Literature Review & Elaboration of Problem
Argu Mint: Picture the Point

Version: 1.2

Project Code	F25-207
Internal Supervisor	Zain-ul-Hassan, Nouman Durrani
External Supervisor	
Project Manager	Syed Muhammad Saad Manzoor
Project Team	Arham Hussain (22K-4080) Aarib Ahmed (22K-4004) Partham Kumar (22K-4079)
Submission Date	December 01, 2025

Document History

Version	*Name of Person*	*Date*	*Description of change*
1.0	Partham Kumar	November 27, 2025	Document Created
1.0	Partham Kumar	November 28, 2025	Introduction Section Completed
1.1	Arham Hussain	November 29, 2025	Abstract and Project Overview
1.1	Arham Hussain	November 29, 2025	Literature review Section Completed
1.2	Aarib Vahidy	November 30, 2025	Document reviewed and finalized

Distribution List

Name	***Role***
Zain ul Hassan	Internal Supervisor
Syed Muhammad Saad Manzoor	Project Manager
Nouman Durrani	Internal Supervisor

Document Sign-Off

Version	*Sign-off Authority*	*Project Role*	*Sign-off Date*

Table Of Contents

1. Abstract

Images serve as powerful tools for conveying meaning and enhancing understanding, while arguments are often perceived as formal and logical constructs. This research project, titled **"Argu Mint: Picture the Point,"** explores the intersection of these domains by developing a text-to-image generation system capable of producing contextually relevant images from textual arguments. The project addresses the Touché 2025 "Image Retrieval/Generation for Arguments" challenge at CLEF, where the task is to generate images that effectively illustrate the central aspects of an argument's claim.

Unlike traditional text-to-image tasks that work with descriptive prompts, argument illustration presents a unique challenge: arguments in our domain consist of single claims without supporting premises (e.g., "Boxing leads to serious injuries" or "Celebrity culture distorts personal values"). The system must interpret the semantic meaning of abstract claims and generate visually compelling images that convey their essence. Our methodology involves fine-tuning Stable Diffusion 3.5 Medium on the CLEF argument-image dataset, leveraging triple text encoder architecture (dual CLIP and T5) for enhanced semantic understanding of argumentative language. The expected outcomes include a specialized T2I model capable of generating contextually appropriate illustrations for arguments, evaluated using NDCG metrics based on manual assessment of key aspect representation.

2. Background and Justification

2.1 Evolution of Text-to-Image Generation

The field of text-to-image generation has witnessed remarkable progress over the past decade, transitioning from domain-specific generative adversarial networks (GANs) to general-purpose transformer and diffusion-based models. Early approaches such as StackGAN and AttnGAN demonstrated the feasibility of generating images from text but were constrained to narrow domains and suffered from artifacts and limited resolution [1][2].

The introduction of DALL-E by OpenAI in 2021 marked a paradigm shift, demonstrating that scaling transformer architectures to 12 billion parameters and training on 250 million image-text pairs could achieve impressive zero-shot generation capabilities [3]. This was followed by significant advances in latent diffusion models (LDMs), which reduced computational requirements by operating in compressed latent spaces rather than pixel space [4].

Current state-of-the-art models including Stable Diffusion 3.5, PixArt-Σ, and HiDream have pushed the boundaries of resolution, prompt adherence, and generation efficiency [5][6][7]. However, these models are primarily trained on descriptive image captions rather than abstract argumentative text, creating a significant domain gap for argument illustration tasks.

2.2 The Argument Illustration Challenge

The Touché 2025 shared task on "Image Retrieval/Generation for Arguments" at CLEF (Conference and Labs of the Evaluation Forum) presents a novel challenge: given an argument consisting of a single claim without supporting premises, systems must find or generate images that help convey central aspects of the argument's claim. This task bridges computational argumentation and visual communication, areas that have traditionally been studied in isolation.

In contrast to traditional text-to-image generation, which typically involves prompts that depict specific visual scenarios (for instance, "a sunset over mountains"), argument illustration necessitates the interpretation of abstract assertions such as "Excessive homework diminishes student well-being" or "Energy drinks interfere with natural sleep patterns" and their conversion into significant visual depictions. The difficulty resides in encapsulating the core of an argument through the imagery produced.

2.3 Project Justification

Argu Mint: Picture the Point addresses this research gap by developing a specialized text-to-image generation system fine-tuned for argument illustration. Our approach builds upon Stable Diffusion 3.5 Medium, leveraging its advanced architecture while adapting it for the unique requirements of argumentative text:

1. Domain-Specific Fine-Tuning: Training on 14,000 argument-image pairs from the CLEF dataset to bridge the gap between argumentative language and visual representation.

2. Enhanced Semantic Understanding: Utilizing the SD3 triple text encoder architecture (dual CLIP and T5 encoders) for comprehensive interpretation of abstract claims.

3. Contextual Generation: Employing classifier-free guidance and flow matching scheduling to generate images that capture key aspects identified in the evaluation criteria.

The practical applications of this research extend to educational technology (visualizing debate topics), journalism (illustrating opinion pieces), legal documentation (representing case arguments), and persuasive communication design.

3. Problem Statement

3.1 Core Problem Definition

The fundamental problem addressed by Argu Mint: Picture the Point is the generation of contextually relevant images from textual arguments that consist of single claims without supporting premises. Unlike traditional text-to-image generation tasks where input prompts describe visual scenes explicitly, argument illustration requires the system to:

1. Interpret Abstract Claims: Arguments such as "Celebrity culture distorts personal values" or "Urban noise pollution damages mental calm" do not directly describe visual content. The system must infer appropriate visual representations from abstract concepts.

2. Capture Key Aspects: According to the Touché 2025 evaluation criteria, generated images must represent the central aspects of an argument's claim. Each argument has two identified key aspects that should be visually conveyed.

Preserve Semantic Integrity: The produced image should faithfully represent the position and implications of the argument, rather than simply illustrating associated but peripheral ideas.

3.2 Technical Challenges

3.2.1 Domain Gap Between Descriptive and Argumentative Text

Existing T2I models are predominantly trained on descriptive image captions from datasets like MS-COCO, LAION, and similar corpora. These captions describe what is visible in an image (e.g., "A dog running in a park"). Argumentative text, however, expresses opinions, claims, and value judgments that do not directly map to visual elements. This domain mismatch results in poor performance when state-of-the-art models are applied directly to argument illustration without fine-tuning.

3.2.2 Semantic Alignment with Abstract Concepts

Large-scale T2I models frequently struggle with attribute binding and compositional understanding [8][17]. For argument illustration, the challenge is amplified because:

- Claims often involve abstract concepts ("values," "well-being," "mental calm")

- The visual metaphors required are not explicitly stated

- Multiple valid interpretations may exist for a single argument

3.2.3 Evaluation Complexity

Unlike standard image generation tasks evaluated through automated metrics (FID, CLIP Score), argument illustration requires assessment of whether generated images effectively convey the argument's key aspects. The Touché challenge employs manual evaluation using NDCG (Normalized Discounted Cumulative Gain) based on human judgment of aspect representation, making optimization and validation more complex.

3.2.4 Quality-Alignment Trade-off

Research has demonstrated that fine-tuning models for improved semantic alignment using human feedback can significantly reduce image quality, producing oversaturated or non-photorealistic outputs [10]. For argument illustration, both aspects are critical, images must be visually compelling while accurately representing the claim.

3.2.5 Limited Training Data

The CLEF argument-image dataset provides a relatively small corpus compared to the billions of image-text pairs used to train foundation models. Effective fine-tuning with limited domain-specific data while preventing overfitting and maintaining generalization is a significant challenge.

3.3 Research Questions

This project seeks to address the following research questions:

1. RQ1: How effectively can diffusion-based T2I models be fine-tuned to generate contextually appropriate images from argumentative claims?

2. RQ2: What architectural modifications or training strategies improve semantic alignment between abstract arguments and generated visual content?

3. RQ3: How can the quality-alignment trade-off be balanced to produce images that are both visually appealing and semantically accurate representations of arguments?

3.4 Scope and Constraints

The project operates within the following scope:

- Input: Single-claim arguments without supporting premises (as defined by the Touché 2025 task)

- Output: Five ranked images per argument for submission

- Base Model: Stable Diffusion 3.5 Medium (selected for its triple encoder architecture and accessibility)

- Training Data: CLEF argument-image dataset (14,000+ samples)

- Evaluation: NDCG based on manual assessment of key aspect representation and MAP for training evaluation

4. Literature Review

This section provides a comprehensive review of the literature on text-to-image generation, organized thematically to trace the evolution of architectures, methodologies, and capabilities.

4.1 Foundational Autoregressive Approaches

4.1.1 Zero-Shot Text-to-Image Generation (DALL-E)

Ramesh et al. [3] introduced DALL-E, a groundbreaking two-stage approach for zero-shot text-to-image generation. The architecture employs a discrete variational autoencoder (dVAE) to compress 256×256 RGB images into 32×32 grids of image tokens, followed by a 12-billion parameter autoregressive transformer that models the joint distribution of text and image tokens. Trained on 250 million image-text pairs collected from the internet, DALL-E demonstrated remarkable generalization capabilities, generating high-quality images on datasets like MS-COCO without task-specific training.

The model's key innovation lies in treating text-to-image generation as a sequence modeling problem, enabling the application of transformer architectures that had proven successful in natural language processing. The reranking strategy, which selects the best image from 32 generated samples using CLIP, further improved output quality. However, the initial tokenization process can lose high-frequency details, and the model exhibits inconsistent performance on compositional tasks such as combining novel concepts (e.g., "a hedgehog wearing a Christmas sweater").

4.1.2 CogView: Transformer-Based Generation

Ding et al. [12] presented CogView, a 4-billion parameter transformer model utilizing a Vector Quantized Variational Autoencoder (VQ-VAE) tokenizer. The model demonstrated scalability and flexibility in fine-tuning for various downstream tasks. However, the autoregressive generation process results in slower inference compared to diffusion-based alternatives, and the initial 256×256 resolution images are compressed to 32×32 tokens, potentially causing blurriness. The training on 30 million Chinese text-image pairs also introduced language and cultural biases in the generated content.

4.1.3 Parti: Pathways Autoregressive Text-to-Image

Yu et al. [11] developed Parti, which conceptualizes image generation as a translation task from text tokens to image tokens. The model excels at content-rich, complex image generation with strong prompt alignment and features a modular architecture facilitating component-wise improvements. However, the autoregressive bottleneck results in slow generation, and the model requires substantial computational resources (TPUv4 clusters), limiting accessibility for research and practical applications.

4.2 Diffusion-Based Architectures

4.2.1 Latent Diffusion Models (Stable Diffusion)

Rombach et al. [4] introduced Latent Diffusion Models (LDMs), representing a significant advance in computational efficiency for diffusion-based generation. The two-stage approach first trains a VAE to compress images into a lower-dimensional latent space, then applies the diffusion process in this efficient representation. A cross-attention mechanism enables flexible conditioning on text prompts, bounding boxes, or other modalities.

LDMs achieved state-of-the-art results on image inpainting and class-conditional synthesis while dramatically reducing training and inference costs. The approach allows the computationally expensive autoencoder to be trained once and reused across multiple diffusion model configurations. However, the compression into latent space can result in loss of fine-grained details, and performance is sensitive to the downsampling factor employed.

4.2.2 GLIDE: Text-Guided Diffusion Models

Nichol et al. [13] presented GLIDE, comparing CLIP guidance and classifier-free guidance for text-conditioned diffusion. The study found that classifier-free guidance produces more photorealistic and caption-aligned images. Despite being smaller than DALL-E, GLIDE achieved superior image quality and demonstrated strong generalization to complex prompts. The primary limitation is slower inference compared to GAN-based approaches, and the lack of publicly available code has limited community adoption and fine-tuning research.

4.2.3 Hierarchical Text-Conditional Generation (unCLIP/DALL-E 2)

Ramesh et al. [8] introduced unCLIP, a two-stage model that first generates CLIP image embeddings from text using an autoregressive prior, then decodes these embeddings into images using a diffusion decoder. The approach achieves improved image diversity compared to systems like GLIDE, with an impressive FID score of 10.39. The CLIP embedding space enables zero-shot semantic modification of images based on text descriptions.

However, the model struggles with attribute binding—associating specific attributes with distinct objects—and cannot reliably generate correctly spelled text within images. The decoder's reliance on 64×64 base resolution limits detail in complex scenes, suggesting avenues for future improvement.

4.3 Diffusion Transformers and Recent Advances

4.3.1 PixArt-Σ: Weak-to-Strong Training

Chen et al. [5] developed PixArt-Σ, a Diffusion Transformer (DiT) capable of directly generating 4K resolution images. The model employs a weak-to-strong training strategy, leveraging pretrained PixArt-α weights for efficient training on higher-quality data. An efficient token compression framework enables high-resolution synthesis while maintaining a relatively small parameter count of 0.6 billion—considerably smaller than models like SDXL.

PixArt-Σ excels at following complex, long text prompts through improved captioning using Share Captioner instead of LLAVA. Despite these advances, the model still struggles with text generation within images and hand/face rendering. Future research directions include enhanced prompt handling, improved face generation, and security considerations for bias mitigation.

4.3.2 HiDream-I1: Sparse Diffusion Transformer

Cai et al. [7] introduced HiDream-I1-Fast, a lightweight variant of a 17-billion parameter sparse diffusion transformer utilizing Mixture-of-Experts (MoE) for efficient generation. The model is optimized for speed and reduced hardware requirements, enabling high-quality synthesis on single-GPU setups while supporting instruction-based image editing through HiDream-E1.

The open-source availability of weights and code facilitates adaptation and research. However, the fast variant exhibits slightly lower fidelity compared to full versions, and integration of editing pipelines requires careful setup. The architecture presents opportunities for fine-tuning on specialized tasks such as argument-to-image mapping.

4.3.3 MMaDA: Multimodal Diffusion Language Models

Yang et al. [14] presented MMaDA, a unified 8-billion parameter diffusion architecture designed for text-to-image generation, multimodal understanding, and textual reasoning. The model employs a modality-agnostic diffusion framework with chain-of-thought fine-tuning and a unified reinforcement learning objective (UniGRPO).

MMaDA outperforms strong baselines including SDXL and Janus on text-to-image benchmarks while maintaining open-source accessibility. The unified architecture may require careful tuning for purely image-focused tasks, but presents potential for adaptation to argument illustration applications requiring integrated reasoning and generation capabilities.

4.4 Controllability and Grounding Approaches

4.4.1 Visual Programming for Generation and Evaluation

Cho et al. [15] introduced VPGEN and VPEVAL, frameworks utilizing visual programming for T2I generation and evaluation. VPGEN implements a three-step process: an LLM generates objects and counts, creates layouts with bounding boxes, and guides a T2I model for final image generation. VPEVAL employs an LLM to generate evaluation programs calling specialized visual modules for scoring and explanation.

These frameworks provide superior control over object counts, spatial relationships, and scales while offering enhanced interpretability through step-by-step generation. The LLM-based approach handles unseen objects beyond training distributions. However, the multi-step pipeline introduces complexity and potential failure points, and output quality depends on underlying layout-to-image model performance.

4.4.2 ObjectDiffusion: Grounded Generation

Süleyman and Biricik [9] developed ObjectDiffusion, grounding T2I diffusion models through explicit layout inputs. Users control object placement and appearance via bounding boxes and object names alongside text prompts. The model achieves high-precision control over spatial relationships and supports open-vocabulary generation, outperforming existing models on controllable generation metrics.

The approach addresses critical controllability requirements for practical applications but requires additional user inputs, increasing workflow complexity compared to zero-shot models. The two-stage architecture adds computational overhead but enables the precise control necessary for structured image synthesis.

4.5 Personalization and Alignment

4.5.1 Imagine Yourself: Tuning-Free Personalization

The Meta GenAI team [16] introduced Imagine Yourself, a tuning-free personalization framework for diffusion-based generation. The approach achieves better prompt alignment for complex instructions without requiring model fine-tuning for individual subjects. However, performance on complex poses and unusual prompt combinations remains imperfect. The framework primarily addresses human subjects, with extension to non-human subjects (pets, art styles, objects, fictional characters) remaining underexplored.

4.5.2 Aligning Models with Human Feedback

Lee and Liu [10] developed a three-stage fine-tuning method using human feedback to enhance text-to-image alignment. Building on the success of RLHF techniques in language models like GPT-3, the approach achieved up to 47% improvement in image-text alignment, generating objects with specified colors, counts, and backgrounds more accurately than pretrained models.

A critical finding is that naive fine-tuning with human feedback can significantly reduce image quality, producing oversaturated or non-photorealistic outputs with reduced diversity. This quality-alignment trade-off represents a fundamental challenge for practical deployment of aligned T2I models.

4.6 Semantic Enhancement with Large Language Models

4.6.1 LLM4GEN: Leveraging LLM Representations

Liu et al. [17] proposed LLM4GEN, a plug-and-play component improving T2I generation from complex instructions. The Cross-Adapter Module (CAM) combines LLM text understanding with standard T2I text encoders, while entity-guided regularization loss ensures correct attribute-object binding.

The approach addresses the limited capacity of CLIP-based text encoders for intricate semantic details and achieves strong performance across benchmarks including FID scores. Unlike previous methods requiring separate LLMs or vast training data, LLM4GEN uses only 10 million data points for efficient training. The introduction of DensePrompts addresses the lack of comprehensive benchmarks for long, detailed descriptions.

4.7 Unified Multimodal Architectures

4.7.1 Chameleon: Mixed-Modal Foundation Models

The Chameleon Team [18] developed a family of early-fusion, token-based mixed models processing arbitrary sequences of text, images, and code. The fully token-based representation enables a unified transformer architecture where images and text are converted to tokens and processed together.

The approach demonstrates strong performance across tasks when trained on large interleaved datasets with stability enhancements. However, optimization challenges arise from different distributions of text and image tokens, causing training instability. Image tokenization (1024 tokens from an 8192-size codebook) creates scalability concerns with increased input lengths affecting training and inference speed.

4.8 Research Gaps and Future Directions

The literature review reveals several persistent research gaps:

1. Attribute Binding: Accurately associating specific attributes with corresponding objects remains challenging

across all architectures [8][17].

2. Text Rendering: Generating correctly spelled text within images is a common failure mode requiring specialized attention [5][8].
3. Quality-Alignment Balance: Improving semantic alignment through fine-tuning often degrades image quality [10].
4. Computational Accessibility: Many state-of-the-art models require substantial resources, limiting research accessibility [11][12].
5. Specialized Applications: Adapting general T2I models for domain-specific tasks such as argument illustration requires further investigation [14][7].
6. Evaluation Metrics: Current metrics may not adequately capture performance on complex, structured generation tasks [15].

4.9 Our Approach: Fine-Tuning SD3.5 for Argument Illustration

Building upon the identified research gaps, our project focuses on fine-tuning Stable Diffusion 3.5 Medium for the specific task of argument illustration. The approach leverages the SD3 architecture's strengths:

- Triple Text Encoder Architecture: Utilizing dual CLIP encoders (CLIPTextModelWithProjection) and a T5 encoder (T5EncoderModel) for comprehensive semantic understanding of argument prompts.
- Flow Matching Scheduler: Implementing FlowMatchEulerDiscreteScheduler for improved generation quality with 28 inference steps.
- Latent Space Generation: Operating in 16-channel latent space (512×512 images → 64×64 latent representations) for computational efficiency.
- Classifier-Free Guidance: Employing guidance scale of 4.5 for balanced prompt adherence and image quality.

Our training pipeline processes 14,000 argument-image pairs from the CLEF dataset, with fine-tuning focused on the SD3Transformer2DModel component while keeping text encoders frozen. Initial validation on prompts such as "Celebrity culture distorts personal values" and "Excessive homework reduces student well-being" demonstrates the model's ability to generate semantically relevant illustrations for abstract argumentative concepts.

5. Appendices

Summary of Reviewed Models

Model	Year	Architecture	Parameters	Key Innovation
DALL-E	2021	Autoregressive Transformer	12B	Zero-shot generation with dVAE
CogView	2021	Transformer + VQ-VAE	4B	Scalable Chinese T2I
GLIDE	2021	Diffusion	3.5B	Classifier-free guidance
Stable Diffusion	2022	Latent Diffusion	~1B	Efficient latent space generation

DALL-E 2 (unCLIP)	2022	Diffusion + Prior	-	CLIP latent generation
Parti	2022	Autoregressive	-	Translation paradigm
PixArt-Σ	2024	Diffusion Transformer	0.6B	4K direct generation
SD 3.5 Medium	2024	Diffusion Transformer	~2B	Triple text encoder architecture
HiDream-I1	2025	Sparse DiT + MoE	17B	Efficient sparse architecture
MMaDA	2025	Unified Diffusion	8B	Multimodal reasoning

Project Implementation Details

Fine-Tuning Configuration:

- Base Model: Stable Diffusion 3.5 Medium
- Training Dataset: CLEF Argument Illustration Dataset (14,000+ samples)
- Training Duration: 29 epochs
- Image Resolution: 512×512 pixels
- Latent Dimensions: 16 channels × 64×64
- Inference Steps: 28
- Guidance Scale: 4.5
- Scheduler: FlowMatchEulerDiscreteScheduler

Text Encoding Pipeline:

- CLIP Encoder 1: CLIPTextModelWithProjection (77 tokens max)
- CLIP Encoder 2: CLIPTextModelWithProjection (77 tokens max)
- T5 Encoder: T5EncoderModel (256 tokens max)
- Combined embedding dimension: 4096

Sample Test Prompts:

- "Celebrity culture distorts personal values"
- "Excessive homework reduces student well-being"
- "Energy drinks disrupt natural sleep cycles"
- "Urban noise pollution damages mental calm"

Evaluation Metrics

- CLIP Score: Evaluates alignment between generated images and text prompts
- Human Evaluation: Assessment of image quality, prompt adherence, and aesthetic appeal

6. References

[1] H. Zhang et al., "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks," arXiv preprint arXiv:1710.10916, 2017.

[2] T. Xu et al., "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks," arXiv preprint arXiv:1910.09399, 2019.

[3] A. Ramesh, M. Pavlov et al., "Zero-Shot Text-to-Image Generation," arXiv preprint arXiv:2102.12092, Feb. 2021.

[4] R. Rombach, A. Blattmann et al., "High-Resolution Image Synthesis with Latent Diffusion Models," arXiv preprint arXiv:2112.10752, Apr. 2022.

[5] J. Chen, C. Ge et al., "PixArt-Σ: Weak-to-Strong Training of Diffusion Transformer for 4K Text-to-Image Generation," arXiv preprint arXiv:2403.04692, Mar. 2024.

[6] I. Goodfellow et al., "Generative Adversarial Nets," Advances in Neural Information Processing Systems, vol. 27, Jun. 2014.

[7] Q. Cai, J. Chen, Y. Chen, and Y. Li, "HiDream-I1: A High-Efficient Image Generative Foundation Model with Sparse Diffusion Transformer," arXiv preprint, May 2025.

[8] A. Ramesh et al., "Hierarchical Text-Conditional Image Generation with CLIP Latents," arXiv preprint arXiv:2204.06125, Apr. 2022.

[9] A. Süleyman and G. Biricik, "Grounding Text-to-Image Diffusion Models for Controlled High-Quality Image Generation," arXiv preprint, Jan. 2025.

[10] K. Lee and H. Liu, "Aligning Text-to-Image Models using Human Feedback," arXiv preprint arXiv:2302.12192, Feb. 2023.

[11] J. Yu, Y. Xu, J. Y. Koh et al., "Scaling Autoregressive Models for Content-Rich Text-to-Image Generation," arXiv preprint arXiv:2206.10789, Jun. 2022.

[12] M. Ding, Z. Yang, W. Hong et al., "CogView: Mastering Text-to-Image Generation via Transformers," arXiv preprint arXiv:2105.13290, Nov. 2021.

[13] A. Nichol, P. Dhariwal, A. Ramesh et al., "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models," arXiv preprint arXiv:2112.10741, Dec. 2021.

[14] L. Yang, Y. Tian, B. Li et al., "MMaDA: Multimodal Large Diffusion Language Models," arXiv preprint,

May 2025.

[15] J. Cho, A. Zala, and M. Bansal, "Visual Programming for Step-by-Step Text-to-Image Generation and Evaluation," arXiv preprint arXiv:2305.13482, May 2023.

[16] GenAI, Meta, "Imagine Yourself: Tuning-Free Personalized Image Generation," arXiv preprint, Sep. 2025.

[17] M. Liu, Y. Ma, and Z. Yang, "LLM4GEN: Leveraging Semantic Representation of LLMs for Text-to-Image Generation," arXiv preprint arXiv:2408.13131, Aug. 2024.

[18] Chameleon Team, "Chameleon: Mixed-Modal Early-Fusion Foundation Models," arXiv preprint, Mar. 2025.

7. Bibliography

- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. "Generative Adversarial Nets." Advances in Neural Information Processing Systems, 2014.
- Vaswani, A. et al. "Attention Is All You Need." Advances in Neural Information Processing Systems, 2017.
- Radford, A. et al. "Learning Transferable Visual Models From Natural Language Supervision." ICML, 2021.
- Ho, J., Jain, A., and Abbeel, P. "Denoising Diffusion Probabilistic Models." Advances in Neural Information Processing Systems, 2020.
- Saharia, C. et al. "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding." NeurIPS, 2022.
- OpenAI. "DALL-E 2." https://openai.com/dall-e-2/, 2022.
- Stability AI. "Stable Diffusion." https://stability.ai/, 2022.
- Hugging Face. "Diffusers Library." https://huggingface.co/docs/diffusers/, 2023.
- Lin, T. Y. et al. "Microsoft COCO: Common Objects in Context." ECCV, 2014.
- Schuhmann, C. et al. "LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models." NeurIPS, 2022.

Document prepared in accordance with IEEE citation and reference guidelines.