Project Title

FYP– I REPORT

BS(AI) Fall 2025

Name: Muhammad Arham Hussain Khan
22k-4080

Name: Aarib Ahmed Vahidy

22k-4004

Name: Partham Kumar

22k-4079

Supervisor: Zain-ul-Hassan

Co-supervisor: Nouman Durrani

Department of Computer Science

FAST-National University of Computer & Emerging Sciences, Karachi

Table of Contents

INTRODUCTION

Images serve as a potent medium for expressing meaning and facilitating comprehension. Conversely, arguments are frequently regarded as structured and logical frameworks. The Touché 2025 shared task at CLEF (Conference and Labs of the Evaluation Forum) seeks to investigate how images can effectively represent essential elements of arguments, thereby improving their clarity and influence.

An argument generally comprises a claim that is substantiated by one or more premises, which offer evidence for its legitimacy. In this research initiative, arguments are crafted to be succinct, with each argument featuring solely a single claim devoid of supporting premises. The task is: *"Given an argument, find (retrieve or generate) images that assist in conveying central aspects of the argument's claim."*

Problem Statement

The challenge addresses the disconnect between textual argumentation and visual communication. While arguments based on text are proficient for logical reasoning, visual representations can greatly enhance understanding and emotional resonance. The objective is to create AI systems capable of producing contextually pertinent images that accurately reflect the fundamental message of textual arguments.

Research Objectives

Our main goals for this research endeavor are:

1. Fine-tune a cutting-edge text-to-image diffusion model (Stable Diffusion 3.5 Medium) on argument-image pairs
2. Develop a training pipeline tailored for limited hardware resources (24GB VRAM)
3. Generate high-quality, contextually precise images from textual argument claims
4. Assess generated images utilizing CLIP alignment scores and BRISQUE quality metrics
5. Submit findings to the Touché 2025 evaluation campaign

Scope and Constraints

Initially, we sought to employ Stable Diffusion 3.5 Large for this endeavor because of its enhanced image generation abilities. Nevertheless, hardware constraints (NVIDIA RTX 4050 with 6GB VRAM and RTX 4090 with 24GB VRAM for training) compelled us to transition to Stable Diffusion 3.5 Medium, a model with 2 billion parameters that offers an ideal equilibrium between quality and computational demands.

METHODOLOGY

Research Approach

This study adopts an experimental quantitative research framework. We implement deep learning methodologies, particularly diffusion models integrated with transformer architectures, to tackle the image generation challenge. The methodology encompasses:

1. Hypothesis: A fine-tuned Stable Diffusion model can generate contextually aware images from argument claims.
2. Experimentation: Systematic training with controlled hyperparameters and evaluation metrics.
3. Quantitative Analysis: Numerical evaluation using CLIP scores, BRISQUE scores, and training loss metrics.

Data Collection and Preprocessing

Dataset Origin:

The training data was derived from the official CLEF 2025 Touché dataset. The dataset contains:

- 128 distinctive arguments in XML format (arguments.xml)
- Corresponding image collection for training purposes
- Each argument contains: ID, topic, and claim

Data Preprocessing Pipeline:

The CLEF dataset provided images accompanied by pre-existing visual descriptions (captions detailing image content). Our objective was to formulate argument-centric training captions by semantically aligning these images with the 128 arguments.

1. Argument Extraction: Extracted argument claims and topics from the XML file (arguments.xml). Each argument was characterized by an ID, topic, and claim statement.

2. Semantic Embedding Generation: Used the GTE-Large sentence transformer model to generate 1024-dimensional embeddings for both argument claims and image visual descriptions. The model was run on GPU with FP16 precision for efficient batch processing.
3. Computed cosine similarity between each image's visual description embedding and all 128 argument embeddings. Images were matched to arguments exceeding a similarity threshold of 0.25, ensuring semantic relevance while maintaining dataset diversity.
4. Argument-Centric Caption Formulation: For each matched image-argument pair, we developed training captions by amalgamating the argument claims with the corresponding visual descriptions:

1. *"An image representing the argument: [CLAIM]. [VISUAL_DESCRIPTION]"*

5. Image Preprocessing:
    a. Initial training conducted at a resolution of 256×256 pixels to facilitate rapid iteration.
    b. Final training performed at a resolution of 512×512 pixels to ensure high-quality output.
    c. Data augmentation techniques employed include: color jitter (0.1), random horizontal flipping (30%), and random affine transformations.
6. JSON Output: The training data is organized in JSONL format adhering with submission requirements.

Model Architecture and Training

Base Model:

Stable Diffusion 3.5 Medium (stabilityai/stable-diffusion-3.5-medium)

Model Components:

- ➢ Transformer (MMDiT): 2 billion parameters, trained with LoRA adapters
- ➢ Text Encoders:
    - o CLIP ViT-L/14 (text_encoder_1)
    - o CLIP ViT-G/14 (text_encoder_2)
    - o T5-XXL (text_encoder_3) with 8-bit quantization
- ➢ VAE: An autoencoder designed for encoding and decoding in latent space.
- ➢ Scheduler: FlowMatchEulerDiscreteScheduler

Training Configuration:

Parameter	Value
Base Resolution	512×512 pixels
Batch Size	4 (with gradient accumulation)
Gradient Accumulation Steps	4
Learning Rate	1e-4
Optimizer	AdamW 8-bit
Mixed Precision	BFloat16
LoRA Rank	8
LoRA Alpha	8
Training Epochs	30
Training Samples	5,000 out of 14,000

How we conducted Training:

1. Initial fine-tuning at 256×256 resolution for rapid experimentation
2. Evaluation of contextual accuracy at lower resolution
3. Once satisfied with contextual understanding, scaled to 512×512 for final training
4. Gradient checkpointing enabled for memory efficiency
5. Early stopping with patience of 5 epochs

Evaluation Metrics Used

1. CLIP Score

This metric assesses the semantic alignment between the images generated and the input prompts by utilizing CLIP embeddings. It is computed as the cosine similarity between the embeddings of the image and the text.

2. Aesthetic Score

This score evaluates the visual appeal of the images through CLIP-based aesthetic predictors. It takes into account factors such as composition, color harmony, and the overall quality of the visuals.

3. BRISQUE Score

The Blind Image Spatial Quality Evaluator measures the technical quality of images, which includes:

- Sharpness (measured by Laplacian variance)

- Contrast (assessed through standard deviation)

- Brightness uniformity (evaluated via histogram entropy)

4. Training Loss

This refers to the Mean Squared Error (MSE) calculated between the predicted noise and the actual noise in the diffusion process, which is tracked over multiple epochs.

Note: The official evaluation metric NDCG@5 (Normalized Discounted Cumulative Gain) requires human evaluators and will be assessed by the CLEF community during the official evaluation phase.

TESTING AND RESULTS

Training Results

The model was trained for 30 epochs on 5,000 samples with the following outcomes:

Metric	Result
Training Accuracy	61%
Final Training Loss	Converged (0.3901)
Training Time	~192 hours
Hardware Used	NVIDIA RTX 4090

Validation Results

Validation was performed on all 128 arguments from the test set. For each argument, 5 sample images were generated with different seeds, and the best sample was selected based on combined scoring.

Metric	Score
Average CLIP Score	61.67
Average BRISQUE Score	67.81
Samples per Argument	5
Total Arguments	128
Final Selected Images	128 (best of 5)

Comparison with Last Year's Results - CLEF 2024

The table below presents the outcomes from last year's competition for your reference:

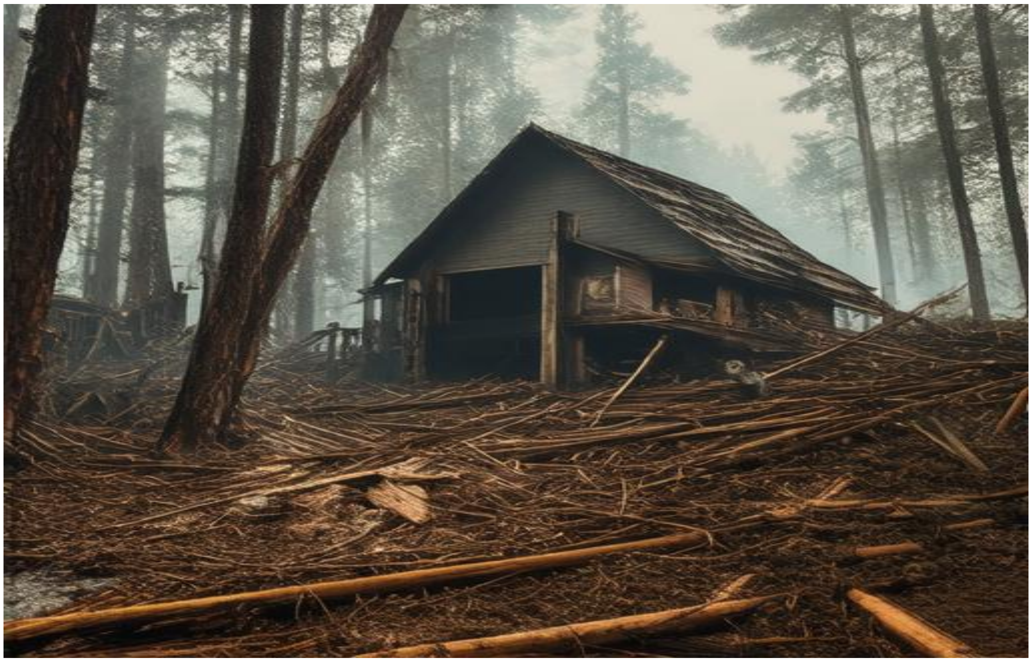
Team	Approach	NDCG@5
HTW-DIL	Ada-Summary	0.428
HTW-DIL	Moondream-Text	0.363
HTW-DIL	Moondream-Image-Text-Default	0.293
Baseline	BM25	0.284
Baseline	SBERT	0.232
DS@GT	Generated-Image-CLIP	0.180
HTW-DIL	Moondream-Image-Text-3epochs	0.150
HTW-DIL	Moondream-Image	0.146
DS@GT	Base-CLIP	0.123
HTW-DIL	Moondream-Image-Text-2epochs	0.120

Note: Our results for CLEF 2025 are pending official evaluation. The NDCG scores will be determined by human evaluators from the CLEF community.

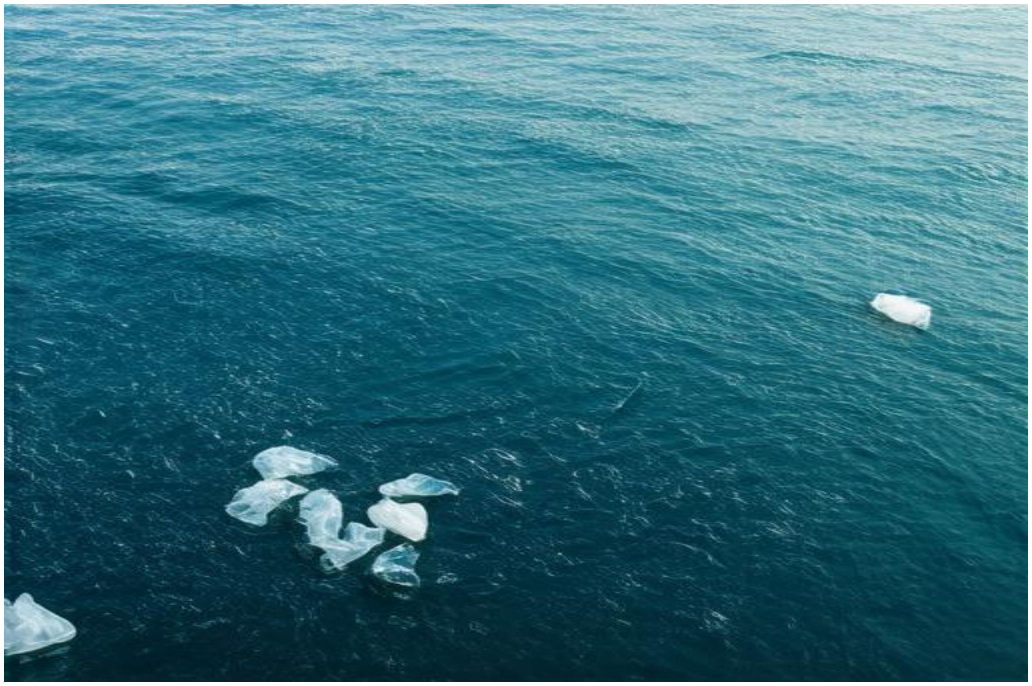
Sample Generated Images

The generated images demonstrate strong contextual understanding of argument claims. Examples include:

➢ Argument: “Cutting down old forests destroys the homes of countless animals.”

.


- Argument: "Plastic waste floating in the ocean endangers marine life."

.


Observations

1. The model successfully learned to generate contextually relevant images from argument claims except for some scenarios where it generated facial features incorrectly.

2. CLIP scores indicate good semantic alignment between prompts and generated images but might not represent true human nature contextual mapping.

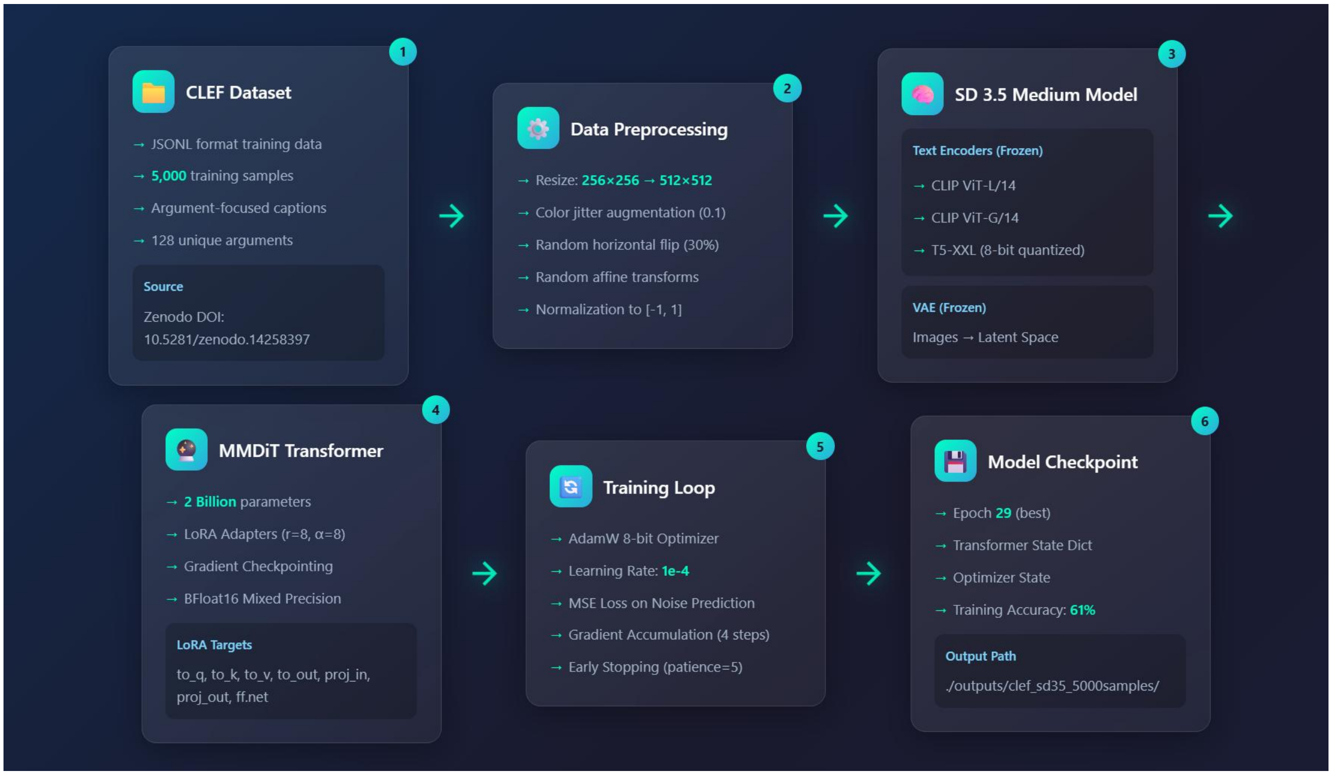
3. BRISQUE scores suggest technically competent image quality

4. The multi-sample selection strategy (5 samples per argument) improved final output quality

5. Progressive resolution training (256×256 to 512×512) proved effective for balancing speed and quality

SYSTEM DIAGRAM

Training Pipeline Architecture



1 CLEF Dataset

→ JSONL format training data

→ 5,000 training samples

→ Argument-focused captions

→ 128 unique arguments

Source

Zenodo DOI: 10.5281/zenodo.14258397

2 Data Preprocessing

→ Resize: 256×256 → 512×512

→ Color jitter augmentation (0.1)

→ Random horizontal flip (30%)

→ Random affine transforms

→ Normalization to [-1, 1]

3 SD 3.5 Medium Model

Text Encoders (Frozen)

→ CLIP ViT-L/14

→ CLIP ViT-G/14

→ T5-XXL (8-bit quantized)

VAE (Frozen)

Images → Latent Space

4 MMDiT Transformer

→ 2 Billion parameters

→ LoRA Adapters (r=8, α=8)

→ Gradient Checkpointing

→ BFloat16 Mixed Precision

LoRA Targets

to_q, to_k, to_v, to_out, proj_in, proj_out, ff.net

5 Training Loop

→ AdamW 8-bit Optimizer

→ Learning Rate: 1e-4

→ MSE Loss on Noise Prediction

→ Gradient Accumulation (4 steps)

→ Early Stopping (patience=5)

6 Model Checkpoint

→ Epoch 29 (best)

→ Transformer State Dict

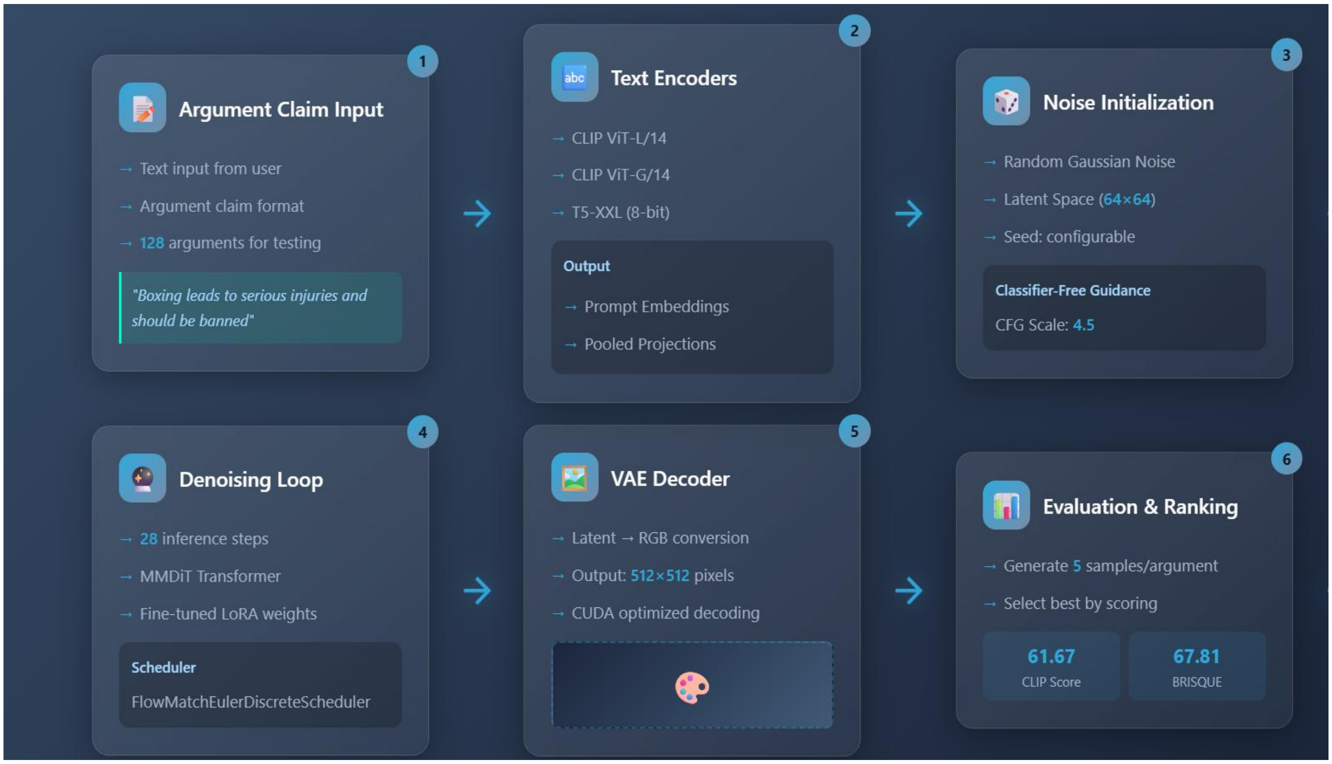
→ Optimizer State

→ Training Accuracy: 61%

Output Path

./outputs/clef_sd35_5000samples/

Inference Pipeline



1 Argument Claim Input

→ Text input from user

→ Argument claim format

→ 128 arguments for testing

"Boxing leads to serious injuries and should be banned"

2 Text Encoders

→ CLIP ViT-L/14

→ CLIP ViT-G/14

→ T5-XXL (8-bit)

Output

→ Prompt Embeddings

→ Pooled Projections

3 Noise Initialization

→ Random Gaussian Noise

→ Latent Space (64×64)

→ Seed: configurable

Classifier-Free Guidance

CFG Scale: 4.5

4 Denoising Loop

→ 28 inference steps

→ MMDiT Transformer

→ Fine-tuned LoRA weights

Scheduler

FlowMatchEulerDiscreteScheduler

5 VAE Decoder

→ Latent → RGB conversion

→ Output: 512×512 pixels

→ CUDA optimized decoding

6 Evaluation & Ranking

→ Generate 5 samples/argument

→ Select best by scoring

61.67 CLIP Score

67.81 BRISQUE

GOALS FOR FYP-II

Building upon the foundation established in FYP-I, the following objectives are planned for FYP-II:

Model Optimization Techniques

1. Exponential Moving Average: Utilize EMA for model weights during the training process, which has been demonstrated to significantly improve overall accuracy and stability in diffusion models. EMA keeps a running average of model parameters.
2. Learning Rate Scheduling: Employ cosine annealing or warm restarts to achieve more effective training dynamics.
3. Advanced LoRA Configurations: Explore higher ranks and various target modules to enhance fine-tuning.

Training Enhancements

1. Larger Dataset: Broaden the training dataset with additional argument-image pairs.
2. Multi-Resolution Training: Conduct training across multiple resolutions simultaneously to achieve better scale invariance.
3. Caption Augmentation: Create a wider array of caption variations to enhance generalization.

Evaluation Improvements

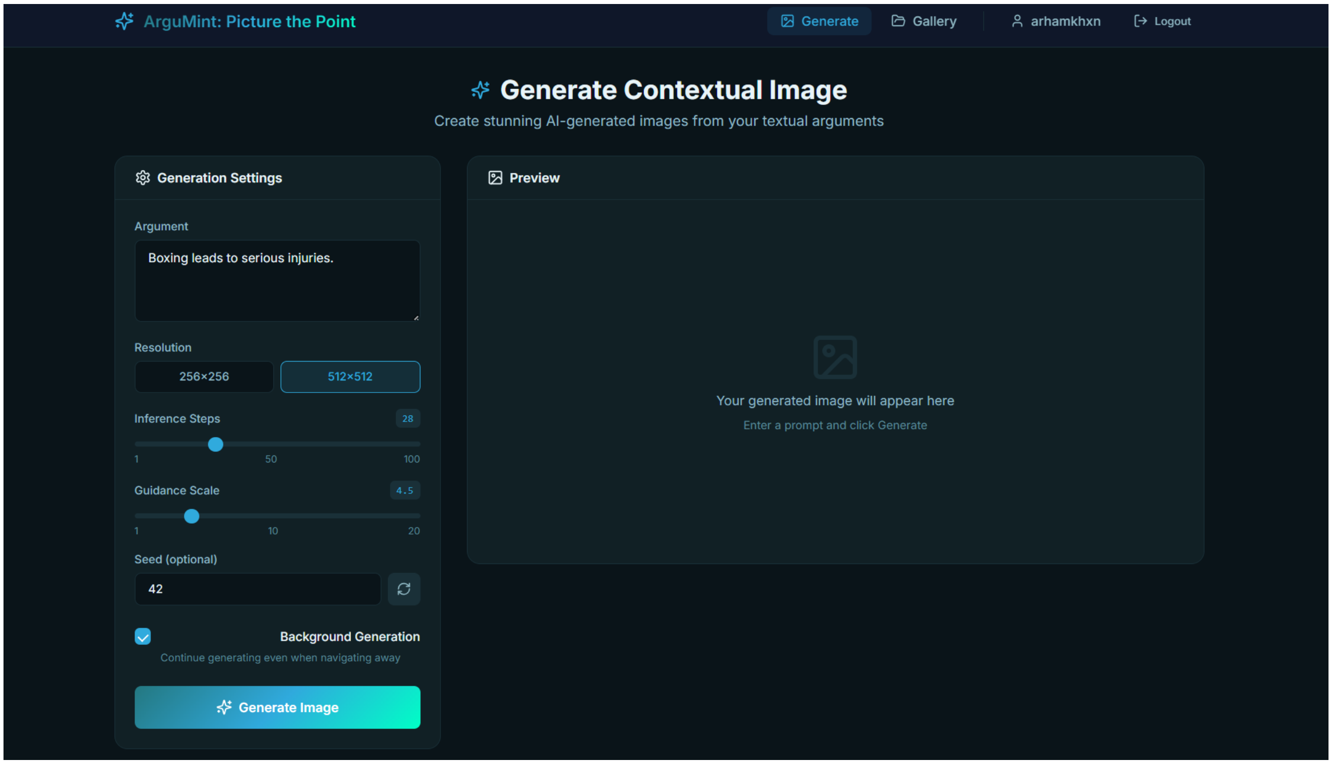
1. Human Evaluation: Perform validation checks through online Touche evaluator that will be accepting responses in future.
2. Additional Metrics: Introduce FID (Fréchet Inception Distance) and IS (Inception Score) for a thorough evaluation or introduce NDCG-equivalent (Normalized Discounted Cumulative Gain) metric.
3. Ablation Studies: Conduct systematic evaluations of various training configurations.

Web Application Development

- Complete the full-stack web application with:

- User authentication and session management
- Real-time image generation with progress tracking
- Image gallery with search and filtering
- API documentation for external integration

➢ Homepage:



Generate Contextual Image

Create stunning AI-generated images from your textual arguments

Argument: Boxing leads to serious injuries.

Resolution: 256×256 | 512×512

Inference Steps: 28 (1 – 50 – 100)

Guidance Scale: 4.5 (1 – 10 – 20)

Seed (optional): 42

☑ Background Generation
Continue generating even when navigating away

Your generated image will appear here
Enter a prompt and click Generate

Competition Goals

1. Analyze official CLEF 2025 evaluation results upon release
2. Submit participant paper to CLEF 2025 workshop

CONCLUSION

This FYP-I project successfully addressed the Touché 2025 shared task on Arguments to Image Generation at CLEF. We developed and fine-tuned a Stable Diffusion 3.5 Medium model capable of generating contextually relevant images from textual argument claims.

Key Achievements:

1. Successful Model Fine-tuning: Trained SD 3.5 Medium with LoRA adapters, achieving 61% training accuracy on argument-image alignment.
2. Effective Evaluation Pipeline: Developed automated evaluation using CLIP scores (61.67 average) and BRISQUE quality metrics (67.81 average).
3. Hardware Optimization: Successfully adapted training pipeline for limited GPU memory using 8-bit quantization, gradient checkpointing, and mixed precision training.
4. Multi-Sample Selection: Implemented strategy to generate 5 samples per argument and select the best based on combined scoring.
5. Web Application Framework: Developed a React + Flask application for interactive image generation.

Challenges Addressed:

- Overcame hardware limitations by switching from SD 3.5 Large to SD 3.5 Medium
- Developed progressive resolution training strategy for efficient experimentation
- Created comprehensive data preprocessing pipeline for argument-image pairs

Current Status:

Our submission to Touché 2025 @ CLEF is complete, and we are awaiting official evaluation results. The NDCG@5 scores will be determined by human evaluators from the CLEF community, with results expected before the CLEF 2025 conference in Madrid.

This project demonstrates the feasibility of using modern diffusion models for argument visualization and lays the groundwork for further improvements in FYP-II.

REFERENCES

[1] M. Heinrich, J. Kiesel, M. Wolter, M. Potthast, and B. Stein, "Image Retrieval/Generation for Arguments 2025," Touché @ CLEF 2025. [Online]. Available: https://touche.webis.de/clef25/touche25-web/image-retrieval-for-arguments.html

[2] Stability AI, "Stable Diffusion 3.5 Medium," Hugging Face, 2024. [Online]. Accessible at: https://huggingface.co/stabilityai/stable-diffusion-3.5-medium

[3] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," arXiv preprint arXiv:2106.09685, 2021.

[4] P. Esser et al., "Scaling Rectified Flow Transformers for High-Resolution Image Synthesis," presented at the 41st International Conference on Machine Learning, 2024.

[5] CLEF Initiative, "CLEF 2025 Conference," Madrid, Spain. [Online]. Accessible at: https://clef2025.clef-initiative.eu/

[6] Touché Dataset, "Image Retrieval/Generation for Arguments 2025 Dataset," Zenodo, 2024. DOI: 10.5281/zenodo.14258397

[7] T. Dettmers et al., "8-bit Optimizers via Block-wise Quantization," presented at the 10th International Conference on Learning Representations, 2022.

[8] Hugging Face, "Diffusers: State-of-the-art diffusion models," 2024. [Online]. Accessible at: https://github.com/huggingface/diffusers

[9] Hugging Face, "PEFT: Parameter-Efficient Fine-Tuning," 2024. [Online]. Accessible at: https://github.com/huggingface/peft

Acknowledgments

We wish to convey our heartfelt appreciation to our Supervisor, Mr. Zain-ul-Hassan, and our Co-supervisor, Mr. Nouman Durrani, for their invaluable guidance, unwavering support, and encouragement throughout this project. Their expertise and mentorship have played a crucial role in the successful completion of this work.

We also extend our gratitude to the Touché Task Committee and the CLEF community for organizing this shared task and for providing the dataset and evaluation infrastructure.