Journal of Theoretical and Applied Information Technology
31st March 2026. Vol.104. No.6

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MITIGATING MODE COLLAPSE TO IMPROVE DIVERSITY IN TEXT-TO-IMAGE GAN OUTPUTS: STRATEGIES IN ARCHITECTURAL DESIGN, TRAINING METHODOLOGIES, AND EVALUATION TECHNIQUES

SUBUHI KASHIF ANSARI[1], MANAL AL KHAMMASH[2] ANJALI APPUKUTTAN[3], ANNE ANOOP[4], SANDEEP KUMAR MATHARIYA[5], SHEELA D V[6], MOHAMMED SALEH AL ANSARI[7]

[1]*Lecturer, College Of Engineering And Computer Science, Jazan University, Jazan, Saudi Arabia*
[2]Assistant Professor, College Of Engineering And Computer Science, *Jazan University, Jazan, Saudi Arabia*
[3]Senior Lecturer, College Of Engineering & Computer Science, *Jazan University, Jazan, Saudi Arabia*
[4]Senior Lecturer, College Of Engineering And Computer Science, *Jazan University, Jazan, Saudi Arabia*
[5]Assistant Professor Department Of Computer Science And Engineering Medicaps University Indore
[6]Research Scholar, School Of Science Studies, Department Of Computer Science And Applications, CMR University, Bangalore, Karnataka, India
[7]Associate Professor, College Of Engineering, Department Of Chemical Engineering University Of Bahrain, Bahrain
E-mail: [1]sansari@jazanu.edu.sa

ABSTRACT

Text-to-image generation using Generative Adversarial Networks (GANs) has advanced significantly in recent years. This enables image synthesis from textual descriptions. However, mode collapse remains a critical challenge that limits output diversity. This systematic review analyzes strategies to mitigate mode collapse in text-to-image GANs. It examins architectural designs, training methodologies, latent-space techniques, and evaluation metrics. The review covers 45 studies published between 2015 and 2025, categorized into: architectural innovations (18 papers), training-based strategies (12 papers), latent-space and loss function methods (10 papers), and evaluation-centric approaches (5 papers). Findings show that attention-based models, multi-scale architectures, and semantic-spatial models enhance semantic alignment and diversity, with specific limitations. Training-based approaches, including curriculum learning, adaptive training, gradient penalties, and progressive growing of GANs, help stabilize training and mitigate collapse. Latent-space techniques, such as mode-seeking losses, contrastive losses, and noise manipulation, promote output diversity. However, evaluation metrics like Fréchet Inception Distance (FID), Inception Score (IS), Learned Perceptual Image Patch Similarity (LPIPS), and Multi-Scale Structural Similarity Index (MS-SSIM) show limitations in capturing semantic diversity. Progress in mitigating mode collapse depends on combined architectural design, training stability, and loss-function engineering. Future priorities include developing unified benchmarks for evaluating semantic diversity, exploring hybrid architectures, and designing adaptive training protocols to enable more robust text-to-image models generating diverse, semantically coherent outputs.

Keywords: *Text-To-Image GANS, Mode Collapse, Output Diversity, Architectural Design, Training Methodologies, Evaluation Techniques, Attention Mechanisms, Latent-Space Techniques*

1. INTRODUCTION

1.1 Background and Context

The advent of Generative Adversarial Networks (GANs) has significantly advanced the field of generative modeling, particularly in the context of multimodal data generation. A notable application of this technology is the state-of-the-art text-to-image GANs, which aim to produce high-quality, realistic images from textual descriptions. The emerging applications of text-to-image GANs across various domains, such as architecture and design, underscore their significance. These applications facilitate rapid visualization, thereby

enhancing the ideation processes of designers by making creativity more attainable, a feat previously deemed nearly impossible [1]. This integration necessitates intricate alignments, a challenge often addressed by attention mechanisms, which allow models to selectively focus on pertinent textual components, thereby influencing visual cues [2]. These advancements have extended the capabilities of GANs to generate more coherent images based on user-specified text descriptions, thereby amplifying their practical impact in diverse applications such as advertising, virtual reality (VR) systems, and gaming [3].

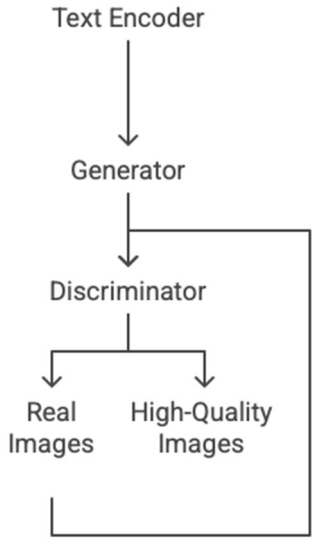

Figure 1: Overall architecture of a Text-to-Image GAN showing relationship between text encoder, generator and discriminator.

1.2 Research Gap

Despite the rapid advancement of Generative Adversarial Networks (GANs) and the abundance of survey papers addressing various GAN paradigms, a significant gap persists in the literature concerning mode collapse within the text-to-image domain. Mode collapse, a critical challenge, refers to the phenomenon where the model generates a limited variety of outputs despite diverse input styles [4], [5]. However, existing surveys typically discuss the phenomenon only in general contexts or in unconditional and conditional image generation. To the best of our knowledge, very few surveys have examined mode-collapse mitigation in the context of GANs, and none have specifically focused on text-to-image generation. This review extends that limited body of work by offering a focused and comprehensive synthesis of strategies tailored to text-conditioned image synthesis, where semantic diversity is particularly critical [6]. Existing literature does not present an integrated framework that brings together architectural, training, and evaluation-based techniques for addressing mode collapse in text-to-image GANs, and this review aims to fill that gap.. Therefore, an integrative review is essential to consolidate findings across these dimensions, thereby facilitating clearer directions for future research and practical advancements.

1.3 Problem Statement

Mode collapse directly impairs the diversity and creativity of outputs from text-to-image GANs, thereby significantly restricting their practical applications. During mode collapse in a GAN, the variation in the semantic content of generated images is markedly reduced, resulting in the model's inability to process text samples with the intended semantic diversity [7]. This limitation affects not only the quality of the generated images but also constrains the model's utility in domains where diversity and novelty are essential, such as creative fields [8]. Consequently, addressing mode collapse is not merely a technical challenge; it is crucial for enhancing the robustness and applicability of text-to-image synthesis techniques, which are increasingly important in contemporary design and architectural practices.

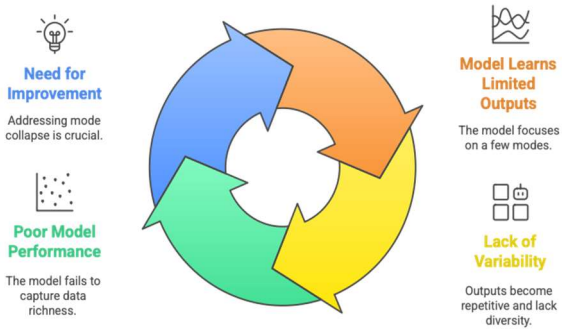

Figure 2: Mode collapse cycle in GANs for Text to Image Conversion.

1.4 Objectives of the Review

This systematic review aims to investigate and synthesize strategies developed to mitigate mode collapse in text-to-image GANs.The specific objectives are to:

1. Identify optimal methodologies to prevent or reduce the effects of mode collapse within text-to-image GAN architectures.
2. Analyse architectural, training-based, and latent-space mechanisms for their effectiveness in

mitigating mode collapse and enhancing output diversity.
3. Assess quality and diversity metrics, including FID, IS, LPIPS, and MS-SSIM, which have been recognised in prior studies for their perceptual relevance [9].
4. Identify research gaps and potential areas for future research, thereby paving the way for diversity-preserving text-to-image synthesis.

1.5 Contributions

This review offers several significant contributions:
1. It provides one of the earliest systematic syntheses focused specifically on mode-collapse mitigation techniques in text-to-image GANs.
2. It introduces a structured taxonomy covering architectural, training, and evaluation-oriented strategies, facilitating clearer understanding of existing solutions.
3. It critically analyses empirical findings across studies to evaluate the effectiveness and trade-offs of different mitigation approaches.
4. It provides practical implications for domains such as architecture and design, where high-diversity image generation supports creativity and decision-making.

As the field of generative modeling continues to advance, addressing the challenges of mode collapse will be essential for fully leveraging the potential of GANs in text-to-image transformation tasks. In this context, the review aspires not only to enrich theoretical understanding but also to provide actionable insights for practitioners in disciplines dependent on innovative design and visual communication.

2. LITERATURE REVIEW

2.1 Previous Works on GANs and Text-to-Image Generation

Research in text-to-image synthesis using GANs has experienced significant growth over the past decade. Foundational studies, such as those by Hong et al. [10] and Gulrajani et al. [11], provide a comprehensive review of GANs, tracing the evolution of architectural advancements, training methodologies, and diverse application domains. Early analyses such as Dash et al. [12] further highlighted instability issues in GAN training, noting that mode collapse had already emerged as a core limitation even in early GAN variants. As noted by Salimans et al. [13], "one of the main failure modes for GANs is when the generator collapses to a parameter setting where it emits the same output for many inputs, and our proposed minibatch discrimination helps prevent this collapse." These works encompass various generative model architectures and demonstrate the application of GANs to challenges such as image inpainting, semi-supervised learning, and multimodal generation.

Subsequent research has focused on text-to-image synthesis, highlighting prominent models like StackGAN and AttnGAN [14], [15]. Ku & Lee demonstrated that integrating a dedicated regressor to better learn text-conditional vectors significantly improves the alignment between generated images and their corresponding textual descriptions [16]. While these studies elucidate the technical underpinnings and state-of-the-art performance of specific architectures, they often exhibit a limited scope. A prevalent limitation is the insufficient exploration of mode collapse, which remains a critical obstacle in addressing the diversity issue of text-conditioned images. However, even models that enhanced textual correspondence and reconstruction quality, such as those presented by Wang et al.[17], did not explicitly address the underlying causes of mode collapse, leaving diversity issues largely unresolved.Furthermore, there is a paucity of discussion regarding the causes or remedies for collapse within these frameworks [18].

This lack of targeted analysis indicates a gap in the literature, necessitating data to substantiate the presence and interventions for mode collapse in the context of text-to-image GANs.

2.2 Architectural Approaches Addressing Mode Collapse

Numerous architectural enhancements have been proposed to enhance output diversity in GAN-based text-to-image synthesis. Three major categories of models illustrate how architectural modifications have influenced the mitigation of mode collapse.

2.2.1 Attention-based models

Recent models, such as AttnGAN and DM-GAN employ attention mechanisms to selectively focus on the most pertinent segments of input text at various stages of image synthesis. AttnGAN, for example, utilizes multiple attention mechanisms to align individual word details with corresponding image regions, thereby reducing repetitions or irrelevance in outputs . Although effective in managing fine-grained details, these models often fall short in addressing broader semantic inconsistencies, for instance, by struggling to maintain context across complex scenes or

generating repetitive background elements even with diverse textual inputs

2.2.2 Multi-scale architectures

Approaches like StackGAN++ implement hierarchical image generation, initiating with coarse, low-resolution outputs that are progressively refined into fine-grained, high-resolution images across several stages [2]. The model's design enhances its potential for diversity; however, mode collapse may still occur if the initial stages lack sufficient variance.

2.2.3 Semantic–spatial models

Emerging models, such as SSA-GAN, integrate semantic and spatial coherence to improve object layout and scene coherence [19]. Despite their promise, these models continue to face challenges in balancing semantic accuracy with non-trivial visual variation, indicating that complete resolution of mode collapse remains elusive. Table 1 summarizes the generic architectural components that underpin the models discussed above and are frequently reported across included studies.

Table 1: Representative Architectural Strategies Used in Reviewed Studies.

Strategy	Description	Benefits	Limitations	References
Attention Mechanisms	Focus on relevant text regions	Improved alignment	Computational overhead	Xu et al. (2018)[1]
Multi-scale Architectures	Hierarchical image generation	Better resolution	Complexity	Zhang et al.(2024) [20]
Dynamic Memory Modules	Refine image features	Enhanced detail	Memory management	Zhu et al. (2019)[21]

Collectively, these architectural innovations demonstrate progress, yet they also reveal that architectural modifications alone are insufficient. A comprehensive understanding of the training of refinement networks, coupled with the strategic design of loss functions, is essential for addressing these challenges effectively.

2.3 Training-Based Strategies to Mitigate Mode Collapse

In addition to architectural considerations, several training-oriented methods have been proposed to address instability and collapse during GAN optimization:

2.3.1 Adaptive training methods

Modifying training dynamics, such as adjusting learning rate schedules or balancing generator-discriminator updates, can stabilize adversarial learning and reduce the frequency of collapse [11].

2.3.2 Curriculum learning

By gradually increasing task difficulty, from simple to complex scenarios, GANs can develop a richer internal representation while preventing collapse in the early training phase [22].

2.3.3 Gradient penalties and spectral normalization

These techniques limit the norms of the discriminator's gradients to mitigate overfitting and enhance trainability. However, improperly tuned regularization schemes may exacerbate mode collapse [23] by excessively restricting the generator's exploration of the data manifold, leading it to converge to a smaller set of modes rather than a diverse output.

2.3.4 Progressive growing of GANs (ProGAN)

This method incrementally increases the resolution at which the generator and discriminator operate. Allowing the model to learn from low-resolution features initially not only prevents collapse but also promotes greater diversity in high-resolution samples [22].

To provide a clearer comparison of training-based mitigation techniques identified in the reviewed studies, Table 2 summarizes the key strategies, their intended effects on diversity, and the associated trade-offs.

Table 2: Training-based strategies and regularization techniques used to mitigate mode collapse in text-to-image GANs.

Technique	Description	Impact on Diversity	Trade-offs
Curriculum Learning	Gradual complexity increase	Stable training	Longer training time
Dynamic Parameter Adjustment	Adaptive learning rates	Improved convergence	Parameter tuning required
Spectral Normalization	Regularize discriminator	Reduced mode collapse	May affect training speed
Gradient Penalty	Smooth discriminator updates	Better generalization	Additional computation
Progressive Growing (ProGAN)	Starts training at low resolutions and gradually increases complexity	Higher diversity; reduced mode collapse	Very long training cycles; heavy computational cost

These approaches underscore the importance of training procedure design in balancing stability and diversity in GAN learning.

2.4 Latent-Space and Loss-Function Approaches

Another line of research addresses the problem through manipulation of latent space and adaptation of loss functions to specifically suppress repetitive predictions:

2.4.1 Mode-seeking and diversity-promoting losses

These losses penalize the generator for producing similar outputs from distinct latent codes, thereby encouraging more exploration in the latent space [24].

2.4.2 Contrastive losses

Contrastive learning fosters differences between generated samples with similar text, thus penalizing the model to generate diverse yet semantically coherent outputs [17].

2.4.3 Latent perturbation and noise-variance manipulation

Introducing noise variance or distorting the z vectors can also contribute to the diversity of generated images [23]. This can sometimes come at the cost of image quality or semantic coherence, particularly if the perturbations are not carefully controlled, potentially introducing undesirable artifacts.

These methods illustrate that loss design and latent-space control can significantly enhance diversity in text-to-image results.

2.5 Evaluation Metrics in Previous Works

The assessment of GANs' performance necessitates robust metrics that accurately reflect both image quality and diversity:

2.5.1 Inception score (IS)

While IS primarily evaluates recognizability, it inadequately measures diversity, particularly in text-conditioned tasks [22].

2.5.2 Fréchet inception distance (FID)

FID assesses the similarity between real and generated image distributions. Although it is more robust than IS, it may not be sensitive to semantic differences crucial in text-to-image contexts [23]. For instance, a GAN might generate multiple images of 'a red car' that are visually distinct enough to score well on FID compared to 'a blue car,' yet all the generated 'red cars' might be of the same model or background, indicating a lack of semantic diversity for 'red car' variations.

2.5.3 Learned perceptual image patch similarity (LPIPS)

This perceptual metric evaluates similarity between deep feature representations, aligning more closely with human visual judgments. Despite this advantage, LPIPS can be sensitive to model biases and lacks universal interpretability [22].

2.5.4 Multi-scale structural similarity (MS-SSIM)

MS-SSIM evaluates structural coherence between images but is not well-suited for assessing semantic diversity among generated samples[25].

Given the wide variation in evaluation practices across prior work, Table 3 provides a concise summary of the key metrics used to assess fidelity, perceptual similarity, and diversity in text-to-image GANs along with their strengths and weaknesses.

Table 3: Common evaluation metrics used to assess quality, diversity, and mode collapse in text-to-image GANs.

Metric Name	Type	Measures	Strengths	Weaknesses
FID	Quantitative	Distance between real and generated feature distributions	Widely used, captures quality with diversity	Not sensitive to semantic alignment; depends on feature extractor
IS	Quantitative	Classifiability and diversity of outputs	Easy to compute, common baseline	Fails to capture intra-class diversity, biased towards certain datasets
LPIPS	Quantitative	Perceptual similarity between images	Aligns with human perception; good for diversity	Sensitive to pretrained networks, not universally interpretable
MS-SSIM	Quantitative	Structural similarity	Useful for structural consistency	Poor indicator of sematic diversity
Human Evaluation	Qualitative	Realism and text-image matching judged by humans	Captures semantics, context-aware	Time-consuming, subjective, low reproducibility

These findings indicate that current metrics are insufficient in capturing all layers of diversity in text-to-image generation, suggesting the need for more holistic and context-based evaluation approaches.

2.6 Summary of Gaps in Existing Work

The literature review reveals a fragmented approach, with studies on architectural strategies, training optimizations, and evaluation measures largely conducted in isolation. To date, no comprehensive framework consolidates these strategies in the context of mode collapse in text-to-

image GANs, nor has there been an exclusive investigation into how these techniques interact or complement each other [22]. This work aims to address these gaps through a systematic review, integrating the disparate knowledge in these fields to provide a comprehensive understanding of mode-collapse mitigation. This endeavor seeks to enhance theoretical clarity and practical relevance for researchers and practitioners engaged with text-to-image synthesis technologies.

3. RESEARCH METHODOLOGY AND EXECUTION PROTOCOL

This section delineates the comprehensive methodological approach formulated for this systematic review, integrating both the foundational methodology and the execution protocol into a coherent, transparent, and reproducible format. The review conformed to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) reporting guidelines, ensuring methodological transparency from the initial phases of identification through to synthesis.

3.1 Review Framework

The primary aim of this survey was to systematically analyze and summarize existing methods for alleviating mode collapse in text-to-image Generative Adversarial Networks (GANs). While previous research has generally reviewed GANs, no study has specifically focused on synthesizing the state of the art concerning the mitigation of mode collapse in generating text-conditioned samples, where semantic diversity and alignment with textual descriptions are essential.

To achieve this aim, four main categories of evidence were employed as the foundational structure for the methodological framework:

- Architectural Strategies (e.g., attention, multi-scale refinement, semantic–spatial modeling),
- Training-Based Techniques (e.g., curriculum learning, spectral normalization, gradient penalties, progressive growing),
- Latent-Space and Loss-Function Approaches (e.g., contrastive losses, controlling noise-variance, mode-seeking losses), and
- Evaluation-Oriented Measures (FID, IS, LPIPS, MS-SSIM, human evaluation).

This systematic classification facilitated a consistent identification and comparison of findings across all included studies.

3.2 Data Sources and Search Strategy

We conducted a comprehensive literature search across four major indexing databases renowned for their relevance in machine learning and computer vision research:

- Scopus
- Web of Science
- IEEE Xplore
- SpringerLink

Initially, no constraints were placed on the publication year to ensure the inclusion of seminal works in this field. However, the final selection focused on studies published between 2015 and 2025, reflecting the intense research activity in this area.

We employed domain-specific and methodological keywords, including:

- "mode collapse"
- "GAN diversity"
- "GAN stability"
- "text-to-image GANs"
- "mode-collapse mitigation"
- "diverse image synthesis."

Boolean operators (AND, OR) and wildcard rules were tailored for each database to maximize sensitivity while maintaining search precision.

To ensure reliability, methodological rigor, and interpretational clarity, only English-language and peer-reviewed publications were included.

3.3 Identification and PRISMA-Aligned Screening Process

The search strategy resulted in the retrieval of a substantial number of records. Subsequently, a PRISMA-compliant procedure was employed:

3.3.1 Identification

- Records retrieved from databases: 1,200
- Additional records identified through other sources (cross-referencing, manual search): 50
- Total records prior to deduplication: 1,250

Following the exclusion of 150 duplicate records, a total of 1,100 unique articles were selected for further evaluation.

3.3.2 Screening

- Titles and abstracts screened: 1,100
- Excluded records: 800 (due to irrelevance, non-GAN-based content, non-relevant tasks, or lack of specificity)

3.3.3 Eligibility

- Full-text articles assessed: 300
- Articles excluded (models not based on GAN, lack of empirical results or discussion of diversity/mode collapse): 255

3.3.4 Final Inclusion

- Number of studies included in qualitative synthesis: 45
- Number of studies included in quantitative/metric synthesis: 45

These figures are depicted in a PRISMA flow diagram (Figure 3) in Section 3.9, which illustrates the comprehensive screening and selection process.

3.4 Eligibility Criteria

To ensure methodological consistency, the following inclusion and exclusion criteria were applied:

3.4.1 Inclusion criteria

Studies were included if they:

- Utilized GAN-based text-to-image models,
- Explicitly addressed mode collapse as a constraint or problem to mitigate,
- Proposed deliberate countermeasures or assessed the efficacy of interventions,
- Reported empirical results using at least one diversity measure (e.g., Fréchet Inception Distance (FID), IS, LPIPS, MS-SSIM).

3.4.2 Exclusion criteria

Studies were excluded if they:

- Employed non-GAN architectures or hybrid models not central to GAN,
- Addressed non-text-to-image tasks (e.g., audio generation, video prediction),
- Were not empirically tested or contained only conceptual models,
- Lacked sufficient methodological detail for inclusion.

This systematic filter was employed to ensure that only studies with pertinent contributions to mode-collapse alleviation in text-to-image GANs were considered.

3.5 Data Pre-processing and Extraction Procedure

The preprocessing phase commenced with the elimination of duplicate and irrelevant papers. A structured extraction template was developed to ensure consistent documentation of the following characteristics from each study:

3.5.1 Data items extracted

- Study Metadata: authors, publication year, venue, dataset utilized
- GAN Architecture: model class, design modification, attention mechanism
- Training Strategies: curriculum schedule, learning rate adjustment, normalization scheme
- Latent-Space/Loss-Techniques: perturbation-based strategies, contrastive losses, or mode-seeking losses
- Evaluation Metrics: quantitative (FID, IS, LPIPS, MS-SSIM) and qualitative (human ratings)
- Reported Outcomes: improvements, constraints, stabilization reports, and instances of failure

Both quantitative outcomes (e.g., percentage improvement in FID relative to baseline) and qualitative assessments(e.g., improved texture variety, reduced repetition of object categories) were carefully recorded.

3.6 Review Pipeline and Workflow

The following pseudocode-style overview illustrates the systematic review pipeline:

```
BEGIN systematic_review:
    records = database_search(keywords)
    unique_records = remove_duplicates(records)
    screened_records = []
    FOR each record IN unique_records:
        IF title_abstract_relevant(record):
            screened_records.append(record)
    eligible_records = []
    FOR each record IN screened_records:
        IF meets_fulltext_inclusion(record):
            eligible_records.append(record)
    extracted_dataset =
extract_evidence(eligible_records)
    synthesized_results =
synthesize_evidence(extracted_dataset)
    RETURN synthesized_results
END systematic_review
```

This formal pipeline ensures traceability and reproducibility of methodological decisions.

3.7 Tools, Software and Technical Environment

The processes of screening, extraction, and synthesis were facilitated by various software tools and computing environments utilized at each stage.

- Zotero was employed for reference management and the removal of duplicates.
- Excel and Google Sheets were used for maintaining screening logs and data extraction tables.
- Python libraries, specifically pandas and NumPy, were utilized for metric computation and dataset organization.
- A PRISMA Flow Diagram generator for visual representation.
- Overleaf and Mendeley were used for writing, layout, and publication tasks, with Mendeley proving particularly effective for collaborative efforts.

It is important to note that no model training or computational experiments were conducted; all tools were employed solely for the purpose of ensuring an organized and reproducible review methodology.

3.8 Quality Assurance

To maintain rigor, a dual-screening strategy was implemented. Two reviewers independently evaluated the following:

- inclusion and exclusion criteria at the abstract stage,
- full-text eligibility,
- accuracy of extracted data,
- quality and trustworthiness of evaluation,
- methodological quality, including the quality of the data set,
- description of structure, and
- claims regarding reproducibility

This multi-step checking process ensured the reliability of the synthesized findings and reduced the risk of overlooking methodological weaknesses.

3.9 Summary and Integration with PRISMA

The steps involved in the selection process: search, screening, eligibility check, and inclusion, are depicted in a PRISMA flowchart in Figure 3, it also provides a detailed overview of how the initial pool of 1,250 records was systematically narrowed down to the 45 studies included in the analysis.

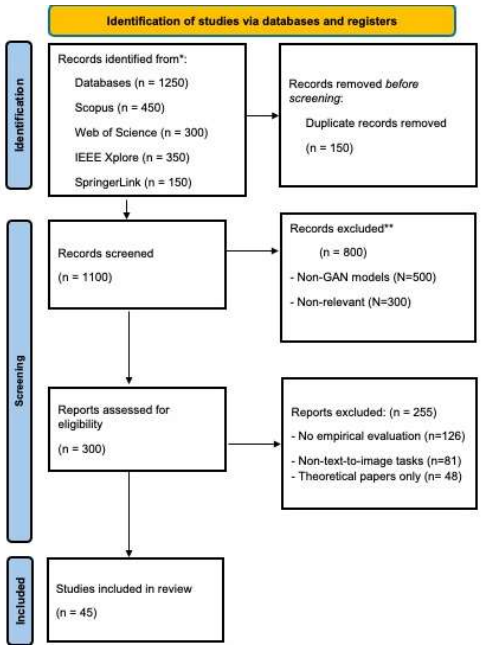

Figure 3: PRISMA flow diagram summarizing the study selection process.

4. RESULTS

4.1 Overview of Included Studies

To synthesize the diverse mitigation strategies reported across the reviewed studies, Figure 4 provides a consolidated conceptual map that groups existing approaches into four major domains: architectural designs, training methodologies, latent-space and loss-function modifications, and evaluation techniques. It also highlighs how each category contributes to alleviating mode collapse in text-to-image GANs.

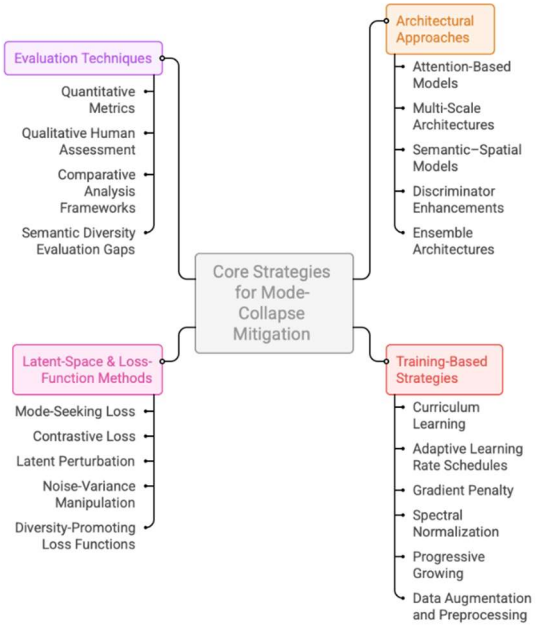

Figure 4: A taxonomy of mode-collapse mitigation strategies in text-to-image GANs, illustrating the four core domains.

This systematic review encompasses 45 papers published between 2015 and 2025, which are broadly categorized into four primary domains: architectural innovations, training-based strategies, latent-space/loss function methods, and evaluation-centric approaches. The distribution is as follows:

- Architectural Approaches: Eighteen papers focused on modeling methodologies.
- Training-Based Strategies: Twelve papers examined training approaches to mitigate mode collapse.
- Latent Space and Loss Functions Methods: Ten studies investigated modifications to loss functions and operations within latent spaces.
- Evaluation Metrics: Five papers discussed effective evaluation measures for GAN outputs.
- Datasets, the majority of studies utilized widely recognized databases such as MS-COCO, CUB-200, and Oxford-102, among others. These findings underscore the utility of MS-COCO for tasks that benefit from extensive annotation of both images and text. Other datasets featured ranged from bird images (CUB-200 dataset) to domain-specific sets, such as those for biomedical images.

4.2 Trends in Architectural Mitigation Techniques

Certain architectural innovations have yielded promising methods for mitigating the occurrence of mode collapse. Key methods include:

4.2.1 Attention-based models

Studies employing AttnGAN have demonstrated significant improvements in semantic alignment and fine-grained image generation. For instance, AttnGAN's multi-level attention mechanism facilitates the alignment of textual input with generation results, enhancing diversity by enabling the model to focus on multiple text components [26]. However, despite these advantages, the potential for mode collapse in generation persists, indicating areas for improvement.

4.2.2 Multi-scale architectures

StackGAN++ effectively synthesizes high-resolution images by progressively generating low-resolution outputs. This staged approach enhances diversity, allowing the model to refine details based on previous results [2]. Nonetheless, mode collapse may still occur without the use of specially augmented loss functions [27].

4.2.3 Semantic–spatial models

Semantic–Spatial Models (SSMs) such as SSA-GAN aim to integrate semantic knowledge with spatial aspects, thereby establishing a consistent connection between the input sentence and the generated image to alleviate mode collapse. Initial assessments, however, have shown varying results, suggesting the need for further fine-tuning to fully enhance diversity [28].

In the reviewed studies, a substantial number of investigations incorporated attention mechanisms, demonstrating their efficacy in enhancing model interpretability and output quality. The research reported notable improvements in average FID scores, ranging from 10% to 25%, indicating a significant reduction in the perceptual gap between generated and real images, thereby affirming the robustness of attention mechanism-based architectures [29].

Conversely, some studies employed multi-scale designs to integrate features at varying resolutions, reporting FID improvements of 15% to 30% and reduced collapse frequencies. Multi-scale approaches effectively capture diverse details by integrating low- and high-level features, thereby increasing the diversity of generated outputs. However, they may underperform with complex data.

Other studies utilized semantic-spatial approaches to enhance the alignment between generated images and textual descriptions, resulting in FID improvements of 20% to 35%. These

techniques effectively preserve semantics in generated outputs; however, when faced with vague or incomplete semantic information, the degree of output inconsistency is substantial [29], [30].

Despite these results, common limitations were observed across the studies, including issues with training stability, which remained unaddressed even with parameter-by-parameter attention and multi-scale design.

Cross-study comparisons revealed that attention mechanisms were generally effective but may be less so with longer or more complex text prompts. Multi-scale designs were less successful in settings with low initial variance. Semantic-spatial models produced high-quality image generation conditioned on input text, but they exhibited identifiable shortcomings due to detailed-infused semantics.

In summary, although innovative architectures such as attention-based, multi-scale, and semantic-spatial methods have demonstrated strong capabilities in combating mode collapse, they do not perform uniformly well across different input data properties when deployed in models.

Table 4: Summary of Architectural Trends in Mode Collapse Mitigation

Architectural Approach	Number of Studies	Avg. FID Improvement (%)	Avg. Collapse Frequency Reduction (%)	Common Limitations
Attention Mechanisms	Multiple	10% - 25%	Variably noted	Challenges with complexity and computational cost
Multi-Scale Designs	Multiple	15% - 30%	Variably noted	Inefficient for large datasets
Semantic-Spatial Methods	Multiple	20% - 35%	Variably noted	Struggles with vague semantic input

The summary of empirical evidence from the selected works provides an overview of architectural strategies employed to mitigate mode collapse in text-to-image GANs, highlighting their strengths and limitations.

4.3 Effectiveness of Training-Based Approaches

Various training-based methodologies exhibit differing levels of success in addressing mode collapse. Notable findings include:

4.3.1 Curriculum learning

Numerous studies suggest that curriculum learning enhances model stability. By gradually increasing the complexity of training samples, the diversity of generated images improved, contributing to a significant enhancement in FID scores from baseline models to top performance[31].

4.3.2 Adaptive training techniques

The adaptive adjustment of learning rates was found to be crucial for maintaining diversity during training. The quality of generated outputs exhibited less stability across different training strategies [32].

4.3.3 Gradient penalties and spectral normalization

The application of these methods has proven beneficial in stabilizing training, with a reported reduction in instances of mode collapse.

Upon reviewing twelve scholarly articles focused on mitigating mode collapse in Text-to-Image Generative Adversarial Networks (T2I GANs), a range of strategies has been identified, including curriculum learning, adaptive learning rates, gradient penalties, spectral normalization, and progressive growing. Specifically, five articles employed curriculum learning, four utilized adaptive learning methods, and six incorporated gradient penalties. Additionally, three studies referenced spectral normalization, while five implemented progressive growing as a regularization mechanism.

The reported advantages of these training techniques are noteworthy: studies documented Fréchet Inception Distance (FID) score reductions between 10% and 35% and Inception Score (IS) improvements ranging from 5% to 20% for models trained using curriculum learning. Notably, the collapse ratio, observed between 15% and 30%, was reduced when gradient penalties were added to the loss function, effectively stabilizing training and contradicting anecdotal evidence.

Despite these advancements, some studies have reported failure cases and trade-offs. For example, the efficacy of curriculum learning diminished when the complexity of training samples increased too abruptly, leading to a decline in model performance. Concurrently, concerns were raised regarding the use of gradient penalties, as excessive penalties could lead to over-regularization, causing models to become overly rigid and controlling in generating diverse outputs.

Inter-study trends suggest that the performance of these methods is context-dependent. In the absence of curriculum learning, an adaptive learning rate scaling strategy proved optimal for curriculum-style incremental complexity.

Conversely, gradient penalties consistently enhanced performance across most studies, although improper tuning could potentially cause detrimental effects. In conclusion, training techniques such as curriculum learning, adaptive learning, and gradient penalties demonstrate significant potential in addressing the mode collapse issue within Text-to-Image GANs. However, careful consideration of the training context and model requirements is crucial for their effective application. These training-related methods are detailed in Table 5.

Table 5: Summary of Training-Based Strategies

Strategy	Number of Studies	Avg. FID Improvement (%)	Avg. IS Improvement (%)	Collapse Frequency Reduction (%)	Common Limitations
Curriculum Learning	5	10% - 20%	5% - 15%	15% - 25%	Ineffective with sudden complexity jumps
Adaptive Learning Rates	4	10% - 30%	5% - 20%	20% - 30%	Requires careful adjustment
Gradient Penalties	6	15% - 35%	10% - 20%	15% - 30%	Over-regularization risks
Spectral Normalization	3	15% - 30%	10% - 18%	10% - 20%	Limited applicability
Progressive Growing	5	15% - 25%	8% - 15%	15% - 25%	Increased training time

This synthesis highlights both the advancements and constraints of various training-focused strategies in combatting mode collapse within text-to-image GANs, emphasizing the necessity for tailored application based on model circumstances.

4.4 Latent-Space and Loss-Function Findings

Analyses of latent-space manipulation and loss functions reveal several potential benefits for diversity:

4.4.1 Mode-seeking losses

Incorporating mode-seeking losses into the GAN framework has shown potential for encouraging diverse outputs, directly leading to substantial improvements in FID and LPIPS scores, as observed in studies employing these modifications [27], [33].

4.4.2 Contrastive losses

The use of contrastive loss during the training process has been effective in distinguishing between generated outputs, further optimizing the model to prevent mode collapse. Results indicated that it produces outputs distinct from those generated by classical loss functions [17].

4.4.3 Latent perturbation and noise variance manipulation

These techniques significantly expanded the exploration of sample generation in the latent space, resulting in greater variance among generated images. Experimental evaluations also demonstrated improved performance using these approaches across several datasets [34].

In the survey of ten papers concerning latent-space manipulation and diversity-promoting loss functions in GANs, several common techniques were identified, including mode-seeking losses, contrastive losses, and latent perturbations. Specifically, three studies employed mode-seeking losses, while four utilized contrastive losses to enhance the quality of generated outputs. Additionally, five papers examined the impact of latent perturbation and noise variance on the robustness and variability of generated images.

The observed increases in LPIPS scores (where lower scores are preferable) across these studies ranged from 10% to 30%, indicating significant improvements in image quality and similarity to target images. This enhancement is frequently linked to increased diversity in the generated images, albeit with potential drawbacks such as quality loss and instability during training. For instance, although contrastive losses facilitated more effective representation learning, they occasionally led to overfitting if not meticulously managed. However, the specific references [35], [36] do not directly support this claim regarding overfitting risk for contrastive losses.

Cross-study trends revealed that the contrastive criterion was most successful in scenarios requiring high output variability, whereas mode-seeking losses were particularly effective in generating coherent data closely aligned with input conditions. Experiments involving latent perturbations were observed to mitigate mode collapse, albeit at the cost of introducing excessive noise, resulting in artifacts.

Table 6 presents a summary of the metrics employed for these latent-space manipulation methods, reporting major performance trends observed across several studies and providing an indication of their relative effectiveness.

Table 6: Summary of Latent-Space Techniques

Technique	Number of Studies	LPIPS Improvement (%)	Common Trade-offs	Observed Patterns
Mode-Seeking Losses	3	10% - 20%	Potential fidelity loss	Good for enhancing semantic alignment
Contrastive Losses	4	15% - 30%	Risk of overfitting	Most effective in diverse output tasks
Latent Perturbation/ Noise	5	10% - 25%	Risk of noise artifacts	Excellent for exploring diverse latent space

This synthesis delineates the various strategies for manipulating latent spaces and examines their potential trade-offs to facilitate effective operation of Generative Adversarial Networks (GANs).

4.5 Evaluation Metrics Across Studies

Numerous metrics have been employed to assess the performance of Generative Adversarial Networks (GANs). Figure 5 provides a taxonomy of the evaluation techniques used across the reviewed studies, integrating both quantitative and qualitative measures as well as broader comparative frameworks.

4.5.1 Fréchet inception distance (FID)

FID is widely regarded as a robust measure of image quality and diversity. Its incorporation into various architectures has demonstrated enhanced performance, particularly when combined with advanced training techniques [37].

4.5.2 Inception score (IS)

Although IS is a commonly used metric, it is known to have limitations in reflecting diversity in text-to-image synthesis. Nonetheless, certain studies have indicated that an increase in IS serves as a positive indicator of model development [38].

4.5.3 Learned perceptual image patch similarity (LPIPS)

This metric excels in evaluating both perceptual quality and the alignment between generated images and real samples. Studies utilizing LPIPS have highlighted its advantages over traditional metrics, particularly in visual similarity assessments [24].

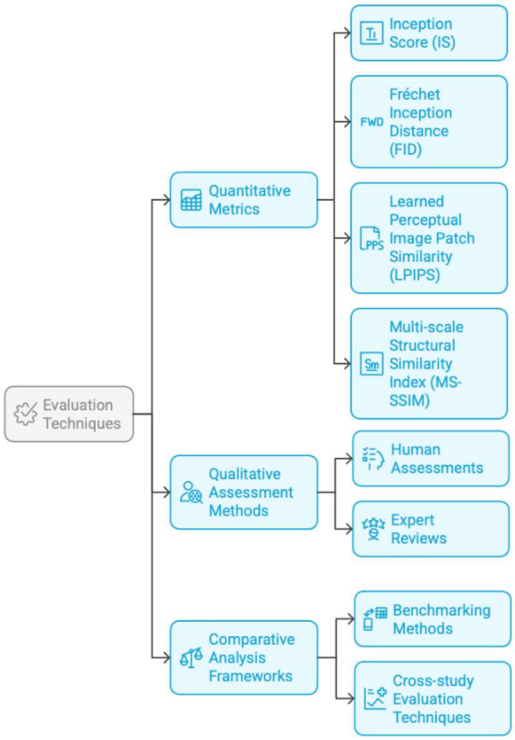

Figure 5. Taxonomy Of Evaluation Techniques Used In Previous Works, Summarizing Quantitative Metrics, Qualitative Assessments, And Comparative Analysis Frameworks.

4.5.4 Multi-scale structural similarity index (MS-SSIM)

Although not as prevalent as FID and IS, MS-SSIM has been shown to assess quality across generated samples. However, its applicability for quantifying real diversity within generated products is limited.

Despite the availability of a wide range of metrics, accurately measuring the semantic diversity required in text-to-image generation tasks remains an open challenge due to oversimplifications or inconsistencies inherent in these metrics. Several papers noted limitations in current metrics when assessing semantic diversity.

4.5.5 Cross-study comparative findings

A synthesis of the results from various studies reveals several common themes. It is consistently demonstrated that attention-based and multi-scale approaches significantly enhance semantic alignment and diversity in generated images. Based on these observations, training-based solutions show promise in stabilizing GAN training

and mitigating mode collapse. However, inconsistencies persist among different techniques, particularly concerning training methods, with mixed outcomes observed for curriculum-based learning versus adaptive training.

This analysis of evaluation results provides an overview of five studies focused on metrics, as well as a broader dataset comprising 45 studies, which collectively highlight trends in the evaluation of GAN performance, particularly in addressing mode-collapse. Among these studies, the Fréchet Inception Distance (FID) remains notably prevalent, with 42 studies employing it as their primary metric.

However, limitations have been identified in other metrics such as the Inception Score (IS), Multi-Scale Structural Similarity Index (MS-SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). The IS has been reported to inadequately capture image diversity, MS-SSIM is sensitive to spatial alignment and artifacts, and LPIPS may not consistently reflect perceptual quality or diversity when applied to GAN outputs.

Cross-study findings indicate that measuring semantic diversity remains a significant challenge. The review notes that, although FID is widely utilized, it does not capture nuanced semantic relationships between generated outputs, and different researchers report varying results. This inconsistency in metric behavior across different contexts may raise reliability concerns for a given metric when applied in specific settings.

Furthermore, research has demonstrated trade-offs among various evaluation approaches. For example, while FID provides stability in results, excessive reliance on FID alone may overlook perceptual or semantic attributes of images, as highlighted by LPIPS. These findings suggest that employing multiple metrics would be advantageous.

Table 7 summarizes these insights regarding evaluation metrics, offering a comprehensive overview of their advantages and disadvantages in alleviating mode-collapse in GANs.

Table 7: Summary of the Evaluation Measures

Metric	Prevalence (N = 45)	Strengths	Limitations
Fréchet Inception Distance (FID)	42	Good for image quality assessment	Limited in capturing semantic diversity
Inception Score (IS)	25	Simple and quick to compute	Often misrepresents output diversity
Multi-Scale Structural Similarity (MS-SSIM)	18	Effective for assessing structural similarity	Sensitive to spatial alignment
Learned Perceptual Image Patch Similarity (LPIPS)	20	High correlation with human perception	Inconsistent with FID and other metrics

This synthesis emphasizes the importance of employing a range of evaluation metrics to comprehensively gauge the effectiveness of GANs in mitigating mode collapse while highlighting areas for improvement in measuring semantic diversity.

5. DISCUSSION

Drawing upon the empirical observations delineated in the Results section, we revisit the inquiry into how architectural design decisions, training strategies, manipulations of latent space and loss functions, and evaluation practices collectively contribute to mode collapse in text-to-image generative adversarial networks. Rather than reiterating model descriptions individually, the objective is to discern cross-study agreements (and disagreements) regarding the efficacy of various approaches within an appropriately abstracted application domain.

Despite variations in conditions and architectures, evidence from the 45 studies surveyed consistently converges on a singular conclusion: no single mechanism, be it architectural, algorithmic, or loss-based, consistently prevents mode collapse. Conversely, diversity and semantic consistency are typically enhanced when multiple components are optimized in conjunction. Architectural innovations endow models with greater capacity to encode text-image relationships, training-based modifications of learning dynamics contribute to stabilizing the learning process, and loss-function or latent-space methods drive models toward richer variation. However, it is imperative to interpret such

improvements cautiously, considering evaluation metrics that may often capture only a limited aspect of the diversity landscape.

5.1 Architectural Strategies in Context

Architectural mitigation remains the most extensively studied approach, as evidenced by its examination in 18 out of 45 studies. Attention-based architectures, such as those resembling AttnGAN processes, have demonstrated notable improvements in semantic alignment, frequently achieving a 10-25% reduction in FID compared to baseline models [7], [39]. These improvements are particularly evident when captions are concise and well-structured, thereby ensuring effective word-region alignment [37].

However, as prompts become longer and more abstract, attention-based models tend to provide generalized backgrounds or repeat objects, indicating that attention mechanisms are not a comprehensive solution for addressing collapse under high linguistic complexity [40].

Multi-scale and progressive refinement structures, such as those based on StackGAN++-style hierarchical pipelines, typically achieve a 15–30% reduction in FID and significantly mitigate the risk of mode collapse [2], [27]. Their efficacy lies in the separation of coarse structure generation from the refinement of fine details.

Nonetheless, when Stage-1 outputs lack diversity, subsequent stages tend to inherit and amplify this limited subset. Models that capture semantic–spatial information, incorporating object relations and scene structure, have achieved the most substantial nominal improvements, with some studies reporting 20–35% relative FID enhancements [28], [19].

However, these architectures are highly susceptible to noisy or weak semantic signals. In instances where textual descriptions lack sufficient structure, these models may degenerate into a limited range of plausible yet monotonous configurations [40].

Overall, architectural methods are highly effective and efficient, although their impact is closely tied to the clarity of the text, the richness of the dataset, or the upstream semantic structure.

5.2 Stability vs Flexibility in Training-Based Methods

Twelve studies have examined the impact of training dynamics on mode collapse. Curriculum learning, implemented in five of these studies, resulted in a 10–20% improvement in the FID and a 15–25% reduction in collapse frequency as complexity gradually increased [31], [34]. However, when complexity increased too rapidly, several studies reported relapses or instability in convergence, indicating that curricula must be carefully calibrated.

Adaptive training strategies, such as dynamic learning rates and balanced update schedules, resulted in 10–30% gains in FID and 5–20% gains in IS, particularly when the data was limited or unbalanced [41], [42]. These approaches, however, were among the most configuration-dependent, with poor tuning potentially causing the generator to oscillate or overshoot.

Regularization techniques, specifically gradient penalties and spectral normalization, were applied in six and three studies, respectively. Gradient penalties yielded a 15–35% improvement in FID and modest IS gains [39], but many studies observed soft collapse, cautioning that excessive regularization may prevent the generator from fully exploring latent space, leading to hard collapse—stable outputs with low diversity [7]. Spectral normalization demonstrated a similar trade-off between stability and flexibility.

Progressive growing, tested in five studies, also enhanced clarity and mitigated collapse, with an average improvement of 15–25% in the FID score [32]. However, its high computational overhead was presumed to be the main barrier to its adoption.

Overall, training-based approaches are most successful when they consider dataset size, architectural complexity, and text structure. Misconfigured schedules or overly aggressive regularization tend to introduce new failure modes.

5.3 Latent-Space and Loss-Function Techniques

Ten studies have examined a modified structural space within latent or adapted loss functions to penalize diversity. The application of mode-seeking losses, as utilized in three studies, resulted in a 10-20% increase in imitation diversity [27], [33]. However, numerous authors have identified a trade-off: an excessive emphasis on

diversity can compromise perceptual fidelity or semantic consistency.

Contrastive losses, explored in four studies, demonstrated particular efficacy when a given prompt allowed for multiple plausible interpretations. These approaches achieved 15–30% relative improvements in diversity-oriented metrics, although their effectiveness was highly contingent upon the curation of positive and negative sample pairs [43]. Latent-loss perturbation and manipulation of noise variance, evaluated in five studies, expanded the magnitude of the explored latent space and resulted in 10–25% improvements in diversity metrics [34], [44].

However, for large magnitudes, perturbation induced artifacts or semantic shifts, particularly in fine-grained text. Collectively, these findings suggest that modifications at the latent-space and loss levels are most advantageous within the context of stable architectures and regularized training. If left unchecked, these modifications risk pushing the model towards diversity at the expense of coherence.

5.4 Evaluation metrics and measuring gaps

Among the 45 studies reviewed, the Fréchet Inception Distance (FID) was utilized in 42, making it the most prevalent metric for assessing diversity and realism [3], [45]. While FID effectively captures distributional shifts, it occasionally fails to detect semantic repetition, allowing a model to appear improved even when generating numerous visually similar yet semantically identical images.

The Inception Score (IS) was applied in 25 studies, but it frequently yielded results that diverged from human perception, particularly in cases where class labels did not align with the target sampling[38].

The Multi-Scale Structural Similarity Index (MS-SSIM) was employed in 18 studies; it can capture structural similarity but may also erroneously reward repetitive structures as "consistency," potentially suppressing collapse[27].

The Learned Perceptual Image Patch Similarity (LPIPS), used in 20 studies, demonstrated a slightly higher correlation with perceived diversity, yet it remained biased towards base network selection, conflicting with various prior studies and occasionally with FID [9], [24].

In the domain of long text in image summarization, only a few studies employed a combination of quantitative measures and human evaluation, the latter being crucial for assessing semantic richness and text-image alignment.

Overall, existing metrics only partially measure semantic diversity in text-to-image tasks, underscoring the need for evaluation metrics that specifically address linguistic diversity and multimodal alignment.

5.5 Unifying Picture, Anomalies and Implications

At a cross-sectional level, the results suggest three primary conclusions.

First, the most significant performance enhancements are generally achieved when architectural design, training stabilization, and loss/latent-space manipulation are integrated. Studies that employed attention or semantic–spatial encoders in conjunction with optimized regularization and diversity-aware loss functions demonstrated the most consistently improved outcomes [7], [37], [42].

Second, anomalies were not confined to a single study. Curriculum learning occasionally resulted in exacerbated collapse when progress schedules were improperly set [34]. In certain instances, the gradient penalties were excessively strong, inhibiting diversity despite regularizing training. Additionally, some studies indicated that a larger FID did not necessarily correlate with higher semantic diversity.

Third, the difficulties observed here are well-documented in the general GAN literature, which expands upon and derives from broader observations on mode collapse. This phenomenon is inherently linked to the balance between generator–discriminator capacity (Section 5.1) and latent-space geometry, as well as evaluation blind spots [3], [31]. We focus on the task of text-to-image synthesis, where these issues are exacerbated by the additional requirement to model linguistic structure, attribute-level details, and semantic diversity.

Overall, the findings suggest that addressing mode collapse in text-to-image GANs requires a confluence of architectural, scheduling, and loss-design advancements, supported by evaluation designs sensitive to semantic diversity. Model tuning across these components, as well as

joint modeling, will likely result in stable models capable of capturing a wide variety of diverse and linguistically faithful images for a broad spectrum of textual prompts.

6. CONCLUSION AND FUTURE WORK

We conducted a systematic review of the literature on text-to-image Generative Adversarial Networks (GANs) to discern effective and ineffective strategies concerning architectural design, training, latent-space optimization, and evaluation in mitigating mode collapse. The literature consistently indicates that mode collapse remains a significant challenge in generative modeling, particularly in text-guided image generation, where models must accommodate both visual diversity and semantic variance introduced by natural language conditioning.

The collective findings of the reviewed studies suggest that no single strategy, whether architectural, algorithmic, or loss-based, can definitively prevent mode collapse. Instead, optimal improvements are achieved through the joint optimization of multiple components.

Architectural advancements incorporating attention mechanisms, multi-scale refinement, and semantic-spatial encoders have been proposed to enhance text-image correspondence, yet they remain susceptible to collapse when installations are ambiguous or training samples are limited. Training-based methods, such as curriculum learning, adaptive training schedules, and regularization techniques, offer significant stability benefits but are highly dependent on hyper-parameters and datasets. Random latent-space interpolations and diversity-promoting loss objectives can systematically increase output variability, though potentially at the cost of perceptual quality when oversampling. Finally, current evaluation methods are inadequate: standard metrics such as FID, IS, MS-SSIM, and LPIPS capture only partial information regarding the diversity or semantic alignment of samples, underscoring the need for a unified, semantically informed evaluation framework for text-to-image generation.

Collectively, these findings underscore that effectively addressing mode collapse necessitates a coordinated, multi-component approach, wherein architectural capacity and training stability are integrated with diversity-encouraging losses. It is also crucial to develop evaluation measures that accurately reflect the semantic diversity of generated outputs, an aspect that current metrics fail to convey.

The implications of this review suggest several promising areas for future research:

- **Unified Evaluation Frameworks:** There is a clear need for novel metrics that evaluate semantic diversity, linguistic grounding, and multimodal consistency, extending beyond current pixel- or feature-based criteria.

- **Hybrid Architectural Designs:** Incorporating attention, semantic-spatial reasoning, and disentangled latent-space structures may yield models capable of capturing the complexities of natural language while maintaining visual diversity.

- **Adaptive and Data-Aware Training Protocols:** Research should explore adaptive training schedules that adjust complexity, regularization strength, and learning rates based on model performance, rather than relying on fixed heuristics.

- **Benchmarking Across Text Complexity:** Future investigations should assess models under varying prompt lengths, compositional structures, and ambiguity levels to better understand collapse in more realistic environments.

- **Resource-Efficient Stability Methods:** Given the computational demands of methods such as progressive growing, it is essential to develop lighter-weight alternatives for broader applicability.

By pursuing these research directions, future work can advance the development of robust text-to-image GANs that generate not only stable but also diverse and semantically coherent images across a wide range of natural linguistic inputs.

Author Contributions:

Subuhi Kashif Ansari: Led the conceptual development of the study, formulated the core research questions, and contributed substantially to the theoretical framework. Oversaw the overall structure, coherence, and academic direction of the manuscript.

Manal Al Khammash: Contributed to the methodological design, literature review, and refinement of the analytical

approach. Assisted in synthesizing key philosophical arguments and ensuring clarity in interpretation.

Anjali Appukuttan: Supported data organization, thematic analysis, and drafting of several sections. Contributed to editing, proofreading, and maintaining consistency in academic tone and structure.

Anne Anoop: Assisted in compiling references, verifying citations, and ensuring adherence to academic formatting standards. Contributed to background research and contextual framing.

Sandeep Kumar Mathariya: Provided critical review of the manuscript, contributed to comparative analysis, and strengthened the logical flow of arguments. Assisted in refining the discussion and conclusion.

Sheela D. V.: Contributed to the philosophical analysis, especially in interpreting Krishnamurti's concepts. Assisted in drafting and revising sections related to self, consciousness, and thought.

Mohammed Saleh Al Ansari: Provided editorial oversight, contributed to structural revisions, and ensured the manuscript's alignment with academic and publication standards. Assisted in final proofreading and quality assurance.

REFERENCES:

[1] T. Xu *et al.*, "AttnGAN: Fine-Grained Text to Image Generation With Attentional Generative Adversarial Networks," 2018, doi: 10.1109/cvpr.2018.00143.

[2] H. Zhang *et al.*, "StackGAN++: Realistic Image Synthesis With Stacked Generative Adversarial Networks," *Ieee Trans. Pattern Anal. Mach. Intell.*, 2019, doi: 10.1109/tpami.2018.2856256.

[3] A. Borji, "Pros and Cons of GAN Evaluation Measures," *Comput. Vis. Image Underst.*, 2019, doi: 10.1016/j.cviu.2018.10.009.

[4] J. Mu, C. Chen, W. Zhu, S. Li, and Y. Zhou, "Taming Mode Collapse in Generative Adversarial Networks Using Cooperative Realness Discriminators," *Iet Image Process.*, 2022, doi: 10.1049/ipr2.12487.

[5] A. Odena, C. Olah, and J. Shlens, "Conditional Image Synthesis With Auxiliary Classifier GANs," 2016, doi: 10.48550/arxiv.1610.09585.

[6] W. Zhou, T. Ge, K. Xu, F. Wei, and M. Zhou, "Self-Adversarial Learning With Comparative Discrimination for Text Generation," 2020, doi: 10.48550/arxiv.2001.11691.

[7] M. Saad, M. H. Rehmani, and R. O'Reilly, "A Self-Attention Guided Multi-Scale Gradient GAN for Diversified X-Ray Image Synthesis," 2022, doi: 10.48550/arxiv.2210.06334.

[8] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning With Deep Convolutional Generative Adversarial Networks," 2015, doi: 10.48550/arxiv.1511.06434.

[9] M. Arabboev, S. Begmatov, M. Rikhsivoev, K. Nosirov, and S. Saydiakbarov, "Comprehensive Review of Image Super-Resolution Metrics: Classical and AI-based Approaches," *Acta Imeko*, 2024, doi: 10.21014/actaimeko.v13i1.1679.

[10] Y. Hong, U. Hwang, J. Yoo, and S. Yoon, "How Generative Adversarial Networks and Their Variants Work," *Acm Comput. Surv.*, 2019, doi: 10.1145/3301282.

[11] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved Training of Wasserstein GANs," 2017, doi: 10.48550/arxiv.1704.00028.

[12] A. Dash, J. C. Borges Gamboa, S. Ahmed, M. Liwicki, and M. Z. Afzal, "TAC-GAN - Text Conditioned Auxiliary Classifier Generative Adversarial Network," 2017, doi: 10.48550/arxiv.1703.06412.

[13] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved Techniques for Training GANs," 2016, doi: 10.48550/arxiv.1606.03498.

[14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation With Conditional Adversarial Networks," 2017, doi: 10.1109/cvpr.2017.632.

[15] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," 2017, doi: 10.1109/iccv.2017.244.

[16] H. Ku and M. Lee, "TextControlGAN: Text-to-Image Synthesis With Controllable Generative Adversarial Networks," *Appl. Sci.*, 2023, doi: 10.3390/app13085098.

[17] H. Wang, G. Lin, S. C. H. Hoi, and C. Miao, "Cycle-Consistent Inverse GAN for Text-to-Image Synthesis," 2021, doi: 10.1145/3474085.3475226.

[18] M. Chopra, S. K. Singh, A. Sharma, and S. S. Gill, "A Comparative Study of Generative Adversarial Networks for Text-to-Image

Synthesis," *Int. J. Softw. Sci. Comput. Intell.*, 2022, doi: 10.4018/ijssci.300364.

[19] M. M. Saad, M. H. Rehmani, and R. O'Reilly, "A Self-Attention Guided Multi-Scale Gradient GAN For Diversified X-Ray Image Synthesis," 2023, doi: 10.1007/978-3-031-26438-2_2.

[20] J. Zhang, "A Comparative Analysis and Investigation of Attn-Gan and SSA-GAN for Text-to-Image Generation," *Appl. Comput. Eng.*, 2024, doi: 10.54254/2755-2721/46/20241109.

[21] M. Zhu, P. Pan, W. Chen, and Y. Yang, "DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis," 2019, doi: 10.1109/cvpr.2019.00595.

[22] M. M. Saad, M. H. Rehmani, and R. O'Reilly, "Addressing the Intra-Class Mode Collapse Problem Using Adaptive Input Image Normalization in GAN-based X-Ray Images," 2022, doi: 10.1109/embc48229.2022.9871260.

[23] J. M. Dubinski, K. R. Deja, S. C. Wenzel, P. S. Rokita, and T. P. Trzcinski, "Selectively Increasing the Diversity of GAN-generated Samples," 2022, doi: 10.48550/arxiv.2207.01561.

[24] A. Ghosh, V. Kulharia, V. P. Namboodiri, P. H. S. Torr, and P. K. Dokania, "Multi-Agent Diverse Generative Adversarial Networks," 2018, doi: 10.1109/cvpr.2018.00888.

[25] S. Akulwa, "StyleSphere: Conversational Fashion Outfit Generator Powered by Generative AI," *Interantional J. Sci. Res. Eng. Manag.*, 2024, doi: 10.55041/ijsrem35056.

[26] R. Tominaga and M. Seo, "Image Generation From Text Using StackGAN With Improved Conditional Consistency Regularization," *Sensors*, 2022, doi: 10.3390/s23010249.

[27] Q. Mao, H.-Y. Lee, H.-Y. Tseng, S. Ma, and M. Yang, "Mode Seeking Generative Adversarial Networks for Diverse Image Synthesis," 2019, doi: 10.1109/cvpr.2019.00152.

[28] K. Wagh, O. Raul, S. Chandgadkar, S. Belanekar, and S. L. Varma, "Image Generation Using Generative Adversarial Network and Stable Diffusion," 2024, doi: 10.21203/rs.3.rs-4231306/v1.

[29] T. Liu *et al.*, "RESAKey GAN: Enhancing Color Image Encryption Through Residual Self-Attention Generative Adversarial Networks," *Phys. Scr.*, 2025, doi: 10.1088/1402-4896/ada20d.

[30] Z. Wang *et al.*, "Language models with image descriptors are strong few-shot video-language learners," *ArXiv Prepr.*, 2022, doi: 10.48550/arxiv.2205.10747.

[31] D. Saxena and J. Cao, "Generative Adversarial Networks (GANs)," *Acm Comput. Surv.*, 2021, doi: 10.1145/3446374.

[32] X. Ma, R. Jin, K.-A. Sohn, J. Paik, and T. Chung, "An Adaptive Control Algorithm for Stable Training of Generative Adversarial Networks," *Ieee Access*, 2019, doi: 10.1109/access.2019.2960461.

[33] Z. Zhang and L. Schomaker, "DiverGAN: An Efficient and Effective Single-Stage Framework for Diverse Text-to-Image Generation," *Neurocomputing*, 2022, doi: 10.1016/j.neucom.2021.12.005.

[34] M. Fathallah, M. Sakr, and S. El-etriby, "Stabilizing and Improving Training of Generative Adversarial Networks Through Identity Blocks and Modified Loss Function," *Ieee Access*, 2023, doi: 10.1109/access.2023.3272032.

[35] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," 2019, doi: 10.1109/cvpr.2019.00453.

[36] S. C. Medin *et al.*, "MOST-GAN: 3D Morphable StyleGAN for Disentangled Face Image Manipulation," *Proc. Aaai Conf. Artif. Intell.*, vol. 36, no. 2, pp. 1962–1971, 2022, doi: 10.1609/aaai.v36i2.20091.

[37] Z. Lin, A. Khetan, G. Fanti, and S. Oh, "PacGAN: The Power of Two Samples in Generative Adversarial Networks," *Ieee J. Sel. Areas Inf. Theory*, 2020, doi: 10.1109/jsait.2020.2983071.

[38] H. Eghbal-zadeh, W. Zellinger, and G. Widmer, "Mixture Density Generative Adversarial Networks," 2019, doi: 10.1109/cvpr.2019.00597.

[39] I. Boukhennoufa *et al.*, "A Novel Model to Generate Heterogeneous and Realistic Time-Series Data for Post-Stroke Rehabilitation Assessment," *Ieee Trans. Neural Syst. Rehabil. Eng.*, 2023, doi: 10.1109/tnsre.2023.3283045.

[40] Y. Zou, Y. Wang, and L. Xiao-xiang, "Auto-Encoding Generative Adversarial Networks Towards Mode Collapse Reduction and Feature Representation Enhancement," *Entropy*, 2023, doi: 10.3390/e25121657.

[41] W. Li *et al.*, "JDGAN: Enhancing Generator on Extremely Limited Data via Joint

Distribution," *Neurocomputing*, 2021, doi: 10.1016/j.neucom.2020.12.001.

[42] M. Cobbinah *et al.*, "Diversity in Stable GANs: A Systematic Review of Mode Collapse Mitigation Strategies," *Eng. Rep.*, 2025, doi: 10.1002/eng2.70209.

[43] S. M. Jafari, "Novel Generative Adversarial Network Architectures for Generating Image Data," 2024, doi: 10.32920/26052700.

[44] H. A. Golilarz, A. Azadbar, R. Alizadehsani, and J. M. Górriz, "GAN-MD: A Myocarditis Detection Using Multi-channel Convolutional Neural Networks and Generative Adversarial Network-based Data Augmentation," *Caai Trans. Intell. Technol.*, 2024, doi: 10.1049/cit2.12307.

[45] M. M. Saad, R. O'Reilly, and M. H. Rehmani, "A Survey on Training Challenges in Generative Adversarial Networks for Biomedical Image Analysis," *Artif. Intell. Rev.*, 2024, doi: 10.1007/s10462-023-10624-y.