce high-quality images with strong semantic alignment with textual prompts [28]. This has led to their recognition as a leading approach in recent T2I developments. Their learning process involves gradually adding noise to an image, starting from Gaussian noise and iteratively refining it into a clean image [3, 28–31]. This mechanism enables controlled and refined image synthesis, allowing for the effective handling of complex scenes and detailed textual prompts.

Compared with diffusion models, GANs are more efficient at inference but suffer from training instability and limited scalability [29].

Of the total dataset of 85 studies, 54 focused on model development, while the remaining 31 comprised non-model contributions, including survey papers and tools/frameworks. Among the model-based studies, transformer-based approaches represented the largest group (21 studies), followed by diffusion models (15 studies) and vision–language models (13 studies), with a smaller number of GAN-based (3 studies) and other architectures (2 studies). This distribution suggests that while transformer-based methods continue to be widely explored because of their flexibility and multimodal capabilities, diffusion models and vision-language systems are also gaining strong traction in advancing image quality and alignment. In contrast, the substantial presence of survey papers (20 studies) and tools (11 studies) indicate a growing interest in evaluation, benchmarking, and sustainability considerations, reflecting a maturing research landscape that extends beyond core model development.

Transformer-based approaches effectively handle multimodal representation learning and autoregressive generation [48]. However, owing to their sequential nature, these models introduce significant computational overheads, particularly for high-resolution outputs. VLMs, such as CLIP, BLIP and Flamingo, are increasingly being integrated into T2I systems. These models enhance cross-modal alignment and enable zero- and few-shot capabilities [28], often serving as conditioning mechanisms rather than standalone generators. The distribution of the model types and their associated characteristics are summarized in Table II. As shown in Fig. 4, transformer-based approaches constitute the largest proportion of model studies, whereas diffusion and vision-language models also represent a significant share, indicating their growing importance in modern T2I systems. Large-scale datasets, such as Common Objects in Context (COCO) and Large-scale Artificial Intelligence Open Network (LAION), contribute significantly to the computational cost owing to their size and training requirements. However, their environmental impact is rarely explicitly measured, indicating a gap in sustainability evaluation.

TABLE II. Model Type vs Efficiency Consideration

Study Category	Model Type	No. of Studies	Efficiency Considered?	Typical Observation	Studies
Model Studies	Transformer-based Models	21	Rarely	High computational cost with limited optimization	[10, 17, 19, 24, 31–47]
	Diffusion Models	15	Rarely	Focus on image quality rather than efficiency	[2–4, 13, 25, 27–30, 48–53]
	Vision-Language Models	13	Partially	Efficiency via pre-trained representations	[5, 54–65]
	GAN-based Models	3	Rarely	Faster inference but limited scalability	[66–68]
	Other Architectures	2	Limited	Varying efficiency characteristics	[15, 69]
Non-Model Studies	Tools/Frameworks	11	Strongly	Designed for efficiency measurement and tracking	[11, 12, 18, 70–77]
	Survey Papers	20	Not Applicable	Provide conceptual analysis rather than implementation	[6, 7, 9, 14, 26, 78–92]

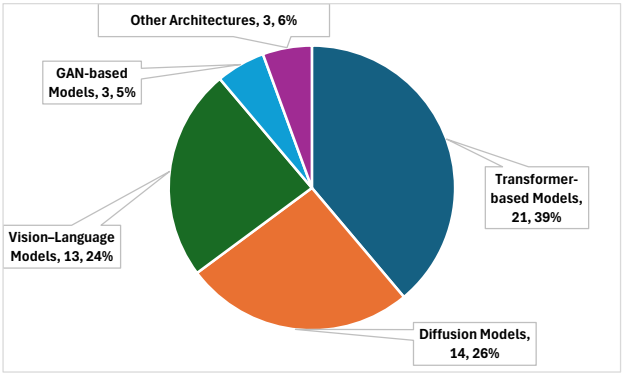

Fig. 4. Distribution of model-based studies across different T2I architectures, highlighting the prevalence of transformer-based, diffusion, and vision-language models.

B. Adaptation Strategies

Adaptation strategies in Text-to-Image (T2I) systems are evolving towards more flexible and computationally efficient approaches. These strategies can be understood from 2 complementary perspectives: learning regimes and adaptation techniques.

1) Learning regimes

Learning regimes define how models are adapted to new tasks.

Zero-shot methods enable direct image generation from textual prompts without requiring additional training [54]. They rely on pre-trained multimodal representation. These models offer rapid deployment and broad generalization [54]. However, they may struggle with rare or highly specific concepts.

Few-shot methods introduce a small number of examples to guide generation, leading to improved fidelity and personalization [41, 63]. They offer enhanced control compared to zero-shot methods, with:

- limited additional computational cost [90].
- Fine-tuning and pre-training involve more extensive updates and are computationally intensive.

These regimes reflect a clear spectrum of computational costs, from zero-shot (minimal) to full training (maximum). As shown in Table III, fine-tuning is the most widely used learning regime (24 studies), followed by pre-training (20 studies) and few-shot approaches (14 studies), while zero-shot methods (8 studies) are less frequently adopted, reflecting a preference for strategies that balance adaptability and performance. The progression of learning regimes in T2I systems is illustrated in Fig. 5, showing a shift from zero-shot approaches with minimal computational cost to pre-training methods that require substantial computational resources.

TABLE III. DISTRIBUTION OF LEARNING REGIMES IN T2I STUDIES

Learning Regime	No. of Studies
Zero-shot	8
Few-shot	14
Fine-tuning	24
Pre-training	20

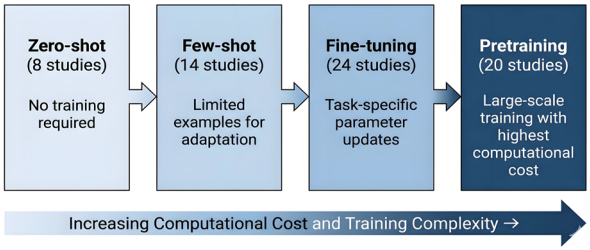

Fig. 5. Learning regime spectrum in T2I systems, illustrating the progression from zero-shot to pre-training based on increasing computational cost and adaptation complexity.

2) Adaptation techniques

From a technical perspective, adaptation strategies emphasize efficiency and scalability. The distribution of techniques used across the studies is summarized in Table IV. The distribution of adaptation techniques across the reviewed studies is illustrated in Fig. 6, where full training emerges as the most used approach, whereas parameter-efficient methods, including adapters, low-rank adaptation, and prompt-based techniques, are increasingly adopted to reduce computational overhead.

TABLE IV. ADAPTATION TECHNIQUE DISTRIBUTION

Technique Type	No. of Studies	Description
Full Training	23	High computational cost, full parameter updates
PEFT-Adapters	12	Lightweight modular updates
PEFT-LoRA	3	Low-rank parameter updates
Prompt-based Methods	12	Input-level conditioning
Hybrid Methods	2	Combined techniques

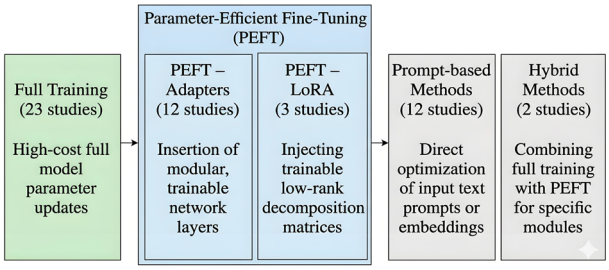

Fig. 6. Distribution of adaptation techniques in T2I systems, highlighting the prevalence of full training and the growing adoption of parameter-efficient methods.

Parameter-Efficient Fine-Tuning (PEFT) methods, including LoRA, adapters, and prompt tuning, represent a significant advancement in model adaptation [13, 41, 64]. These techniques reduce the training costs by updating only a small subset of the model parameters. This enables the scalable adaptation of large models [41].

LoRA is inspired by rank decomposition in Natural Language Processing (NLP) [13]. Adapters can be trained under various conditions to achieve rich control and editing effects in the outputs. They offer composability and generalization [13]. They can also improve the results in few-shot settings. Adapter modules are an alternative to full fine-tuning, especially when dealing with many downstream tasks, because they are parameter-efficient [38]. Different adapter-based methods

include Adapter, Hyperformer, and Compacter [64]. Prompt tuning involves adding trainable prompts to the input [24, 64].

Pre-training prompt-tuning parameters can improve performance with limited labeled data [41]. However, prompt tuning can be tricky, potentially leading to decreased generalization or underperformance compared with zero-shot baselines as training progresses [65].

PEFT methods introduce learnable adapting modules that are plugged into frozen backbone models [33]. Although full fine-tuning of pre-trained models is often effective, it requires many annotated data points and is resource-intensive [24, 39, 54]. PEFT methods address these challenges [49].

Overall, adaptation strategies reflect a shift towards efficiency-driven customization, although their effectiveness remains dependent on the capabilities of large pre-trained models.

C. Efficiency Analysis

The analysis indicates that the computational efficiency of T2I systems varies significantly throughout the model's lifecycle. The pre-training stage is the most resource-intensive. It contributes to the majority of the computational cost, as it demands large-scale datasets, extensive GPU usage, and prolonged training times [48]. Adaptation strategies, such as Parameter-Efficient Fine-Tuning (PEFT) methods, significantly reduce computational overhead [49] during the adaptation phase. These methods aim to develop parameter-efficient adaptation techniques with trainable parameters that are significantly reduced (often by approximately 90–99%) compared to the full model parameters [49]. However, these efficiency gains do not extend to the pre-training stage, which continues to dominate the overall computational and environmental costs. Zero-shot approaches can eliminate training costs entirely at the adaptation stage, whereas few-shot methods introduce minimal additional computation [33]. Adapter models are low-cost. They achieve their purpose by learning the alignment between external control signals and the internal knowledge of pre-trained T2I models, even for larger diffusion models such as Stable Diffusion XL (SDXL) [13].

Despite improvements in other areas, inference efficiency remains challenging. Here, GANs prove to be better than diffusion models. Diffusion models require multiple denoising steps [93] (approximately 10–50 steps). This leads to slower inference compared to GAN-based models [2]. For example, some unoptimized models can take 15 s to sample a single image, which is considerably slower than GAN methods that produce images in a single forward pass (in milliseconds) [2]. Recent optimizations have improved the sampling speed; however, high-quality generation still involves substantial computational effort. However, some latent diffusion models have shown that they can significantly decrease inference costs compared to pixel-based diffusion approaches [48]. The distribution of efficiency considerations across studies is presented in Table V.

TABLE V. EFFICIENCY CONSIDERATION ACROSS STUDIES

Efficiency Considered	Number of Studies	Percentage (%)	Observation
Yes	52	61.2%	Explicit efficiency focus
No	31	36.5%	Performance-focused studies
Implicit	2	2.4%	Limited or indirect consideration

These findings suggest that efficiency improvements primarily lie in adaptation strategies rather than in core model architectures. Techniques such as minimizing the model size can lower the computational requirements and energy consumption. However, smaller models often sacrifice some accuracy compared to their larger counterparts [78]. A drawback of earlier diffusion models, the higher inference cost, is being progressively mitigated by more efficient architectures and scheduling methods [51].

D. Environmental Sustainability

Environmental sustainability is an emerging concern in T2I research. Many studies acknowledge the energy consumption and carbon emissions associated with large-scale models; however, empirical evidence regarding these impacts is limited [12, 86, 89]. The distribution of evidence strengths across the studies is summarized in Table VI.

TABLE VI. EVIDENCE STRENGTH DISTRIBUTION ACROSS STUDIES

Evidence Strength	Number of Studies	Percentage (%)	Typical Paper Types
Strong	30	35.3	PEFT methods, monitoring tools
Moderate	20	23.5	Vision-language models, efficiency studies
Weak	8	9.4	Core T2I models
No Empirical Evidence	27	31.8	Frameworks, performance-focused studies

The results reveal a polarized trend, where a significant number of studies provide strong empirical evidence, while a comparable proportion does not. Strong empirical evidence is primarily associated with parameter-efficient methods and monitoring tools (30 studies), whereas core T2I models often lack detailed environmental evaluation, with a substantial proportion of studies falling into the weak (8) or no evidence (27) categories.

This imbalance highlights a clear gap between efficiency-focused optimization and rigorous environmental evaluation in current T2I research efforts. Tools such as "CarbonTracker", "eco2AI", and "machine learning emissions calculators" are available to track carbon emissions. However, greater incentives are needed for researchers and industry developers to measure and report these findings [89]. Table VII summarizes the representative tools identified in the reviewed studies for tracking energy consumption and carbon emissions during model development and experimentation.

TABLE VII. PRACTICAL TOOLS FOR EFFICIENCY AND SUSTAINABILITY ASSESSMENT

Tool	Purpose	Metrics Reported	Use in T2I/AI Pipeline	Limitation
CarbonTracker [70]	Tracks and predicts the carbon footprint of deep learning training	Energy consumption, CO_2 emissions, training duration	Monitoring training and fine-tuning stages of T2I models	Requires hardware and regional energy information
eco2AI [12]	Tracks environmental impact of machine learning experiments	Carbon emissions, energy usage, hardware utilization	Sustainability monitoring during model experimentation and adaptation	Accuracy depends on system-level monitoring availability
ML CO_2/Emissions Calculator [11]	Estimates carbon emissions associated with machine learning workloads	CO_2 emissions, electricity usage	Estimating environmental impact of model training and deployment	Provides estimated rather than directly measured values

These tools improve the transparency of environmental reporting by enabling researchers to estimate and monitor their computational impact. However, they primarily support measurements and assessments rather than directly reducing computational costs or energy consumption.

Green AI frameworks advocate for lifecycle-based evaluation and carbon tracking [7, 80, 92]. However, such practices have not yet been widely adopted in T2I research. Consequently, sustainability considerations often remain secondary to performance optimization [7, 89]. The environmental impact of AI training, including carbon emissions, is a crucial aspect of AI sustainability that must be explicitly accounted for and evaluated [12, 89].

Therefore, awareness of the impact of AI is growing, and tools for measurement exist. However, comprehensive empirical evaluations and widespread adoption of Green AI practices are still lacking in T2I research, with a continued focus on performance over sustainability. Although several studies have acknowledged the energy consumption and environmental impact of T2I systems, quantitative reporting remains inconsistent and fragmented in the literature. Metrics such as energy usage, carbon emissions, and training cost are not uniformly reported. This limits the ability to perform direct cross-study comparisons of the results. As a result, this review adopts a qualitative synthesis approach while highlighting the need for standardized and reproducible environmental reporting practices

E. Technique → Efficiency → Environmental Mapping

A central contribution of this study is the mapping of adaptation techniques, computational efficiency, and environmental implications. This mapping provides a qualitative framework for understanding the trade-offs across T2I systems. It synthesizes the reported efficiency characteristics and environmental insights from the reviewed studies. Although the study aims to enable a comparative analysis, quantitative comparisons are constrained by the limited availability of consistent empirical data in the literature. To address this, future studies should prioritize the adoption of standardized metrics for energy consumption and carbon emissions, coupled with open reporting practices. Table VIII presents these relationships conceptually, rather than through standardized numerical metrics.

TABLE VIII. TECHNIQUE-EFFICIENCY-ENVIRONMENTAL MAPPING

Approach Category	Example Methods	Efficiency Type	Evidence Strength	Environmental Implication
Full Model Training	Diffusion, CogView	Not optimized	Weak	High computational cost (multi-GPU training, extended training duration, high energy consumption)
PEFT Methods	LoRA, Adapters	Training + Inference	Strong	Reduced energy consumption due to parameter-efficient updates; overall impact is constrained by the high cost of pre-training large models
Prompting	CLIP, DALL-E	Inference	Moderate	Lower computational cost during adaptation
Vision–Language Models	Flamingo, BLIP	Both	Moderate	Moderate environmental cost due to reliance on large pre-trained models
Monitoring Tools	CarbonTracker, eco2AI, ML CO_2 Calculator	Both	Strong	Enables measurement of energy consumption and carbon emissions
Evaluation Frameworks	CLIPScore	None	None	No direct impact
Green AI Studies	Sustainability surveys	Conceptual	Moderate	Highlight environmental concerns
Dataset/ Benchmark	COCO, LAION	Not Applicable	None	Indirect Impact via dataset scale and training cost

Note: Environmental implications are described qualitatively owing to the limited availability of standardized quantitative metrics across the reviewed studies.

Full-model training approaches are associated with the highest computational cost and environmental impact [6, 32, 80]. Training models can incur substantial environmental costs because of the energy required for hardware over extended periods [6]. For example, when GPT-3 was accessed 590 million times in January 2023, it consumed energy equivalent to that of 175,000 people [80].

In contrast, Parameter-Efficient Fine-Tuning (PEFT) methods significantly reduce training requirements by updating only a small subset of parameters [41]. PEFT methods can match the performance of full fine-tuning while updating as little as 0.01% of the model

parameters [41]. This approach drastically lowers the memory and storage requirements for training and saving models [41]. PEFT offers a promising alternative during the adaptation phase by reducing computational and storage requirements [41]. However, it is important to note that these efficiency gains are contingent on the availability of large pre-trained models, the training of which involves substantial computational and environmental costs.

Zero-shot and few-shot approaches further reduce the adaptation cost, although they remain dependent on pre-trained models [41]. PEFT methods have shown favorable performance in few-shot fine-tuning settings [41].

Monitoring tools support sustainability by enabling the measurement of energy consumption, although they do not directly reduce the computational cost.

This mapping demonstrates that efficiency gains are primarily concentrated in adaptation strategies, particularly PEFT methods, rather than in core model architectures. The mapping is shown in Fig. 7. Full model training remains computationally intensive and environmentally costly [6, 32, 80], whereas PEFT techniques significantly reduce resource requirements by updating only a small fraction of the parameters [41].

Therefore, while full model training is computationally and environmentally intensive, PEFT methods offer a more sustainable and efficient alternative. They focus on updating only a fraction of the parameters. Efficiency improvements have been largely observed in these adaptation strategies.

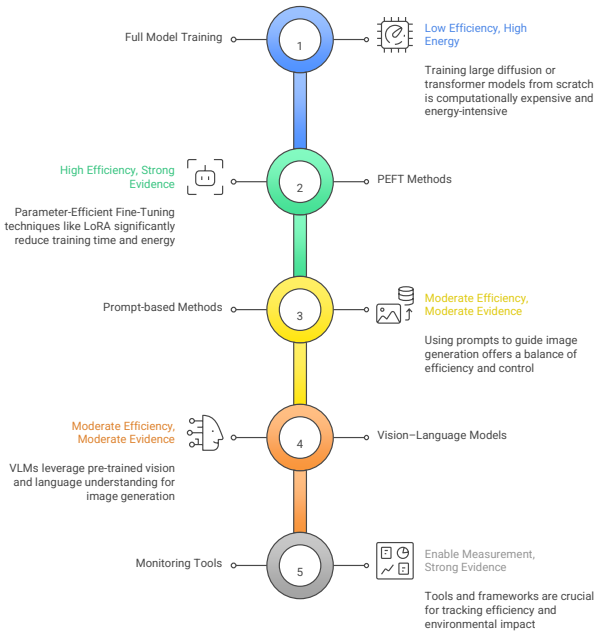

Fig. 7. Technique-efficiency-environmental mapping illustrating the relationship between adaptation strategies, computational efficiency, evidence strength, and environmental implications of T2I systems.

To further illustrate the relationship between model design choices and sustainability implications, Table IX compares the major technology pathways in T2I systems based on their computational cost drivers, efficiency benefits, and environmental concerns.

TABLE IX. TECHNOLOGY PATH VS. ENVIRONMENTAL COST COMPARISON

Technology Path	Main Cost Driver	Efficiency Benefit	Environmental Cost Level	Sustainability Concern
Full model training	Large-scale datasets, multi-GPU training, long training duration	High model capacity and strong generation quality	High	Substantial energy use and carbon emissions during training
Diffusion-based generation	Iterative denoising steps and large model size	High image quality and semantic alignment	High to Moderate	Slow inference and high compute demand, especially in unoptimized models
Transformer-based T2I	Sequential token generation and large-scale pre-training	Strong multimodal representation and controllable generation	High	High memory use, long training time, and scalability challenges
Vision–Language Models	Large pre-trained image–text models and embedding alignment	Efficient transfer through pre-trained representations	Moderate to High	Environmental cost is shifted to large-scale pre-training
PEFT methods	Lightweight adaptation on frozen pre-trained models	Reduced trainable parameters and lower adaptation cost	Low to Moderate	Benefits are mainly limited to adaptation; pre-trained backbone cost remains
Prompt-based methods	Prompt design and inference-time conditioning	Low-cost adaptation without full retraining	Low to Moderate	Still depends on large pre-trained models for generation
Monitoring tools	Measurement and tracking infrastructure	Supports energy and carbon reporting	Low	Enables measurement but does not directly reduce compute
Dataset/benchmark scaling	Large image–text datasets and repeated training cycles	Improves model generalization and evaluation coverage	Moderate to High	Dataset size increases storage, preprocessing, and training energy demand

The comparison highlights that although techniques such as PEFT and prompting improve adaptation efficiency, the environmental burden associated with large-scale pre-training and dataset scaling remains a significant challenge across modern T2I systems.

F. Ethical Considerations: Misuse, Fairness and Environmental Justice

The rapid advancement of Text-to-Image (T2I) systems has introduced several ethical concerns, alongside improvements in generative performance [49].

One major issue is the potential misuse of these systems to generate deceptive or harmful synthetic content [78]. High-quality generated images can be used to create misinformation, manipulated media, deepfakes, and misleading visual narratives [25, 28]. As T2I models become more photorealistic and widely accessible, the risks associated with unauthorized content generation, identity manipulation, and malicious use increase [4]. These concerns highlight the need for stronger safeguards, content moderation mechanisms and responsible deployment practices.

Fairness and bias remain important challenges in T2I systems [25, 78]. Many generative models are trained on large-scale Internet datasets that may contain demographic, cultural, or social biases [25, 28, 29, 31]. Consequently, the generated outputs can unintentionally reinforce stereotypes, underrepresent certain groups, or produce biased visual associations [25, 29]. Bias can also emerge through prompt interpretation and dataset imbalance, affecting the inclusiveness and reliability of generated content [4, 31]. Addressing these issues requires diverse datasets, fairness-aware evaluation frameworks, and bias mitigation strategies during model development and deployment [25, 29].

In addition to the ethical concerns related to content generation, environmental justice has emerged as an important issue in large-scale AI systems. Training and deploying computationally intensive T2I models require substantial energy resources and large-scale data center infrastructure to process the data. The environmental burden associated with energy consumption and carbon emissions is often unevenly distributed across regions and communities. This raises broader concerns regarding equitable access to computational resources, sustainable infrastructure, and the societal impact of large-scale AI development. Therefore, future research should integrate ethical responsibility, fairness, and environmental sustainability into a unified framework for the responsible development of generative AI.

Therefore, the rapid evolution of T2I systems brings significant generative capabilities, as well as critical ethical challenges related to content misuse, pervasive biases in generated outputs, and the environmental impact of large-scale AI. Addressing these issues requires a multifaceted approach encompassing responsible deployment, bias mitigation strategies, and a focus on sustainable development.

G. Key Insights

Three primary insights emerge from an analysis of Text-to-Image systems, focusing on their computational efficiency, performance prioritization, and existing literature gaps.

First, the computational efficiency of T2I systems is largely driven by adaptation techniques. Methods such as LoRA, adapters, and prompt tuning offer strong evidence for their role in reducing training costs [9].

Second, a significant observation is that most T2I models prioritize performance over sustainability. Advances in image quality have often been achieved with limited consideration of environmental impact [32, 80, 89]. The concept of 'Green AI' advocates for promoting efficiency as an evaluation metric, noting that many studies prioritize accuracy improvements without focusing on efficiency [9, 32]. There is a growing call to assess the carbon footprint and environmental impact of AI model training and tuning [89].

Third, a clear gap exists in the current literature regarding the systematic evaluation of efficiency and sustainability of T2I systems. Although these topics are frequently discussed, few studies have provided empirical evaluations or standardized metrics [92].

Research on AI and machine learning for production efficiency and sustainable development has historically been limited, although it is becoming increasingly interdisciplinary [79]. There is a critical need for standardized evaluation practices, particularly for carbon metrics, to fully realize AI's potential of AI across various domains [92].

TABLE X. RESEARCH GAP-FUTURE DIRECTION ROADMAP

Research Gap	Current Limitation	Future Direction	Expected Impact
Limited standardized sustainability metrics	Energy consumption, carbon emissions, FLOPs, and GPU usage are reported inconsistently across studies	Develop standardized reporting protocols for energy use, carbon emissions, hardware settings, and dataset scale	Enables fair comparison of T2I models and improves reproducibility
Sustainability evaluation mostly absents in core T2I models	Many diffusion and transformer-based models prioritize image quality over environmental reporting	Integrate carbon and energy reporting into model evaluation benchmarks	Encourages environmentally responsible model development
Efficiency gains concentrated mainly in adaptation stage	PEFT methods reduce adaptation cost but do not address the high cost of pretraining large models	Develop lifecycle-aware optimization covering pretraining, adaptation, inference, and deployment	Reduces overall computational and environmental burden
Limited dataset-level environmental analysis	Dataset size, preprocessing, storage, and repeated training costs are rarely evaluated	Include dataset scale, storage cost, and data curation energy in sustainability assessments	Improves understanding of dataset-driven environmental impact
Lack of integration between monitoring tools and T2I workflows	Tools such as CarbonTracker and eco2AI are not widely embedded in T2I experimentation	Incorporate monitoring tools into standard T2I training and inference pipelines	Improves transparency and routine reporting of energy and emissions
Inference efficiency remains challenging for diffusion models	Multi-step denoising increases inference time and computational cost	Develop faster sampling methods, distillation, and efficient inference pipelines	Supports scalable deployment with lower energy consumption
Ethical and environmental impacts are treated separately	Bias, misuse, and environmental justice are often discussed independently	Develop responsible AI frameworks that combine fairness, misuse prevention, and sustainability	Supports safer, fairer, and more sustainable generative AI systems

These findings highlight the necessity of integrating performance, efficiency, and sustainability into a unified framework for the future development of T2I technologies. Therefore, based on the reviewed literature and key insights identified in the previous section, Table X presents a roadmap that links major research gaps with future directions and their expected impact on the development of more efficient and sustainable T2I systems.

V. Conclusion and Future Research Directions

This study presents a structured review of Text-to-Image (T2I) generation. It focuses on the relationships between generative architectures, adaptation strategies, computational efficiency, and environmental sustainability. The findings show that diffusion-based models have emerged as a leading approach in terms of performance and recent advancements, although transformer-based and hybrid architectures are widely used in the literature. These are often supported by vision-language models for enhanced multimodal conditioning.

Simultaneously, techniques such as zero-shot and few-shot, often known as adaptation strategies, along with parameter-efficient fine-tuning methods, have enabled more flexible and scalable deployment of these systems. The key contribution of this study is the development of a unified mapping between the techniques, efficiency characteristics, and environmental implications. This mapping highlights that significant progress has been made in improving the adaptation efficiency. This is particularly achieved through methods such as LoRA, adapters, and prompt tuning.

However, the overall computational burden of T2I systems remains high because of their reliance on large pre-trained models.

The findings demonstrate that efficiency improvements are mostly concentrated in the adaptation stage, with limited attention given to optimizing the full model lifecycle. In addition, an evidence gap is found in the literature: the environmental impact of large-scale AI systems is increasingly recognized, but empirical studies that quantify energy consumption and carbon emissions in T2I models are very limited. This is another important contribution of this study. In addition, standardized evaluation frameworks for measuring efficiency and sustainability are largely absent. This makes it difficult to compare models or assess the trade-offs between performance and environmental impact.

The results also highlight a broader trend in T2I research, where performance optimization continues to take precedence over efficiency and sustainability considerations. With the rapid advancements in image quality and controllability, computational demands and environmental concerns are also increasing. Addressing this imbalance requires a shift towards more holistic model design and evaluation practices.

Future research should focus on integrating efficiency and sustainability directly into the model architecture. These should not be treated as secondary considerations. This integration includes the development of energy-aware training methods, optimization of inference pipelines, and adoption of standardized metrics for evaluating environmental impact. In addition, greater emphasis should be placed on combining parameter-efficient techniques with lifecycle-aware monitoring tools. This will enable a more sustainable deployment of generative models.

In conclusion, this study provides a comprehensive and structured perspective on the current state of T2I generation research. This emphasizes the need to balance performance, efficiency, and sustainability. By bridging these dimensions, the paper contributes to the development of more responsible and scalable generative AI systems.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Subuhi K. Ansari conceived the study, designed the review methodology, conducted the literature search and screening, extracted and analyzed the data, developed the tables and figures, interpreted the findings, and drafted the manuscript. Fatma A. Alqahtani contributed to literature screening, study selection, and validation of the extracted data. Chamandeep Kaur assisted with data extraction, data verification, and review of the analytical results. Wafaa Abu Shamlah contributed to data interpretation, critical review of the findings, and manuscript revision. Najla Mohammed Galam assisted with manuscript formatting, quality checking, and editorial revisions. Nedaa Abduaziz Hadi supervised the study, provided methodological guidance, critically reviewed the manuscript, and contributed to the overall refinement of the work. All authors reviewed, approved, and agreed to the published version of the manuscript.

Acknowledgment

The authors wish to thank all the reviewers for their valuable comments and suggestions, which helped improve the quality and clarity of this manuscript.

References

[1] R. Bayoumi, M. Alfonse, and A. B. M. Salem, "Text-to-image synthesis: A comparative study," in *Proc. Digital Transformation Technology*, 2021, pp. 229–251.

[2] A. Nichol, P. Dhariwal, A. Ramesh *et al.*, "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," arXiv preprint, arXiv: 2112.10741, 2021

[3] N. Ruiz, Y. Li, V. Jampani *et al.*, "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 22500–22510.

[4] A. Ramesh, P. Dhariwal, A. Nichol *et al.*, "Hierarchical text-conditional image generation with CLIP latents," arXiv preprint, arXiv: 2204.06125, 2022.

[5] L. Yuan, D. Chen, Y. L. Chen *et al.*, "Florence: A new foundation model for computer vision," arXiv preprint, arXiv: 2111.11432 2021.

[6] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in *Proc. 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3645–3650.

[7] R. Verdecchia, J. Sallou, and L. Cruz, "A systematic review of Green AI," *WIREs Data Min & Knowl*, vol. 13, no. 4, e1507, 2023.

[8] T. Wang, K. Wang, H. Cai *et al.*, "APQ: Joint search for network architecture, pruning and quantization policy," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2075–2084.

[9] R. Schwartz, J. Dodge, N. A. Smith *et al.*, "Green AI," *Commun. ACM*, vol. 63, no. 12, pp. 54–63, 2020.

[10] E. J. Hu, Y. Shen, P. Wallis *et al.*, "LoRA: Low-rank adaptation of large language models," arXiv preprint, arXiv: 2106.09685, 2021.

[11] A. Lacoste, A. Luccioni, V. Schmidt *et al.*, "Quantifying the carbon emissions of machine learning," arXiv preprint, arXiv: 1910.09700, 2019.
[12] S. A. Budennyy, V. D. Lazarev, N. N. Zakharenko *et al.*, "Eco2ai: Carbon emissions tracking of machine learning models as the first step towards sustainable AI," *Dokl. Math.*, vol. 106, pp. S118–S128, 2022.
[13] C. Mou, X. Wang, L. Xie *et al.*, "T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," in *Proc. AAAI Conf. on Artificial Intelligence*, 2024, pp. 4296–4304.
[14] A. Parnami and M. Lee, "Learning from few examples: A summary of approaches to few-shot learning," arXiv preprint, arXiv: 2203.04291, 2022.
[15] C. Li, Z. Ding, D. Zhao *et al.*, "Building energy consumption prediction: An extreme deep learning approach," *Energies*, vol. 10, no. 10, 1525, 2017.
[16] S. K. Ansari and R. Kumar, "Design of augmented diffusion model for text-to-image representation," *IJARCS*, vol. 16, no. 1, pp. 55–60, 2025.
[17] O. Gafni, A. Polyak, O. Ashual *et al.*, "Make-A-Scene: Scene-based text-to-image generation with human priors," in *Proc. Computer Vision-ECCV 2022*, 2022, pp. 89–106.
[18] J. Hessel, A. Holtzman, M. Forbes *et al.*, "CLIPScore: A reference-free evaluation metric for image captioning," in *Proc. 2021 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2021, pp. 7514–7528.
[19] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Proc. 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conf. on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 4582–4597.
[20] S. K. Ansari and R. Kumar, "Design of augmented diffusion model for text-to-image representation using hybrid GAN," in *Proc. International Workshop on Cultural Perspectives of Human-Centered and Technological Innovations*, 2024, pp. 166–176.
[21] Subuhi, K. Ansari, and R. Kumar, "C2I-generative adversarial networks for image captioning and generation: A novel approach to text-image synthesis," *International Journal of Scientific Development and Research*, vol. 10, no. 3, pp. c529–c537, 2025.
[22] S. K. Ansari, M. A. Khammash, A. Appukuttan *et al.*, "Mitigating mode collapse to improve diversity in text-to-image Gan outputs: Strategies in architectural design, training methodologies, and evaluation techniques," *Journal of Theoretical and Applied Information Technology*, vol. 104, no. 6, pp. 14–32, 2026.
[23] C. Chen, S. Dong, Y. Tian *et al.*, "Temporal self-ensembling teacher for semi-supervised object detection," *IEEE Transactions on Multimedia*, vol. 24, pp. 3679–3692, 2022.
[24] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proc. 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 3045–3059.
[25] M. Cheong, E. Abedin, M. Ferreira *et al.*, "Investigating gender and racial biases in DALL-E mini images," *ACM J. Responsib. Comput.*, vol. 1, no. 2, 13, 2024.
[26] E. Triantafillou, T. Zhu, V. Dumoulin *et al.*, "Meta-dataset: A dataset of datasets for learning to learn from few examples," arXiv preprint, arXiv: 1903.03096, 2020.
[27] L. Theis, A. v. d. Oord, and M. Bethge, "A note on the evaluation of generative models," arXiv preprint, arXiv: 1511.01844, 2015.
[28] Y. Balaji, S. Nah, X. Huang *et al.*, "Text-to-image diffusion models with an ensemble of expert denoisers," arXiv preprint, arXiv: 2211.01324, 2023.
[29] C. Saharia, W. Chan, S. Saxena *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in Neural Information Processing Systems*, vol., 35, pp. 36479–36494, 2022.
[30] J. Cho, A. Zala, and M. Bansal, "DALL-eval: Probing the reasoning skills and social biases of text-to-image generation models," in *Proc. EEE/CVF International Conf. on Computer Vision (ICCV)*, 2023, pp. 3043–3054.
[31] J. Yu, Y. Xu, J. Y. Koh *et al.*, "Scaling autoregressive models for content-rich text-to-image generation," arXiv preprint, arXiv: 2206.10789, 2022.
[32] R. Rombach, A. Blattmann, D. Lorenz *et al.*, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10684–10695.
[33] E. M. Bender, T. Gebru, A. McMillan-Major *et al.*, "On the dangers of stochastic parrots: Can language models be too big?" in *Proc. 2021 ACM Conf. on Fairness, Accountability, and Transparency*, 2021, pp. 610–623.
[34] H. Chen, R. Tao, H. Zhang *et al.*, "Conv-adapter: Exploring parameter efficient transfer learning for ConvNets," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024, pp. 1551–1561.
[35] T. Dettmers, A. Pagnoni, A. Holtzman *et al.*, "QLoRA: Efficient finetuning of quantized LLMs," *Advances in Neural Information Processing Systems*, vol. 36, pp. 10088–10115, 2023.
[36] M. Ding, Z. Yang, W. Hong *et al.*, "CogView: Mastering text-to-image generation via transformers," *Advances in Neural Information Processing Systems*, vol. 34, 19822–19835, 2021.
[37] M. Ding, W. Zheng, W. Hong *et al.*, "CogView2: Faster and better text-to-image generation via hierarchical transformers," *Advances in Neural Information Processing Systems*, vol. 35, 16890–16902, 2022.
[38] Y. Gu, X. Han, Z. Liu *et al.*, "PPT: Pre-trained prompt tuning for few-shot learning," in *Proc. 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 8410–8423.
[39] N. Houlsby, A. Giurgiu, S. Jastrzebski *et al.*, "Parameter-efficient transfer learning for NLP," in *Proc. 36th International Conf. on Machine Learning*, 2019.
[40] M. Jia, L. Tang, B. C. Chen *et al.*, "Visual prompt tuning," in *Proc. Computer Vision-ECCV 2022*, 2022, pp. 709–727.
[41] T. Lei, J. Bai, S. Brahma *et al.*, "Conditional adapters: Parameter-efficient transfer learning with fast inference," *Advances in Neural Information Processing Systems*, vol. 36, pp. 8152–8172, 2023.
[42] H. Liu, D. Tam, M. Muqeeth *et al.*, "Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1950–1965, 2022.
[43] A. S. Luccioni, S. Viguier, and A. L. Ligozat, "Estimating the carbon footprint of BLOOM, a 176B parameter language model," *JMLR*, vol. 24, no. 253, pp. 1−15, 2023.
[44] J. Pfeiffer, A. Kamath, A. Rücklé*et al.*, "AdapterFusion: Non-destructive task composition for transfer learning," in *Proc. 16th Conf. of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 487–503.
[45] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
[46] O. Vinyals, C. Blundell, T. Lillicrap *et al.*, "Matching networks for one shot learning," *Advances in Neural Information Processing Systems*, vol. 29, 2016.
[47] J. Wang, Z. Yang, X. Hu *et al.*, "GIT: A generative image-to-text transformer for vision and language," arXiv preprint, arXiv: 2205.14100, 2022.
[48] J. Yuan, H. Chen, S. Tian *et al.*, "Prompt-based learning for few-shot class-incremental learning," *Alexandria Engineering Journal*, vol. 120, pp. 287–295, 2025.
[49] S. Marjit, H. Singh, N. Mathur *et al.*, "DiffuseKronA: A parameter efficient fine-tuning method for personalized diffusion models," in *Proc. 2025 IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV)*, 2025, pp. 3529–3538.
[50] C. Tang, K. Wang, and J. v. d. Weijer, "IterInv: Iterative inversion for pixel-level T2I models," in *Proc. 2024 IEEE International Conf. on Multimedia and Expo (ICME)*, 2024, pp. 1–6.
[51] X. Xu, Z. Wang, E. Zhang *et al.*, "Versatile diffusion: Text, images and variations all in one diffusion model," in *Proc. IEEE/CVF International Conf. on Computer Vision (ICCV)*, 2023, pp. 7754–7765.
[52] L. Zhang A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. IEEE/CVF International Conf. on Computer Vision (ICCV)*, 2023, pp. 3836–3847.
[53] S. Zhao, D. Chen, Y. C. Chen *et al.*, "Uni-ControlNet: All-in-one control to text-to-image diffusion models," *Advances in Neural Information Processing Systems,* vol. 36, pp. 11127–11150, 2023.
[54] J.-B. Alayrac and others, "Flamingo: A visual language model for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23716–23736, 2022.
[55] X. Chen, X. Wang, S. Changpinyo *et al.*, "PaLI: A jointly-scaled

multilingual language-image model," arXiv preprint, arXiv: 2209.06794, 2022.
[56] C. Jia, Y. Yang, Y. Xia *et al.*, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. 38th International Conf. on Machine Learning*, 2021, pp. 4904–4916, 2021.
[57] H.-K. Ko, G. Park, H. Jeon *et al.*, "Large-scale text-to-image generation models for visual artists' creative works," in *Proc. 28th International Conf. on Intelligent User Interfaces*, 2023, pp. 919–933.
[58] J. Li, D. Li, S. Savarese *et al.*, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. 40th International Conf. on Machine Learning*, 2023.
[59] J. Li, D. Li, C. Xiong *et al.*, "BLIP: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. 40th International Conference on Machine Learning*, 2022.
[60] X. Pan, L. Dong, S. Huang *et al.*, "Kosmos-G: Generating images in context with multimodal large language models," in *Proc. International Conf. on Learning Representations*, 2024, pp. 40959–40974.
[61] Z. Peng, W. Wang, L. Dong *et al.*, "Kosmos-2: Grounding multimodal large language models to the world," in *Proc. International Conf. on Learning Representations*, 2024, pp. 51575–51598.
[62] A. Radford, J. W. Kim, C. Hallacy *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. 38th International Conf. on Machine Learning*, 2021, pp. 8748–8763.
[63] F. Shakeri, Y. Huang, J. S. Rodríguez *et al.*, "Few-shot adaptation of medical vision-language models," in *Proc. Medical Image Computing and Computer Assisted Intervention-MICCAI 2024*, 2024, pp. 553–563.
[64] Y. L. Sung, J. Cho, and M. Bansal, "VL-ADAPTER: Parameter-efficient transfer learning for vision-and-language tasks," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5217–5227.
[65] B. Zhu, Y. Niu, Y. Han *et al.*, "Prompt-aligned gradient for prompt tuning," in *Proc. IEEE/CVF International Conf. on Computer Vision (ICCV)*, 2023, pp. 15659−15669.
[66] M. Heusel, H. Ramsauer, T. Unterthiner *et al.*, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in Neural Information Processing Systems*, vol. 30, pp. 6626–6637, 2017.
[67] T. Hinz, M. Fisher, O. Wang *et al.*, "Improved techniques for training single-image GANs," in *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 1300–1309.
[68] T. Salimans, I. Goodfellow, W. Zaremba *et al.*, "Improved techniques for training GANs," *Advances in Neural Information Processing Systems*, vol. 29, 2016.
[69] C. Wah, S. Branson, P. Welinder *et al.* (July 2011). The caltech-UCSD birds-200-2011 dataset. [Online]. Available: https://authors.library.caltech.edu/records/cvm3y-5hh21
[70] L. F. W. Anthony, B. Kanding, and R. Selvan, "Carbontracker: Tracking and predicting the carbon footprint of training deep learning models," arXiv preprint, arXiv: 2007.03051, 2020.
[71] D. D. Silva and D. Alahakoon, "An artificial intelligence life cycle: From conception to production," *Patterns*, vol. 3, no. 6, 100489, 2022.
[72] T. Gebru, J. Morgenstern, B. Vecchione *et al.*, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.
[73] T. Y. Lin, M. Maire, S. Belongie *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Computer Vision-ECCV 2014*, 2014, pp. pp 740–755.
[74] C. Schuhmann, R. Beaumont, R. Vencu *et al.*, "LAION-5B: An open large-scale dataset for training next generation image-text models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25278–25294, 2022.
[75] A. L. Stein. Artificial intelligence and climate change. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3665760
[76] T. J. Yang, Y. H. Chen, J. Emer *et al.*, "A method to estimate the energy consumption of deep neural networks," in *Proc. 2017 51st Asilomar Conf. on Signals, Systems, and Computers*, 2017, pp. 1916–1920.
[77] S. Zhou, M. Gordon, R. Krishna, *et al.*, "HYPE: A benchmark for human eye perceptual evaluation of generative models," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
[78] Y. I. Alzoubi and A. Mishra, "Green artificial intelligence initiatives: Potentials and challenges," *Journal of Cleaner Production*, vol. 468, 143090, 2024.
[79] A. Bitzenis, N. Koutsoupias, and M. Nosios, "Artificial intelligence and machine learning in production efficiency enhancement and sustainable development: A comprehensive bibliometric review," *Front. Sustain.*, vol. 5, 1508647, 2025.
[80] V. Bolón-Canedo, L. Morán-Fernández, B. Cancela *et al.*, "A review of green artificial intelligence: Towards a more sustainable future," *Neurocomputing*, vol. 599, 128096, 2024.
[81] J. Cowls, A. Tsamados, M. Taddeo *et al.*, "The AI gambit: Leveraging artificial intelligence to combat climate change—Opportunities, challenges, and recommendations," *AI & Soc.*, vol. 38, pp. 283–307, 2021.
[82] O. Farooq and A. Khan, "The impact of machine learning on climate change modeling and environmental sustainability," *Artificial Intelligence and Machine Learning Review*, vol. 6, no. 1, pp. 8–16, 2025.
[83] E. García-Martín, C. F. Rodrigues, G. Riley *et al.*, "Estimation of energy consumption in machine learning," *Journal of Parallel and Distributed Computing*, vol. 134, pp. 75–88, 2019.
[84] K. Kirkpatrick, "The carbon footprint of artificial intelligence," *Commun. ACM*, vol. 66, no. 8, pp. 17–19, 2023.
[85] A. Nordgren, "Artificial intelligence and climate change: Ethical issues," *JICES*, vol. 21, no. 1, pp. 1–15, 2023.
[86] C. Plociennik, P. Watjanatepin, K. V. Acker *et al.*, "Life cycle assessment of artificial intelligence applications: Research gaps and opportunities," *Procedia CIRP*, vol. 135, pp. 924–929, 2025.
[87] R. Salama, C. Altrjman, and F. Al-Turjman, "A survey of Machine Learning (ML) in sustainable systems," *Multidisciplinary Advance Sciences and Technology*, vol. 2, no. 3, 2023.
[88] Y. Song, T. Wang, S. K. Mondal *et al.*, "A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities," *ACM Computing Surveys*, vol. 55, no. 13s, 271, 2023.
[89] A. V. Wynsberghe, "Sustainable AI: AI for sustainability and the sustainability of AI," *AI Ethics*, vol. 1, no. 3, pp. 213–218, 2021.
[90] Y. Wang, Q. Yao, J. T. Kwok *et al.*, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surv.*, vol. 53, no. 3, 63, 2020.
[91] W. Zeng and Z. Xiao, "Few-shot learning based on deep learning: A survey," *MBE*, vol. 21, no. 1, pp. 679–711, 2024.
[92] S. Santos, A. L. C. Ottoni, R. Borgo *et al.*, "A systematic review of green machine learning practices and challenges for sustainability," *Artificial Intelligence Review*, vol. 59, 132, 2026.
[93] S. K. Ansari and R. Kumar, "Wing- generative adversarial networks: Design of diffusion model for text-to-image representation," in *Proc. 2024 Conf. on Renewable Energy Technologies and Modern Communications Systems: Future and Challenges*, 2024, pp. 1–6.

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).