IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Enhancing Text-to-Image Generation using Semantic Understanding

Shrinidhi R S A
Department of CSIT Jain (Deemed to be University)

Gopinath G
Department of CSIT Jain (Deemed to be University)

Abstract – Text-to-Image synthesis, image generation from text descriptions, has progressed with deep generative models. Yet, language and visual output alignment is still difficult due to ambiguity and semantic complexity. The review discusses recent approaches that improve semantic alignment with better text encoders, generative architectures, and alignment methods. Methods such as CLIP embedding's, LLM adapters, and semantically conditioned diffusion models are discussed. Models like AttnGan, Stable Diffusion, and OmniDiffusion are compared, along with main datasets, metrics, and open challenges.

Index terms – Text-to-Image synthesis, semantic understanding, generative models, diffusion, language models, multimodal alignment.

INTRODUCTION

Text-to-image synthesis is the task of finding images from natural language descriptions, falling at the crossroads of computer vision and natural language processing. It enables various applications, such as AI-augmented creativity, accessibility, and content creation. Previous methods—particularly those using Generative Adversarial Networks (GANs)—prioritized finding visually realistic images but not semantically accurate ones, usually conflating sentences such as "a red bird with black wings" and "a black bird with red wings" [1] [2]. The space has since been improved with the advent of vision-language corresponding models such as CLIP, which align the text and image modalities in the same embedding space [3]. Innovations like Latent Diffusion Models (LDMs) and autoregressive models such as DALL·E 3 have also furthered the capacity to produce fine-grained semantics and intricate compositions [4] [3]. This work presents a specific review of these advances, highlighting semantic awareness—how models deduce, encode, and decode linguistic meaning into visually sensible outputs.

LITERATURE REVIEW

Text-to-image synthesis has witnessed dramatic change, driven by the marriage of generative modeling and advanced semantic alignment methods. This part summarizes landmark contributions, mapping advancements from attention-based GANs to cutting-edge diffusion models combined with large language models (LLMs) and adaptive modular frameworks.

A. Attention-Based GANs:

One of the earliest seminal works in this area is AttnGAN [5], which uses a char-CNN-RNN text encoder and a stacked GAN generator. It presents a word-level attention mechanism that allows for fine-grained correspondence between textual hints and image regions. While AttnGAN achieves a fairly high Fréchet Inception Distance (FID) of 23.5, it was the first to introduce attention-driven generation in this field.

Based on this idea, DM-GAN [11] incorporates a dynamic memory module within the generative process. This module improves image features by selectively paying attention to uncertain textual areas at synthesis time, leading to enhanced image-text correspondence and an FID of 20.6.

B. Contrastive and Semantic Learning in GANs:

XMC-GAN [2] improves semantic fidelity by adding contrastive losses between image-caption pairs using CLIP embedding's to close the semantic gap. The model has a reduced FID of 19.8, highlighting the advantages of cross-modal pretraining and contrastive alignment.

In parallel, SD-GAN [7] uses a Siamese network structure and Semantic Conditional Batch Normalization (SCBN) to enhance paraphrasing robustness. The BiLSTM-based encoder and two-pathway architecture provide stable

generation over linguistic variations, reducing the FID to 17.4.

C. Latent Diffusion and Modular Frameworks:

The coming of Stable Diffusion [4] was a paradigm shift with the use of Latent Diffusion Models (LDMs) conditioned on CLIP embedding's. Acting in latent space allows high-quality and efficient generation of images, with FID equal to 15.2. Its modular and open-source architecture has catalyzed industry and research adoption.

D. LLM-Driven Diffusion Models:

OmniDiffusion [1] is a major breakthrough in integrating a T5-based LLM trained with adapter modules into a diffusion framework. The architecture is robust at understanding long-form and multilingual inputs, attaining an FID of 12.3 and demonstrating the importance of more profound linguistic understanding in image generation.

SD-XL [6] extends the Stable Diffusion architecture with an XL Adapter and a U-Net++ backbone, improving compositional accuracy and high-resolution generation. The model shows a higher FID of 11.7, resulting from improved prompt conditioning and structural improvements.

E. Adaptive Expert-Based Approaches:

The cutting-edge RAPHAEL model [8] integrates CLIP, LLMs, and a Mixture-of-Experts (MoE) diffusion model. It leverages semantic routing to adaptively choose expert routes depending on prompt difficulty, allowing for scalable and flexible image creation. RAPHAEL realizes the highest FID to date at 11.2, showing the strength of modular, expert-controlled designs.

I. Scope and Contributions

The primary contributions of this review are:

- Survey of semantic-aware text encoding techniques.
- Architectural review of generative architecture.
- Semantic alignment strategy discussion.
- Survey of evaluation datasets and metrics.
- Comparison and open problems.

II. Text Encoding and Semantic Representation

One of the principal pillars of good text-to-image generation is the degree to which a system can represent natural-language input as dense, descriptive semantic representations. These need to capture both the surface meaning (syntax and word-level information) and the underlying intent or context (semantics, common-sense relationships). Encoding approaches have progressed from early recurrent networks to advanced transformer-based models that rely on large-scale pertaining between vision and language modalities.

A. CLIP- Based Encoders

CLIP (Contrastive Language–Image Pre-training) [4]is now the foundation of numerous contemporary text-to-image models owing to its dual-modality learning paradigm.

It learns image and text embedding's jointly by optimizing a contrastive loss over a massive image-caption corpus. The next encoder- most often a pre-trained transformer model on a diverse set of internet-sourced captions- is a fixed length embedding aligned with visual features form a ResNet or ViT image encoder. This joint embedding space facilitates zero-shot image classifications, retrieval, and, more applicably, text-to-image conditioning. CLIP's architecture dos impose some constraints:

- Fixed Token Limit: CLIP only encodes the first 77 tokens, which restricts its capacity to handle long-form, descriptive inputs.
- Monolingual Bias: While trained on multifarious text, CLIP performs best in English and does not have strong multilingual generalization.
- Shortage of Deep Semantics: CLIP embedding's tend to favor patterns of co-occurrences rather than compositional semantics, and this can cause literal but shallowly semantic generations.

In spite of these limitations, models such as Stable Diffusion [4]and SD-XL [6] use CLIP as a backbone because of its high alignment scores and latent diffusion pipeline compatibility.

B. LLM-Powered Adapters

To counteract the stiffness of CLIP-based encoders, recent work has investigated the incorporation of Large Language Models (LLMs) as a source of dense semantic representations. Tan et al. [1] presented OmniDiffusion, a three-stage architecture where a pertained LLM (e.g., T5 or GPT) produces text embedding's, which are fed through a lightweight projection module—termed an adapter—to realign them with CLIP's visual latent space.

The advantages of this strategy are:

- Multilingual support: LLMs pre-trained on multilingual datasets are capable of processing multiple languages, dialects, and cultured allusions.
- Extended prompting: As opposed to CLIP, LLMs take long and sophisticated inputs (e.g., dialogues, directions, or paragraphs).
- Contextual Richness: LLMs embedding's capture higher-order semantics, so it becomes feasible to understand abstract or metaphorical prompts (e.g., "a solitary castle under existential skies").

Adapters are effective bridges that circumvent retraining the entire diffusion model. They add little overhead and can be adapted on task-specific datasets, allowing domain adaptions and

personalization without compromising generalization.

C. Siamese and Contrastive Text Embedding's:

Yin et al. [7] and others suggest employing Siamese networks with contrastive tasks to impose explicitly semantic consistency between paraphrased or semantically equivalent prompts. In such configurations, two identical encoders are used to process two variant descriptions of one image, e.g., "a man on a red motorcycle" and "a person is riding a crimson bike," and the embedding's are induced to be near in latent space.

This has a number of benefits:

- Prompt Robustness to Variations
- Generation Invariant to Paraphrase
- Enhanced Discrimination

This embedding's are commonly combined with semantic-conditioned normalization or discriminations that ensure alignment across input modalities to form the foundation of architectures such as SD-GAN [7].

III. Generative Architectures

The generative foundations in text-to-image synthesis decides the model's capacity to convert semantic embedding's into high-fidelity, coherent images. Architectural design over the years has shifted from GAN-based models centered around realism, to diffusion-based models prioritizing precision and stability, and lately, to hybrid pipelines blending strengths from more than one paradigm. This section discusses important categories: GANs, diffusion models, and hybrid approaches.

A. GAN-Based Models

Generative Adversarial Networks (GANs) [9] were some of the first architectures to be used for text-to-image synthesis. A GAN comprises a generator that generates fake images and a discriminator that learns to distinguish between real and fake samples. The generator, conditioned on text, takes in both text embedding's and noise, generating images consistent with the semantic hints.

1. AttnGAN: This was the first model that incorporated word-level attention mechanisms, enabling the generator to attend to certain words when generating various regions of an image. It employed a multi-stage generation framework, improving coarse to fine-grained outputs.
   Advantages: Improved object localization, explainable attention maps.
   Disadvantages: Restricted resolution (e.g., 256*256), comparatively unstable training.
2. DM-GAN [8]: Enhance AttenGAN with a dynamic memory module that stores and recovers sematic features during refinement processes. The memory enables the generator to re-strongly emphasize important textual details lost during initial generation.
   Innovation: Memory gating improves semantic coherence during refinement.
   Limitation: Memory overhead and increased training time.
3. XMC-GAN (Contrastive GAN) [2]: the model uses contrastive losses between image-image and caption-caption pairs, forcing images produced form paraphrased prompts to be semantically close.
   Strength: Greater semantic alignment compared to conventional GANs
   Challenge: Difficulty of contrastive training and stability in multi-model optimization.

B. Diffusion-Based Models

Diffusion models are a newer generation of generative models that generate images by iteratively by denoising Gaussian noise, with guidance from conditioning inputs like text embeddings.

1. DDPMs (Denoising Diffusion Probabilistic Models) [10]: These models invert a stationary Markov process that progressively adds noise to an image. While accurate, they are computationally costly because of the lengthy sampling chains (typically 1000+ steps).
   Pros: High image quality, stable training.
   Cons: slow inference and costly compute requirements.
2. Latent Diffusion Models (LDMs) [4]: These save computation by working in latent space that is compressed (with auto encoders such as VQ-GAN). Generation is done in latent space and then decoded to pixel space.
   Efficiency: Orders of magnitude faster than pixel-space diffusion.
   Integration: Can be conditioned easily with CLIP or LLM- based embedding's.
3. Stable Diffusion [4]: One of the most commonly used LDMs, integrating CLIP

C. Hybrid and Emerging Models

Since GANs and diffusion models both possess strengths and limitations, there have been various emerging models which hybridize characteristics of both to design more solid systems.

1. DALL-E 3 [3] an OpenAI hybrid pipeline employing autoregressive decoding for coarse image layout and diffusion sampling for fine-tuning. Deep Integration with LLMs such as GPT-4 allows conversational prompting and edit ability.

Key Benefit: Strong alignment with human language; support for image editing.

Use Case: Interactive generation and semantic editing.

2. Stable Diffusion XL (SD-XL) [6]: Augments stable Diffusion with a bigger CLIP encoder, several conditioning pathways, and architectural improvements. It produces state-of-the-art resolution and realism and allows for longer prompts.

   Advantages: Flexible conditioning, better compositionality

   Disadvantages: Increased model size and training requirement.

3. RAPHAEL [8]: Adds a mixture-of-Experts (MoE) mechanism to diffusion pipelines, allowing for specialized modules to process various aspects of image synthesis, e.g., texture, object layout, and style.

   Innovation: Scalable capacity with expert modularity

   Research Frontier: Requires expert routing mechanisms and balanced training.

These generative strategies each reflect trade-offs between realism, semantic control, training complexity, and inference speed. Table 1 below outlines side-by-side the different architecture types and their mechanisms for semantic integration.

IV. Semantic Alignment Techniques

Semantic alignment refers to ensuring that the generated image faithfully reflects the meaning of the input text prompt- no just visually, but conceptually and contextually. Even high-quality images can fail to convey the correct semantics if the model lacks an understanding of linguistic nuances, especially under paraphrased, long-form, or multilingual prompts. This section examines key strategies developed to improve alignment between text and visual outputs in text-to-image synthesis models.

A. Adapter-Based Alignment

Adapter modules are lightweight neural layers introduced between pre-trained large language models (LLMs) and the image synthesis backbone. Their role is to map rich language embedding's (e.g., from T5 or GPT variants) into the conditioning space compatible with visual generation frameworks such as Stable Diffusion.

- OmniDiffusion [1] is a prime example: it replaces CLIP's limited text encoder with a full LLM, followed by an adapter that projects high-dimensional textual representations into CLIP's latent visual space.
- Adapters offer modular extensibility, enabling fine-tuning on domain-specific data (e.g., medical, multilingual) without retraining the base model.
- Benefits:
  - Extend beyond CLIP's 77-token limit.
  - Retain context from longer, structured prompts.
  - Support for multiple languages and fine-grained modifiers.
- Limitations:
  - Alignment relies on adapter calibration, which can drift if not carefully tuned

B. Semantic-Conditioned Batch Normalization (SCBN)

Semantic-conditioned normalization techniques incorporate text features directly into the intermediate stages of the image synthesis process via modulated normalization layers.

- SCBN (Semantic- Conditioned Batch Normalization) modifies batch norm layers so that the scale (γ) and Shift (β) parameters are functions of textual embedding's.
- First proposed in [7], this technique allows the generator to dynamically alter image features (e.g., texture, colour, object presence) in response to linguistic content.
- Later variations include Instance Norm (AdaIN) or Layer Norm adaptations, enabling per-sample modulation.
- Advantages:
  - Improved semantic style transfers from text to image.
  - Fine-grained, localized control over visual details (e.g., bird colour, object position).
- Applications:
  - Used in SD-GAN [7] and other GAN-based pipelines where alignment is challenging under high visual variation.
- Trade-offs:
  - Requires carefully tuned normalization schemes to prevent mode collapse or visual artefacts.

C. Contrastive Learning for Semantic Consistency

Contrastive learning aims to bring semantically similar items closer in embedding space while pushing dissimilar ones apart. In text-to-image generation, this technique helps ensure consistent image outputs for semantically equivalent prompts.

- XMC-GAN [2] introduced caption–caption and image–image contrastive losses, treating paraphrased captions or generated images from the same prompt as positive pairs.

- Later models (e.g., SD-GAN [7]) applied Siamese networks to learn consistent representations across diverse input prompts, using shared weights and contrastive losses.
- Implementations:
  - Use cosine similarity or InfoNCE loss between text/image embedding's.
  - Enforce invariance under paraphrasing or rewording.
- Benefits:
  - Increases robustness to linguistic variation.
  - Reduces generation inconsistency from prompt phrasing differences.
- Limitations:
  - Requires large datasets of paraphrased text–image pairs.
  - Contrastive learning may interfere with adversarial losses if not balanced properly.

II. Datasets and Evaluation Metrics

Robust evaluation of text-to-image generation models requires diverse, high-quality datasets and reliable metrics that reflect both visual fidelity and semantic alignment. However, most existing metrics focus more on image quality than on semantic alignment. However, most existing metrics focus more on image quality than on semantic accuracy, and datasets often vary in complexity, domain coverage, and annotation richness. This section provides a detailed overview of commonly used datasets and evaluation protocols.

A. Common Datasets

1. CUB-200-2011 [12]:
- Domain: Fine-grained bird species
- Images: ~ 11,800 images of 200 bird species
- Captions: 10 per image, human-annotated
- Use Case: Fine-grained attribute generation (e.g., wing colour, beak length)
- Limitations: Domain-specific, lacks scene diversity

2. MS-COCO (Microsoft Common Objects in Context) [13] :
- Domain: Everyday scenes with multiple objects
- Images: ~330,000 images, 80 object categories
- Captions: 5 per image, open-vocabulary descriptions
- Use Case: General-purpose generation and captioning
- Advantages: Rich scene diversity, multiple captions enable paraphrasing tests

3. Laion-5B [4]:
- Domain: Broad, web-scraped images and captions
- Size: Over 5 billion image-text pairs
- Language Support: Multilingual
- Use Case: Large-scale pertaining for models like Stable Diffusion
- Benefits: Scale enables robust generalization
- Caveats: Noisy annotations, inconsistent caption quality, biases

4. Other Specialized Datasets:
- FashionGen (clothing-focused), Oxford-102 Flowers, PaintSkills (fine-grained skill control)
- Useful for domain adaptation and task-specific tuning

B. Evaluation Metrics

While high-resolution, realistic images are important, true success in text-to-image generation hinges on how well the image reflects the input prompt. Here are key metrics:

1. FID (Fréchet Inception Distance) [14]:
   - Measures distance between distributions of real and generated image features
   - Lower is better; FID < 15 often considered high quality
   - Strength: Sensitive to realism and diversity
   - Limitation: Not language-aware; doesn't assess semantic correctness
2. IS (Inception Score) [15]:
   - Captures both confidence and diversity of generated images using a pretrained classifier
   - Higher is better; but performance can be inflated by generic-looking outputs
   - Weakness: No grounding in text; doesn't evaluate text-image match
3. R-precision [5]:
   - Measures retrieval accuracy of matching the correct caption from a pool of candidates given a generated image
   - Reflects how well the image aligns with its intended text
   - Limitation: Can be misleading if captions are too similar or ambiguous
4. CLIP-Score [16]:
   - Uses CLIP embeddings to compute cosine similarity between text and image representations
   - Reference-free evaluation: doesn't need ground truth images
   - Advantages: Correlates well with human judgment, scalable

- Weakness: Inherits CLIP's biases; may favor literal matches over conceptual understanding

5. Human Evaluation:
   - Human annotators assess generated images for relevance, aesthetics, and coherence
   - Most reliable, but expensive and slow
   - Often used as the gold standard in large-scale model comparisons (e.g., DALL·E 3, Midjourney)

VI. Comparative Analysis

Comparing text-to-image generation models involves understanding trade-offs between architectural design, semantic alignment techniques, computational efficiency, and output quality. This section presents a structured analysis of key methods, highlighting how semantic understanding contributes to improved generation performance.

A. Method Comparison Overview

The following table synthesizes critical attributes of representative models based on encoder type, generation architecture, semantic alignment strategy, and FID score (where available).

Table . Comparative Summary of Key Text-to-Image Generation Models

Method	Text Encoder	Generator Type	Semantic Strategy	FID ↓
AttnGAN	char-CNN-RNN	Stacked GAN	Word-level attention	23.5
DM-GAN	RNN + Memory Module	GAN	Dynamic memory refinement	20.6
XMC-GAN	CLIP	GAN	Contrastive caption–image loss	19.8
SD-GAN	BiLSTM + SCBN	Siamese GAN	Siamese net + semantic BN	17.4
Stable Diffusion	CLIP	Latent Diffusion	CLIP conditioning	15.2
OmniDiffusion	LLM (T5) + Adapter	Diffusion	Adapter-based LLM alignment	12.3
SD-XL	CLIP + XL Adapter	LDM (U-Net++ XL)	Enhanced prompt conditioning	11.7
RAPHAEL	CLIP + LLM + MoE	MoE Diffusion	Mixture-of-Experts + semantic routing	11.2

Note: FID values vary based on dataset and evaluation protocols; shown here for comparison using MS-COCO unless otherwise stated.

B. Semantic Fidelity vs. Image Realism

The best-performing models (e.g., SD-XL, RAPHAEL) excel at both semantic accuracy and visual realism. Their success is attributed to:

- Rich text encoders (CLIP + LLM fusion)
- Fine-grained conditioning mechanisms (e.g., adapters, normalization)
- Scalable diffusion processes with architectural enhancements

Meanwhile, older GAN-based systems lag in capturing abstract or complex scenes, despite faster inference.

VII. Challenges and future Directions

Despite remarkable progress in text-to-image generation, several persistent challenges hinder its full potential, especially regarding semantic understanding, controllability, and scalability. This section highlights major research hurdles and outlines promising directions for future investigation.

1. Long-Form and Multilingual Prompts

Challenge: Most models are limited by text encoders like CLIP, which support a maximum of 77 tokens and are primarily trained on English datasets. This constraint restricts the system's ability to handle descriptive, narrative, or multilingual prompts.

Future Direction:

- Integration of LLMs (e.g., T5, GPT) with cross-lingual capabilities allows support for extended and multilingual prompts.
- Adapter-based alignment modules [1] can bridge the semantic gap between long-text embedding's and image-generating modules without retraining the entire pipeline.
- Building multilingual datasets (e.g., LAION-X) with aligned captions is also essential.

2. Fine-Grained Attribute Control

Challenge: Generating specific attributes (e.g., "a red apple on a blue plate under sunlight") or modifying parts of an image (e.g., changing only the background) is difficult for many models. Current conditioning techniques struggle with localized control.

Future Direction:

- Use of **segmentation-aware guidance**, **spatial attention**, and **region-based masks** can improve localized generation.
- Incorporation of **language-guided editing** and **interactive user inputs** can enable real-time refinement (e.g., inpainting or outpainting with prompt updates).
- Layered or compositional generation approaches (e.g., ControlNet) are promising for structured control.

3. Evaluation Standardization

Challenge: Existing evaluation metrics like FID and IS primarily measure image realism and diversity, but do not fully capture semantic alignment or task relevance. Human evaluations are costly and inconsistent.

Future Direction:

- Development of reference-free, text-image alignment metrics (e.g., CLIP-Score [16]) that better reflect human judgments.
- Task-specific benchmarks (e.g., story-to-comic generation, logo synthesis) could enable more targeted evaluations.
- Hybrid protocols combining automated and crowdsourced assessments could provide scalable, robust evaluations.

4. Ethical and Bias Considerations

Challenge: Models trained on web-scale datasets often inherit and amplify social, gender, or racial biases. They may also generate offensive or misleading content if not properly controlled.

Future Direction:

- Curating diverse, balanced training datasets and integrating bias detection tools in the pipeline.
- Implementing prompt filters, content safety classifiers, and user consent layers before synthesis.
- Explainable AI techniques can improve transparency and traceability of model decisions.

5. Resource Efficiency

Challenge: High-performing models (e.g., SD-XL, RAPHAEL) require massive compute for training and inference, making them inaccessible to smaller labs or real-time applications.

Future Direction:

- Model distillation, quantization, and sparsity-based pruning can reduce size and latency.
- Designing modular and adaptive architectures that activate only necessary submodules per prompt (e.g., MoE in RAPHAEL [8]).
- Leveraging edge-compatible inference solutions for deployment on mobile and low-power devices.

Summary of Open Research Areas:

Challenge	Research Needs
Prompt length and language	Multilingual LLM adapters, cross-lingual pretraining
Fine-grained control	Interactive editing, semantic masking, region-level attention
Evaluation	CLIP-based scores, human-aligned and task-aware metrics
Ethical synthesis	Dataset filtering, bias audits, explainability
Computational scalability	Distillation, low-rank optimization, mixture-of-experts routing

VIII. Challenges and future Directions

Text-to-image synthesis has developed substantially, from initial GAN-based approaches that yielded low-resolution and semantically constrained results to diffusion models that can produce high-quality and contextually rich images. Of central importance among these developments is semantic understanding's integration, which allows more precise alignment between sophisticated linguistic input and generated visual output.

This review emphasized the significance of semantic alignment and representation, exemplifying techniques like CLIP-based encoders, LLM-integrated architectures such as OmniDiffusion, and adapter models that enable efficient multimodal fusion with fewer retraining needs. Latent diffusion models (LDMs) have become prevalent architecture because of their scalability and output quality, whereas hybrid systems like DALL-E 3 illustrate the strengths of blending semantic conditioning with autoregressive and diffusion methods.

In spite of all these developments, issues persist in processing long prompts, providing fine-grained control, guaranteeing ethical content generation, and constructing standardized semantic evaluation metrics. Tighter LLM integration, modular controllable architectures, human-aligned evaluation procedures, and responsible deployment methodologies need to be the focus of future studies. Semantic understanding will remain the foundation for constructing reliable, creative, and intelligent generative systems.

ACKNOWLEDGEMENTS

We would like to thank Prof. Yashaswini for proofreading the paper.

REFERENCES

[1] Z. Tan et al., "An Empirical Study and Analysis of Text-to-Image Generation Using LLM-Powered Textual Representation," Proc. NeurIPS, 2024.

[2] H. Ye et al., "Improving Text-to-Image Synthesis Using Contrastive Learning," ICLR, 2021.

[3] A. Ramesh et al., "Hierarchical Text-Conditional Image Generation with CLIP Latents," arXiv preprint arXiv:2212.11962, 2023.

[4] R. Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models," CVPR, 2022.

[5] P. Xu et al., "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks," CVPR, 2018.

[6] T. Ho et al., "Stable Diffusion XL: Scaling Latent Diffusion Models for High-Resolution Synthesis," 2024.

[7] G. Yin et al., "Semantics Disentangling for Text-to-Image Generation," ECCV, 2019.

[8] X. Xu et al., "DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis," CVPR, 2019.

[9] J. Ho et al., "Denoising Diffusion Probabilistic Models," NeurIPS, 2020.

[10] Z. Xue et al., "RAPHAEL: Text-Conditional Image Diffusion via Mixture-of-Experts," ICLR, 2024.

[11] K. Patel & P. Shah, "Semantic-aware Mapping for Text-to-Image Synthesis," IJSEM, vol. 10, no. 2, 2025.

[12] P. Wah et al., "The Caltech-UCSD Birds-200-2011 Dataset," Technical Report, 2011.

[13] T.-Y. Lin et al., "Microsoft COCO: Common Objects in Context," ECCV, 2014.

[14] M. Heusel et al., "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," NeurIPS, 2017.

[15] T. Salimans et al., "Improved Techniques for Training GANs," NeurIPS, 2016.

[16] O. Mokady et al., "CLIP-Score: A Reference-Free Evaluation Metric for Image Captioning," ECCV, 2022.