FYP

# TextPixs (Text to Image)

Research Based Project


Syeda Anshrah Gillani (1337-2021)
Umema Mujeeb (2396-2021)
Maheen Ali (1589-2021)

**Internal Supervisor**
Sir Osama Ahmed Khan
**External Supervisor & Project Sponsor**
Mirza Samad Ahmed Baig

Department of Computing, FEST
Hamdard University

# Summary



- Problem Statement
- Objective
- FYP Scope
- Literature Review
  - Gap Analysis
  - Comparitive Analysis
- Our methodology
- Architecture
- Our Project Plan (Time lines)
- Budget / Costing
  - Vast.ai Creds
  - Cost Breakdown
- FYP Deliverables
- References

# Problem Statement



"Text-to-image generation challenges AI with accurately translating text into realistic visuals. Current methods struggle with nuanced semantics and visual fidelity. "Text Pixs" aims to pioneer RC-GAN, integrating NLP and computer vision to advance image synthesis quality, impacting content creation, education, and virtual reality."

# Objective



- Develop a Recurrent Convolutional Generative Adversarial Network (RC-GAN) to improve the fidelity and accuracy of generating images from textual descriptions.
- Address the limitations of current text-to-image generation methods in capturing semantic nuances and producing high-quality visual outputs.
- Conduct empirical evaluations to assess the performance of RC-GAN across diverse datasets and application scenarios.
- Enhance the utility of AI in content creation, education, and virtual reality through advanced text-to-image synthesis capabilities.

# FYP Scope



"The scope of the project "Text Pixs" involves developing and evaluating a Recurrent Convolutional Generative Adversarial Network (RC-GAN) tailored for text-to-image generation. This initiative aims to address existing challenges in accurately translating textual descriptions into visually realistic images by leveraging advancements in natural language processing (NLP) and computer vision. By focusing on improving the fidelity and semantic coherence of generated images, "Text Pixs" seeks to advance the capabilities of AI in enhancing visual content synthesis and application across various domains."

# Literature Review (Gap analysis)



In summary, the gap analysis across all three members' research reveals several key areas for improvement in the text-to-image field, such as:

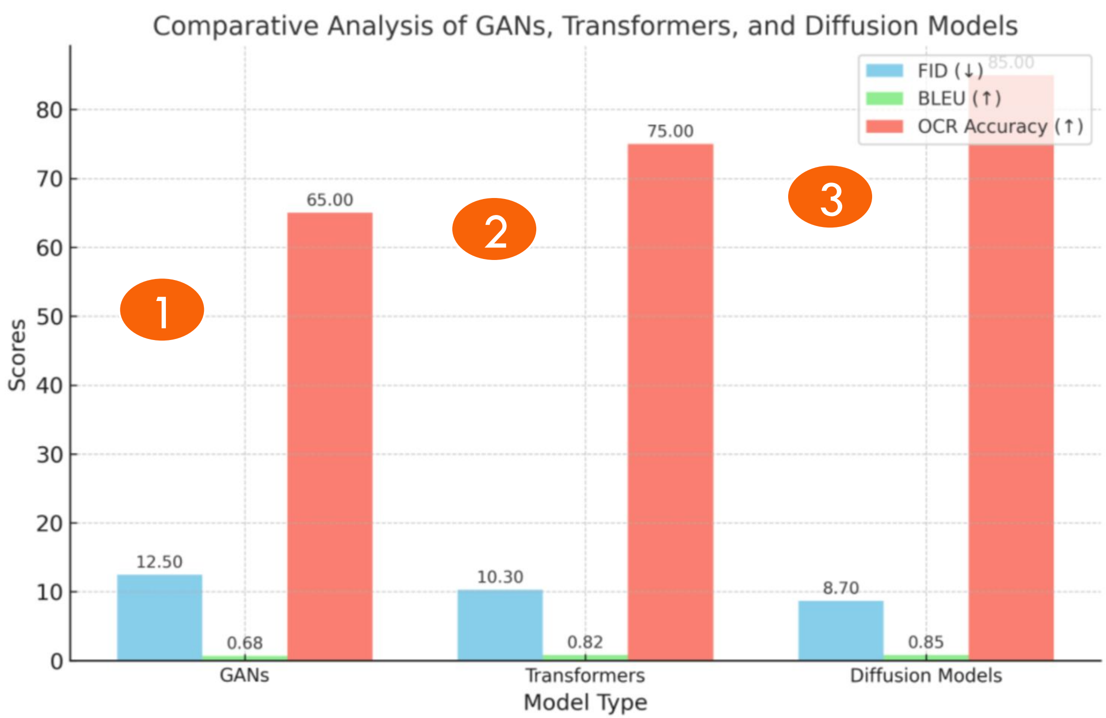
1. Annotation and Dataset Complexity:
Reducing reliance on annotated data and exploring semi-supervised methods.

2. Text-Image Coherence:
Enhancing the connection between the textual input and visual features, especially for complex or abstract text.

3. Evaluation Metrics:
Developing more comprehensive evaluation methods that go beyond quantitative metrics and incorporate subjective, human-centered assessment.

# Literature Review (Comparative Analysis)






**LEGEND**

**FID (Fréchet Inception Distance)**:
Measures the quality and realism of generated images by comparing their distribution to real images.

**BLEU (Bilingual Evaluation Understudy)**:
Evaluates the accuracy of machine-translated text by comparing it to human references using n-gram overlap.

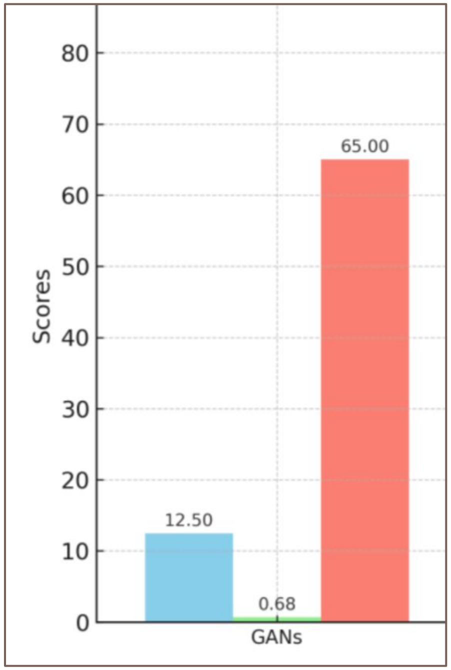
**OCR Accuracy**:
Assesses how accurately an Optical Character Recognition (OCR) system converts scanned text into digital text.

Figure 4: Bar graph showing the comparative analysis of GANs, Transformers, and Diffusion Models based on FID, BLEU, and OCR accuracy. The graph highlights the performance differences across the three model types.

# Literature Review (Comparative Analysis)



1




**GANS** - Random samples generated for human faces.

# Literature Review (Comparative Analysis)



2

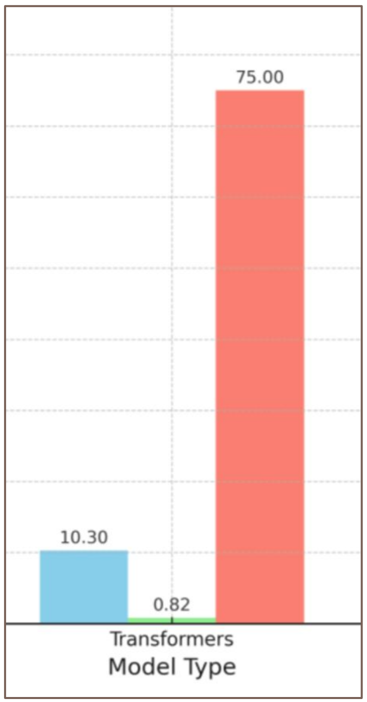



**Image Transformer**

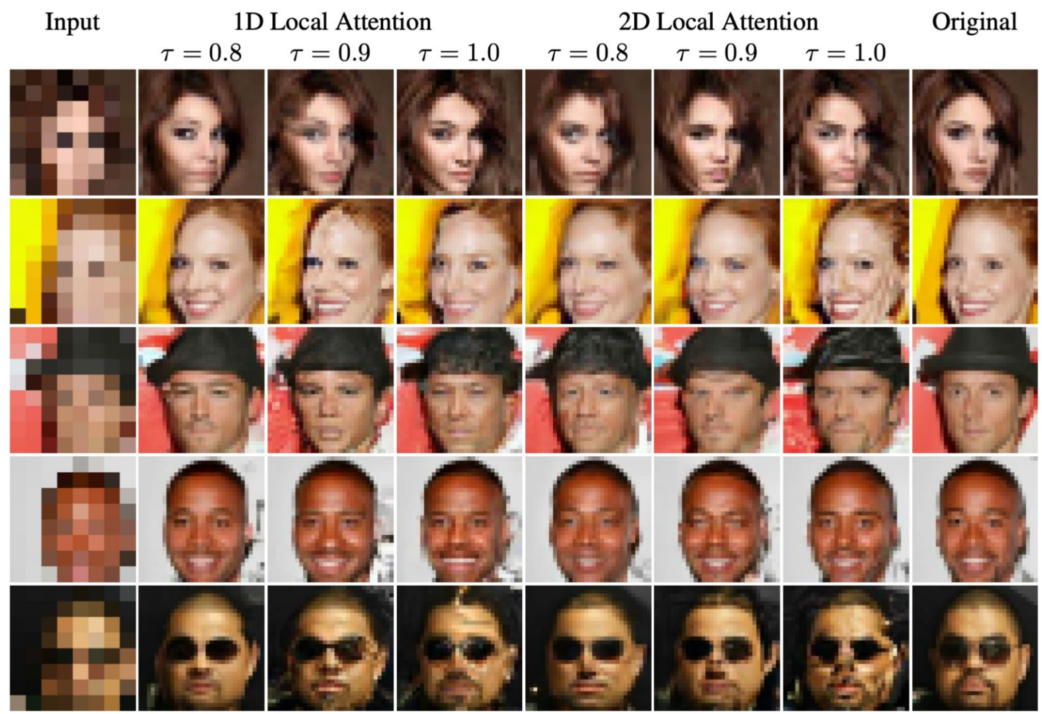



*Table 6.* Images from our 1D and 2D local attention super-resolution models trained on CelebA, sampled with different temperatures. 2D local attention with $\tau = 0.9$ scored highest in our human evaluation study.

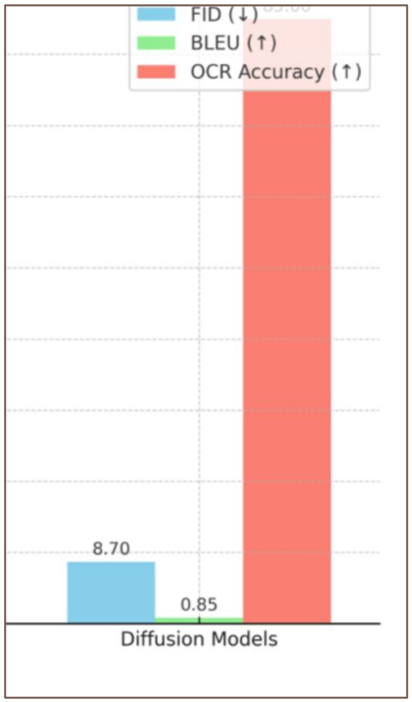
**TRANSFORMERS** - Random samples generated for human faces.

# Literature Review (Comparative Analysis)



3

## Multi-Modal-Driven Face Generation





**DIFFUSION** - Random samples generated for human faces.

# Challenges in existing systems



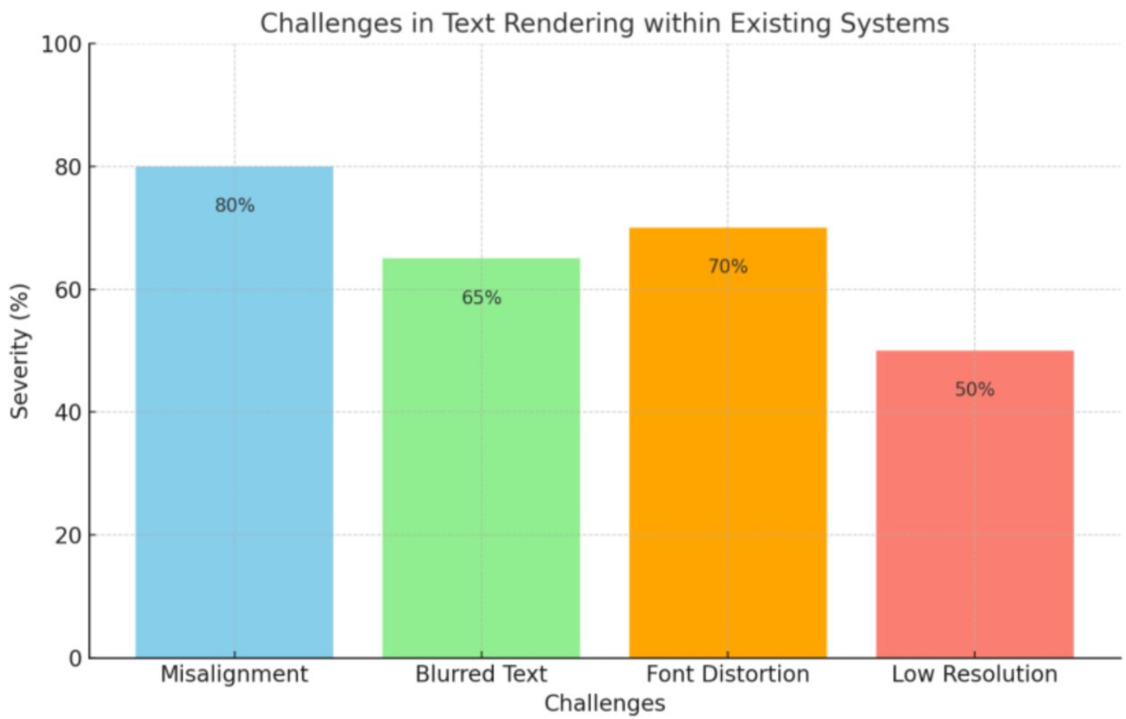
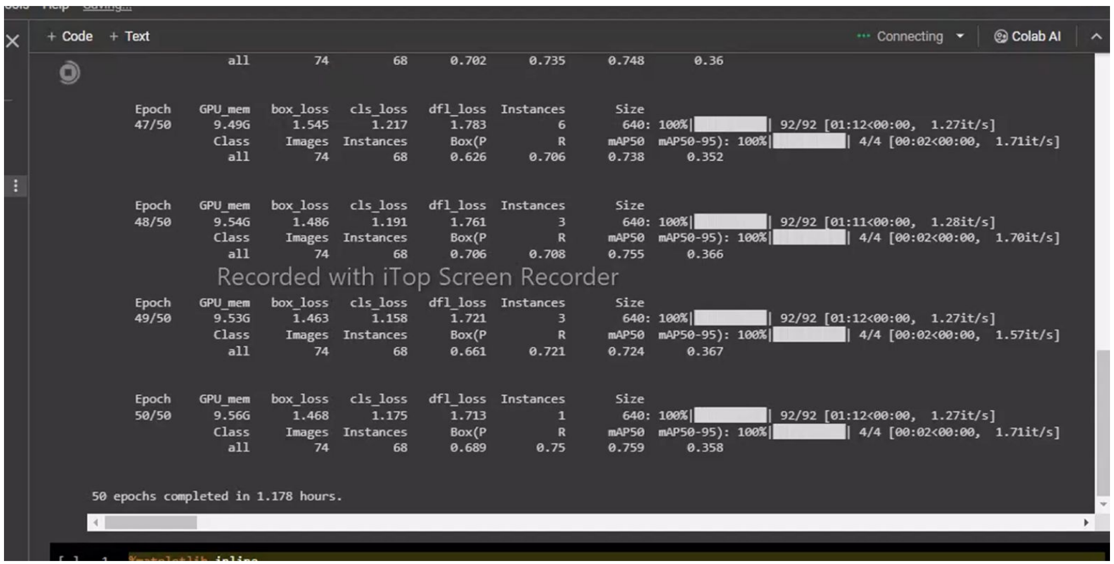



Figure 1: Illustration of text rendering challenges in existing systems, highlighting issues such as misalignment, blurred text, font distortion, and low resolution.

# Model Training / Fine-Tuning

## Fine Tuning / Training



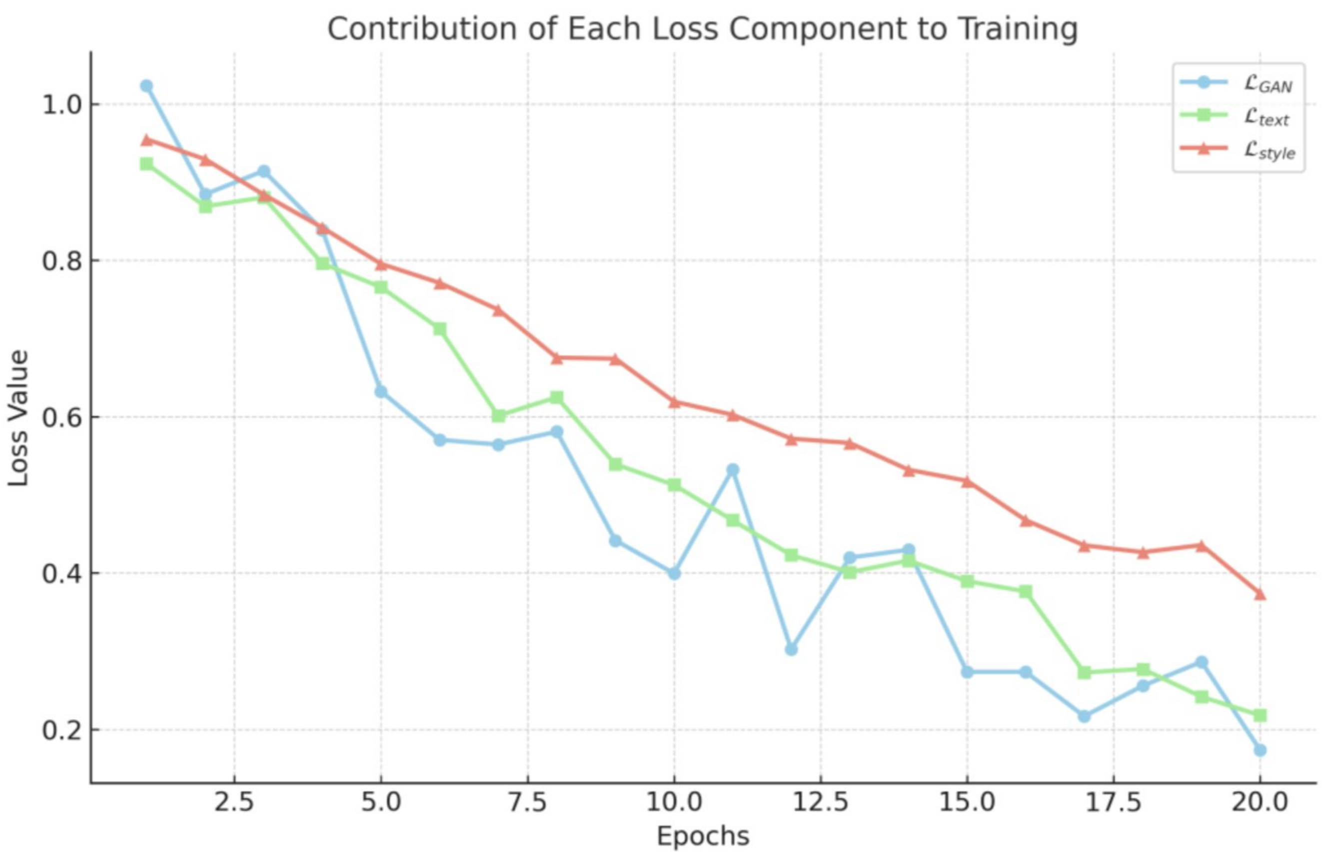

**Medium dataset (~50k-100k samples)** → 20-70 epochs

## Result


Prompt:
Generate a logo for project named "Textpixs" for uiux purpose.

# Contribution of each loss component to training



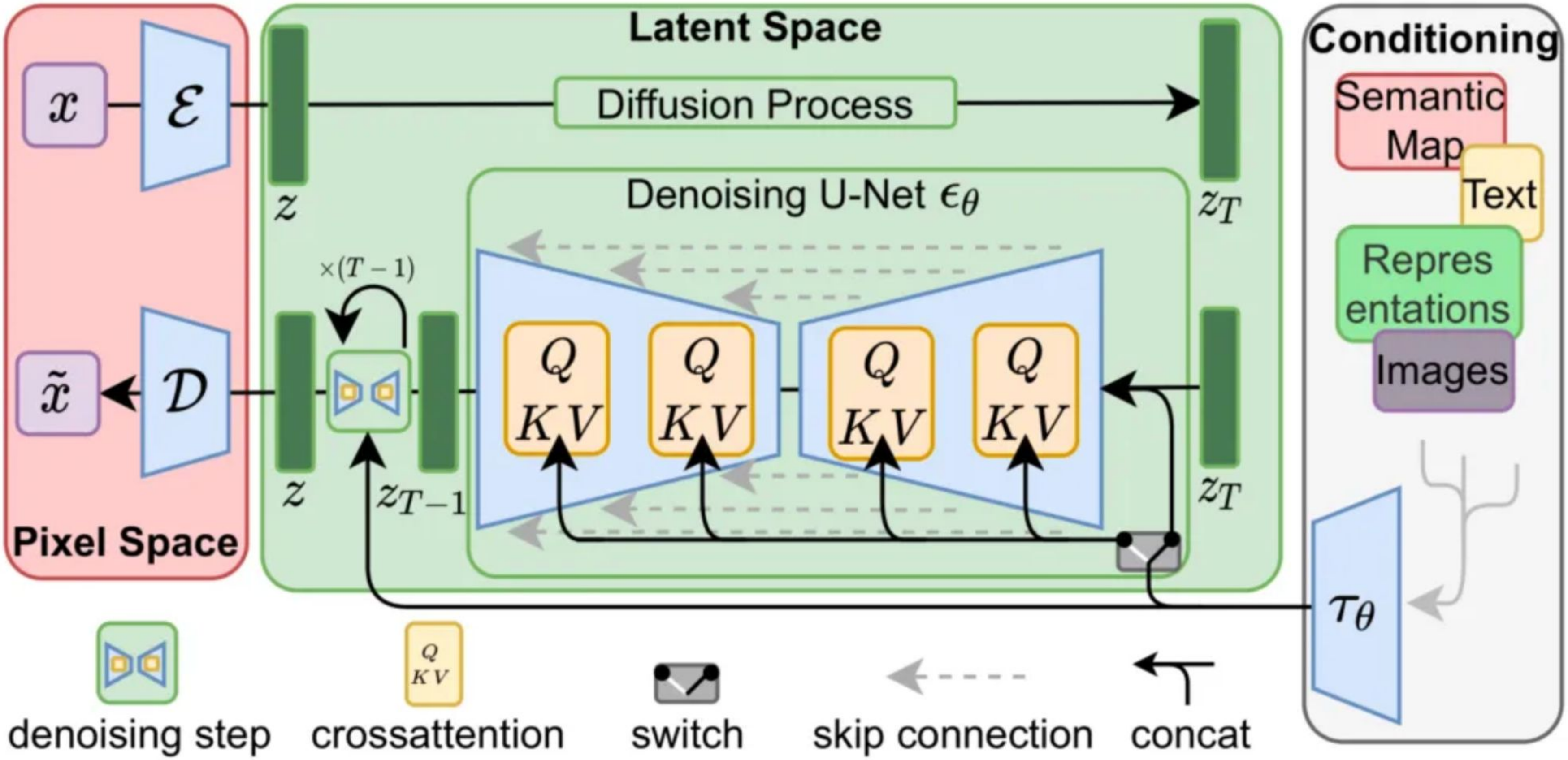



**LEGEND**

- LGAN : Adversarial loss for image synthesis.
- Ltext: Text rendering loss.
- Lstyle: Style consistency loss.

# How it works - Architecture

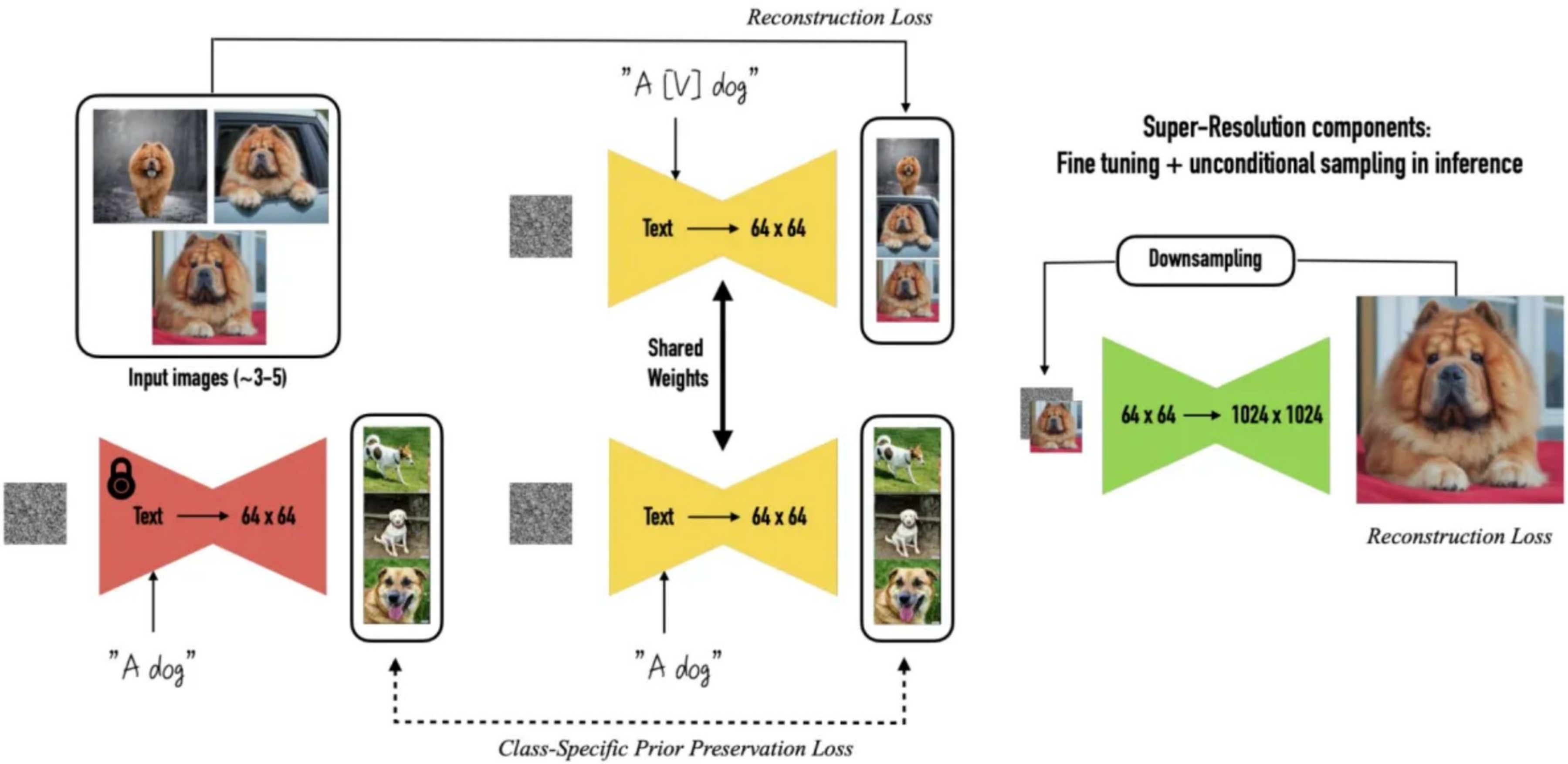





Architecture of Stable Diffusion Model

# How it works - Architecture






Fine tuning a stable diffusion model

# Our Methodology



- The Agile Scrum methodology is ideally suited for the "Text Pixs" project, aiming to advance text-to-image generation technology.

**WHY?**

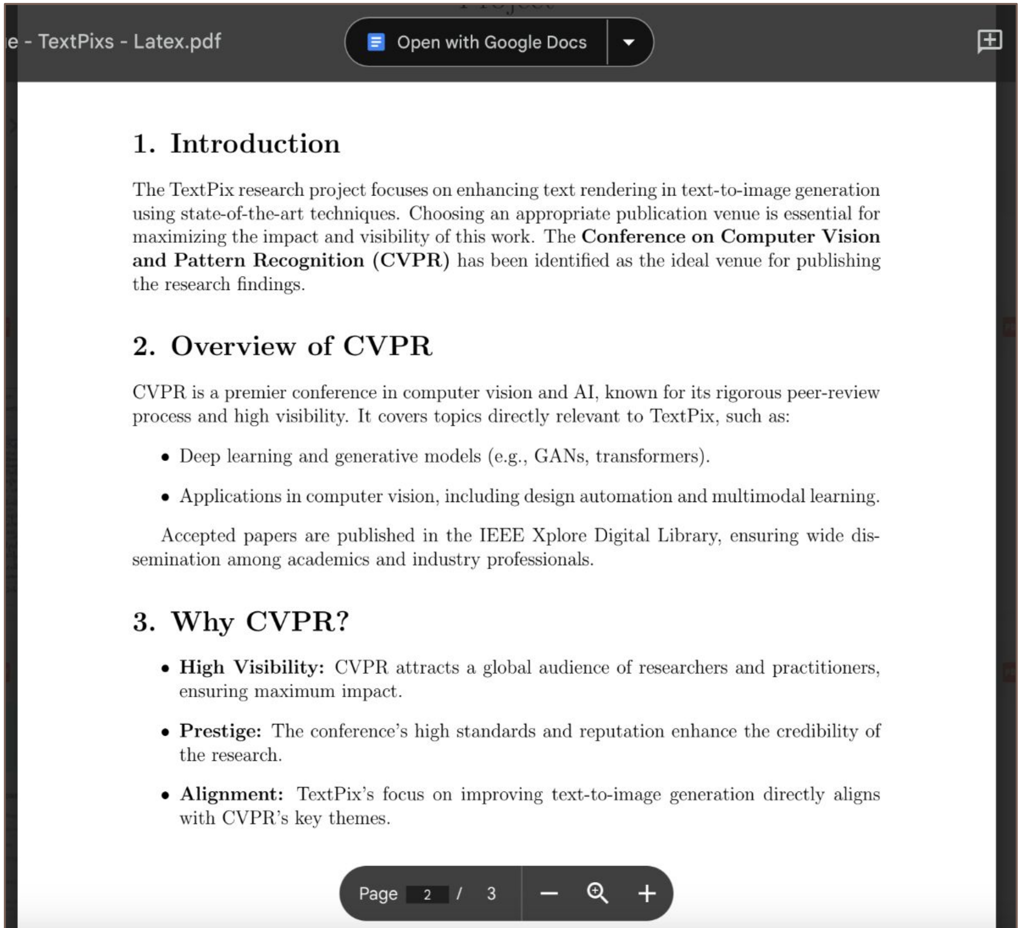
- Agile Scrum is chosen because it facilitates flexibility, iterative development, and continuous feedback loops, which are essential for refining the Recurrent Convolutional Generative Adversarial Network (RC-GAN). The project involves diverse tasks such as data preprocessing, model refinement, and performance evaluation, each requiring focused development cycles. By breaking down these tasks into manageable sprints, the team can prioritize effectively and adjust strategies based on evolving requirements and feedback. This iterative approach ensures that the RC-GAN model can evolve dynamically, meeting the project's goals of enhancing image fidelity and semantic accuracy.

# Our Project Plan



# Selected Publication Venue - CVPR





e - TextPixs - Latex.pdf

Open with Google Docs

## 1. Introduction

The TextPix research project focuses on enhancing text rendering in text-to-image generation using state-of-the-art techniques. Choosing an appropriate publication venue is essential for maximizing the impact and visibility of this work. The **Conference on Computer Vision and Pattern Recognition (CVPR)** has been identified as the ideal venue for publishing the research findings.

## 2. Overview of CVPR

CVPR is a premier conference in computer vision and AI, known for its rigorous peer-review process and high visibility. It covers topics directly relevant to TextPix, such as:

- Deep learning and generative models (e.g., GANs, transformers).
- Applications in computer vision, including design automation and multimodal learning.

Accepted papers are published in the IEEE Xplore Digital Library, ensuring wide dissemination among academics and industry professionals.

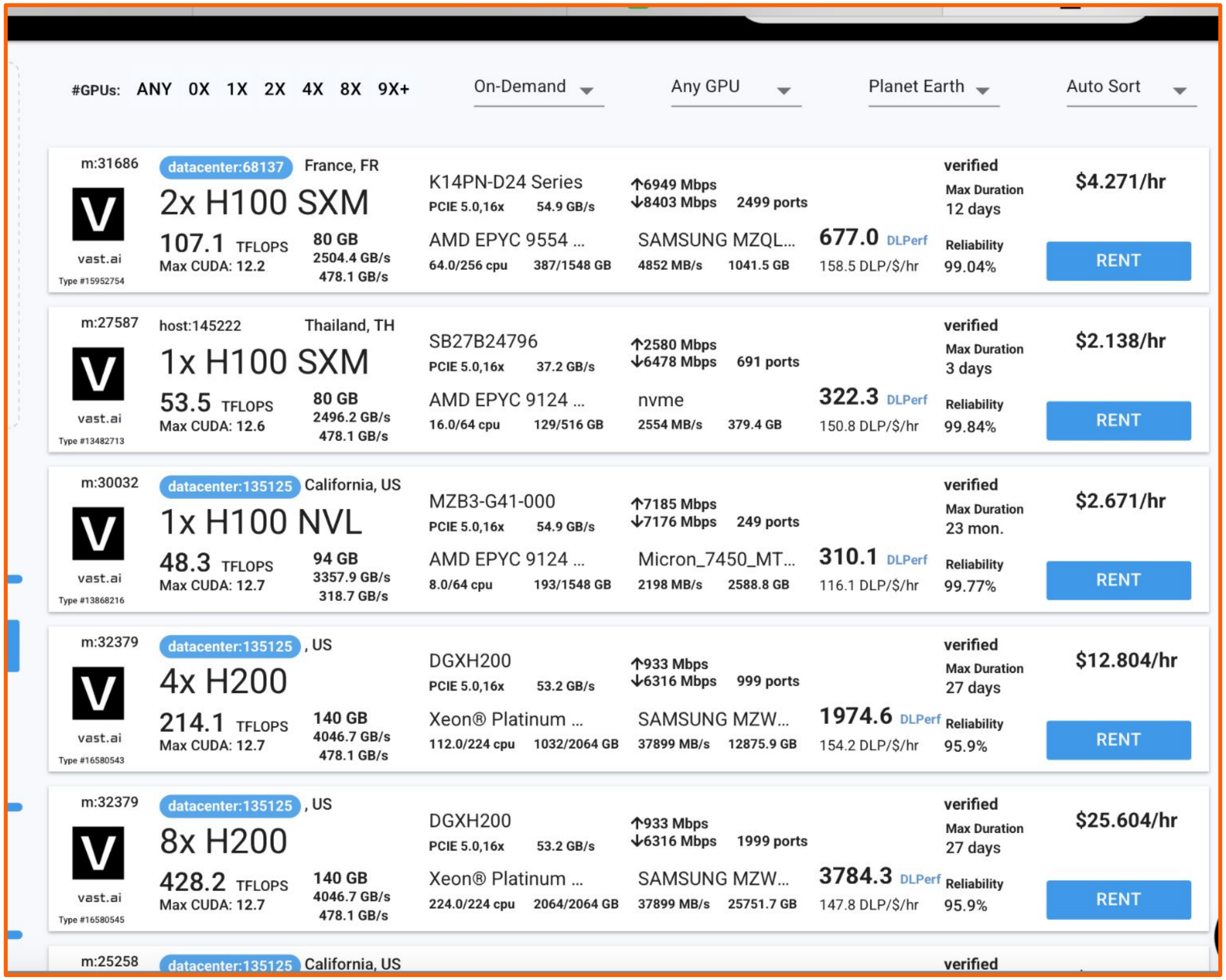
## 3. Why CVPR?

- **High Visibility:** CVPR attracts a global audience of researchers and practitioners, ensuring maximum impact.
- **Prestige:** The conference's high standards and reputation enhance the credibility of the research.
- **Alignment:** TextPix's focus on improving text-to-image generation directly aligns with CVPR's key themes.

Page 2 / 3


# Budget / Costing - Vast.ai





# Budget / Costing

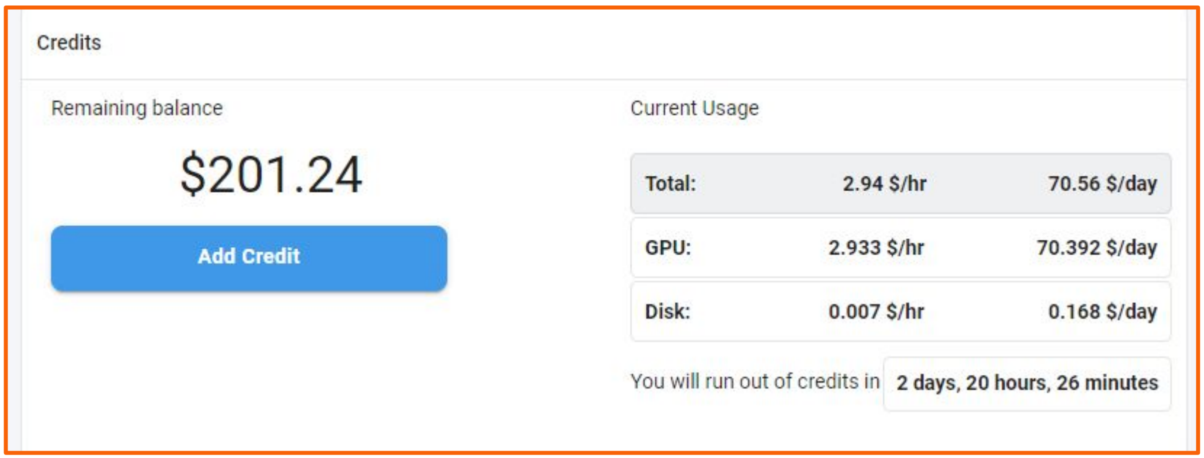


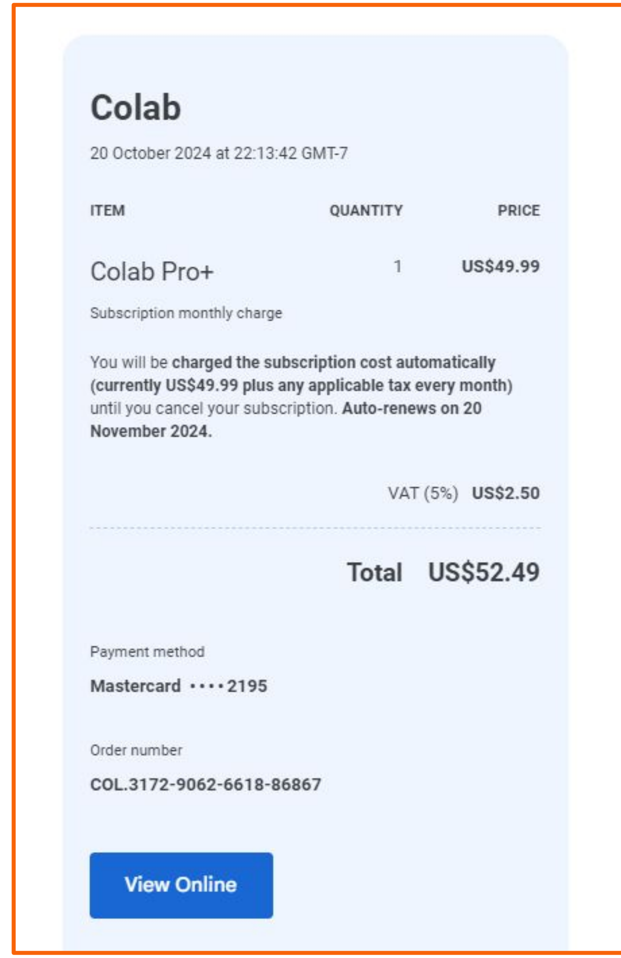
1. GPU/Vast.ai

   We initially added $700 this semester in vast.ai

2. Colab Pro+

   Subscribed to colab pro+ for daily heavy tasks



# Budget / Costing



- **Cloud GPU Cost Detail:**
  - **Nebius H100 SXM 5 GPU:** $3.15 per hour
    - **Estimated Usage:** 500 - 1,000 hours
    - **Cost Range:** $1,575 - $3,150
- **Total Cost: $4,675 - $11,700**

| Category | Item | Estimated Cost |
|---|---|---|
| Hardware and Software | High-Performance Laptop/Desktop | $1,200 - $2,500 |
| | Dedicated GPU | $800 - $2,500 |
| | Cloud Computing Credits | $1,575 - $3,150 |
| | Software Licenses | $200 - $400 |
| | Storage Devices | $100 - $300 |
| Data and Resources | Dataset Acquisition | $100 - $500 |
| | Data Annotation | $200 - $600 |
| Team and Development | Research Materials | $50 - $150 |
| | Conference and Workshop Fees | $100 - $300 |
| Miscellaneous | Printing and Stationery | $50 - $100 |
| | Travel Expenses | $100 - $300 |
| | Contingency | $200 - $400 |
| Total | | $4,675 - $11,700 |

# FYP Deliverables



## FYP-I

- Project Proposal, Scope and Plan
- Definition Literature
- Gap/Comparative Analysis
- Model Selection and Initial Training
- Research Paper Draft
- Selection of Publishing Venue
- Feedback Incorporation

## FYP-II

- Feedback Incorporation
- Algorithm refinement and improvement
- Documentation of Research
- Algorithm Testing and Validation
- Final Research Draft.
- Submission to Journal for Publication
- Final Research
- Final Presentation and Submission

# Some References



**References**


1. **Ramesh et al. (2021)**: *Zero-Shot Text-to-Image Generation*. arXiv:2102.12092
2. **Xu et al. (2018)**: *AttnGAN: Fine-Grained Text-to-Image Generation with Attentional GANs*. CVPR
3. **Saharia et al. (2022)**: *Photorealistic Text-to-Image Diffusion Models*. arXiv:2205.11487
4. **Goodfellow et al. (2014)**: *Generative Adversarial Networks (GANs)*. arXiv:1406.2661
5. **Zhang et al. (2017)**: *StackGAN: Text to Photo-Realistic Image Synthesis with Stacked GANs*. ICCV
6. **Dosovitskiy et al. (2021)**: *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. ICLR

# THANK YOU!