

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



2

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

Problem Statement

3

“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

4

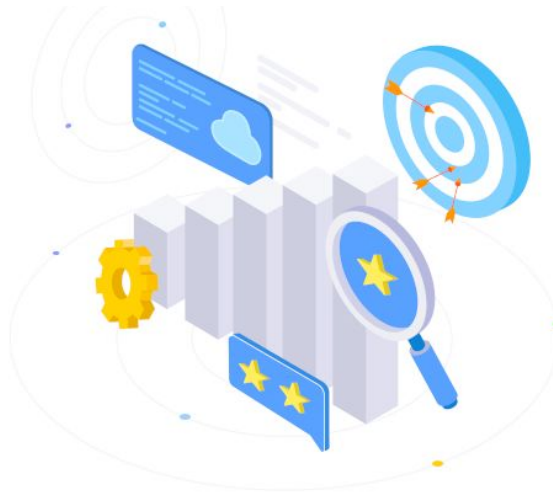
- ❑ Develop a Recurrent Convolutional Generative Adversarial Network (RC-GAN) to improve the fidelity and accuracy of generating images from textual descriptions.
- ❑ Address the limitations of current text-to-image generation methods in capturing semantic nuances and producing high-quality visual outputs.
- ❑ 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

5

“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)

6

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:

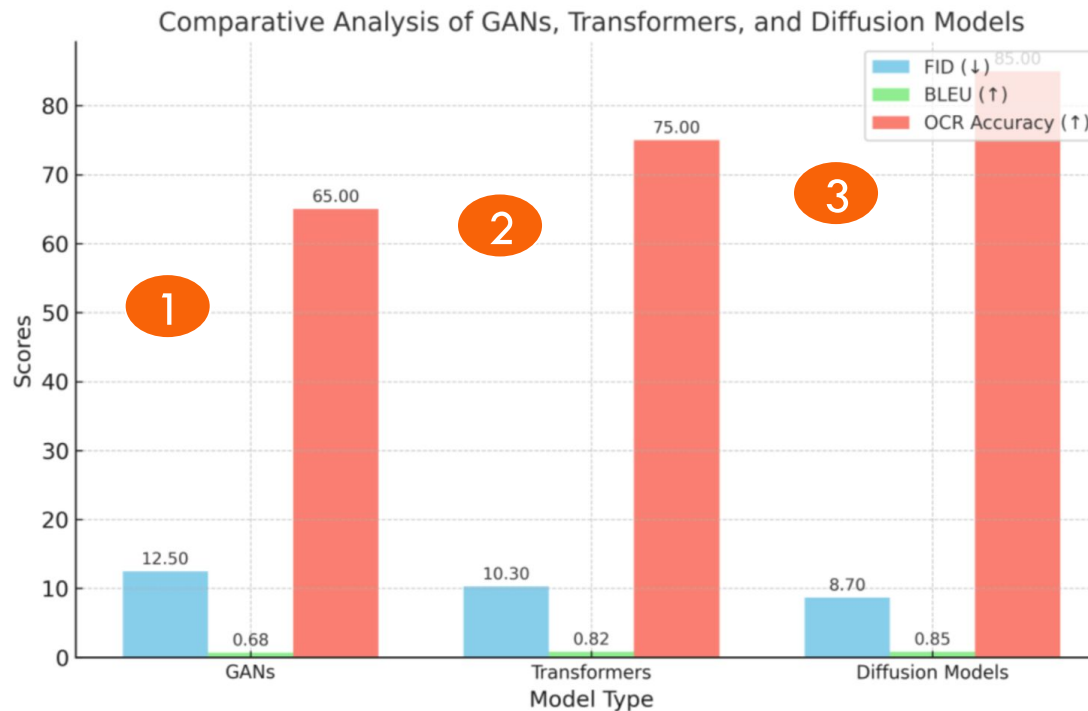
Developing more comprehensive evaluation methods that go beyond quantitative metrics and incorporate subjective, human-centered assessment.



Literature Review (Comparative Analysis)



7



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.

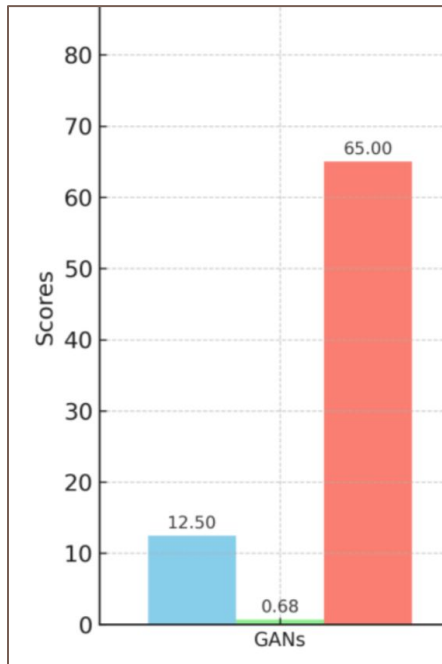
Figure 4: Bar graph showing the comparative analysis of GANs, Transformers, and Diffusion Models based on FID, BLEU, and OCR accuracy. The graph highlights the performance differences across the three model types.

Literature Review (Comparative Analysis)



8

1



GANs - Random samples generated for human faces.

Literature Review (Comparative Analysis)

9

2

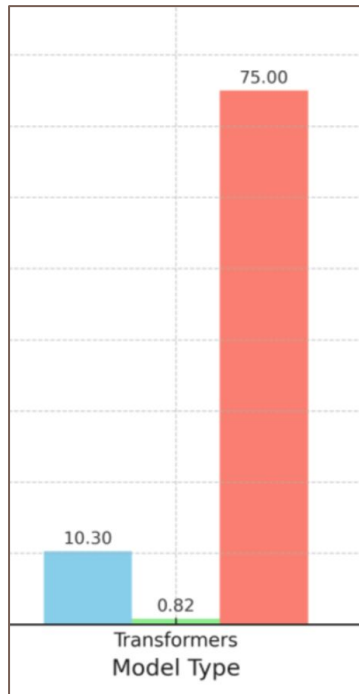


Image Transformer

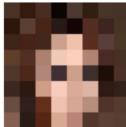


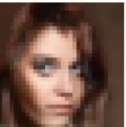

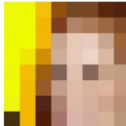






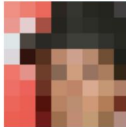





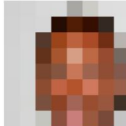













Input	1D Local Attention			2D Local Attention			Original
	$\tau = 0.8$	$\tau = 0.9$	$\tau = 1.0$	$\tau = 0.8$	$\tau = 0.9$	$\tau = 1.0$	
							
							
							
							
							

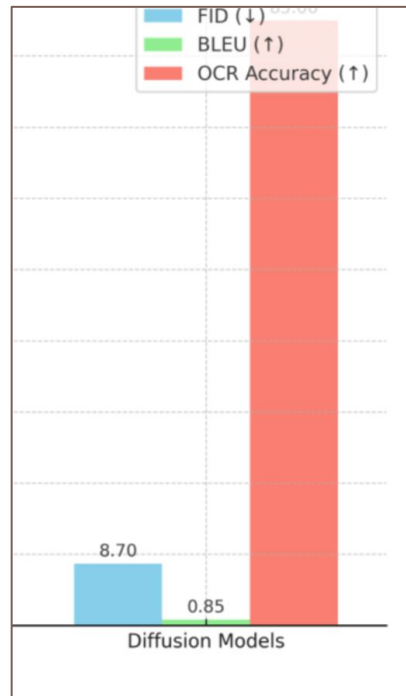
Table 6. Images from our 1D and 2D local attention super-resolution models trained on CelebA, sampled with different temperatures. 2D local attention with $\tau = 0.9$ scored highest in our human evaluation study.

TRANSFORMERS - Random samples generated for human faces.

Literature Review (Comparative Analysis)

10

3



Multi-Modal-Driven Face Generation



DIFFUSION - Random samples generated for human faces.

Challenges in existing systems

11

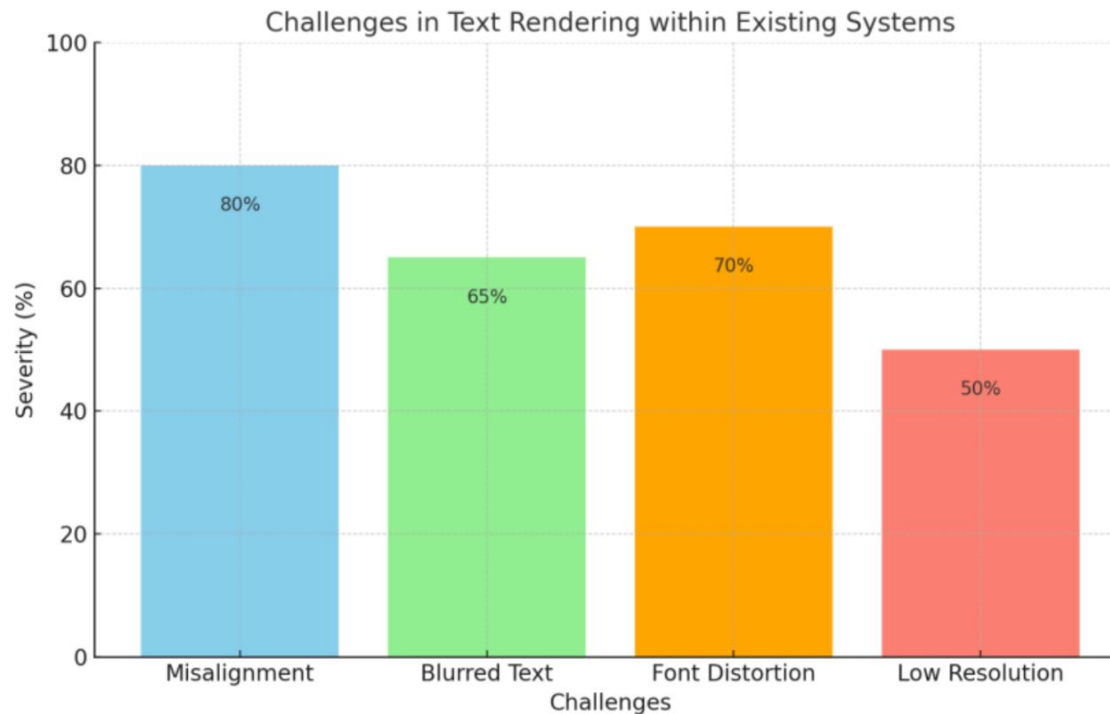


Figure 1: Illustration of text rendering challenges in existing systems, highlighting issues such as misalignment, blurred text, font distortion, and low resolution.

Model Training / Fine-Tuning

12

Fine Tuning / Training

Colab AI

	all	74	68	0.702	0.735	0.748	0.36
Epoch	GPU mem	box_loss	cls_loss	dfl_loss	Instances	Size	
47/50	9.49G	1.545	1.217	1.783	6	640: 100%	92/92 [01:12<00:00, 1.27it/s]
Class	Images	Instances	Box(P	R	mAP50	mAP50-95): 100%	4/4 [00:02<00:00, 1.71it/s]
all	74	68	0.626	0.706	0.738	0.352	
Epoch	GPU mem	box_loss	cls_loss	dfl_loss	Instances	Size	
48/50	9.54G	1.486	1.191	1.761	3	640: 100%	92/92 [01:11<00:00, 1.28it/s]
Class	Images	Instances	Box(P	R	mAP50	mAP50-95): 100%	4/4 [00:02<00:00, 1.70it/s]
all	74	68	0.706	0.708	0.755	0.366	
Recorded with iTop Screen Recorder							
Epoch	GPU mem	box_loss	cls_loss	dfl_loss	Instances	Size	
49/50	9.53G	1.463	1.158	1.721	3	640: 100%	92/92 [01:12<00:00, 1.27it/s]
Class	Images	Instances	Box(P	R	mAP50	mAP50-95): 100%	4/4 [00:02<00:00, 1.57it/s]
all	74	68	0.661	0.721	0.724	0.367	
Epoch	GPU mem	box_loss	cls_loss	dfl_loss	Instances	Size	
50/50	9.56G	1.468	1.175	1.713	1	640: 100%	92/92 [01:12<00:00, 1.27it/s]
Class	Images	Instances	Box(P	R	mAP50	mAP50-95): 100%	4/4 [00:02<00:00, 1.71it/s]
all	74	68	0.689	0.75	0.759	0.358	

50 epochs completed in 1.178 hours.

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

Result



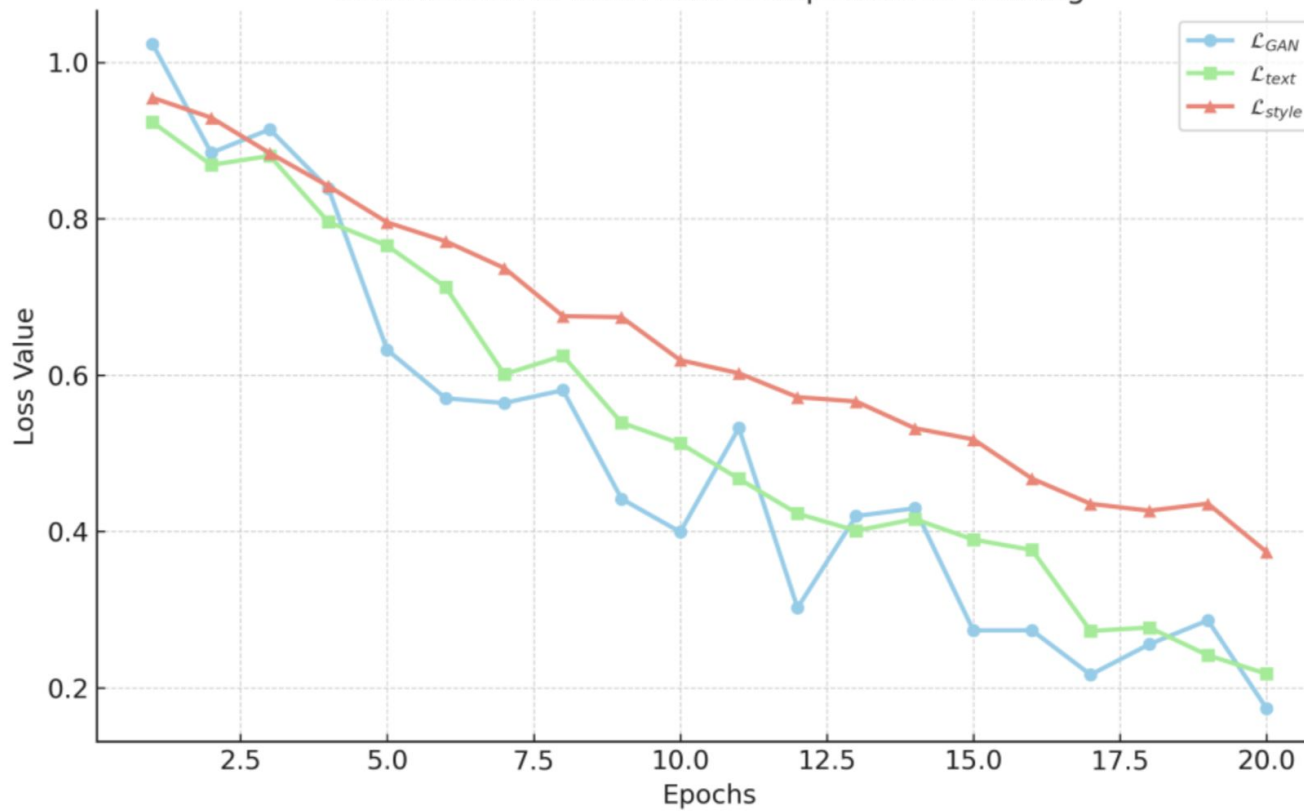
Prompt:

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

Contribution of each loss component to training

13

Contribution of Each Loss Component to Training

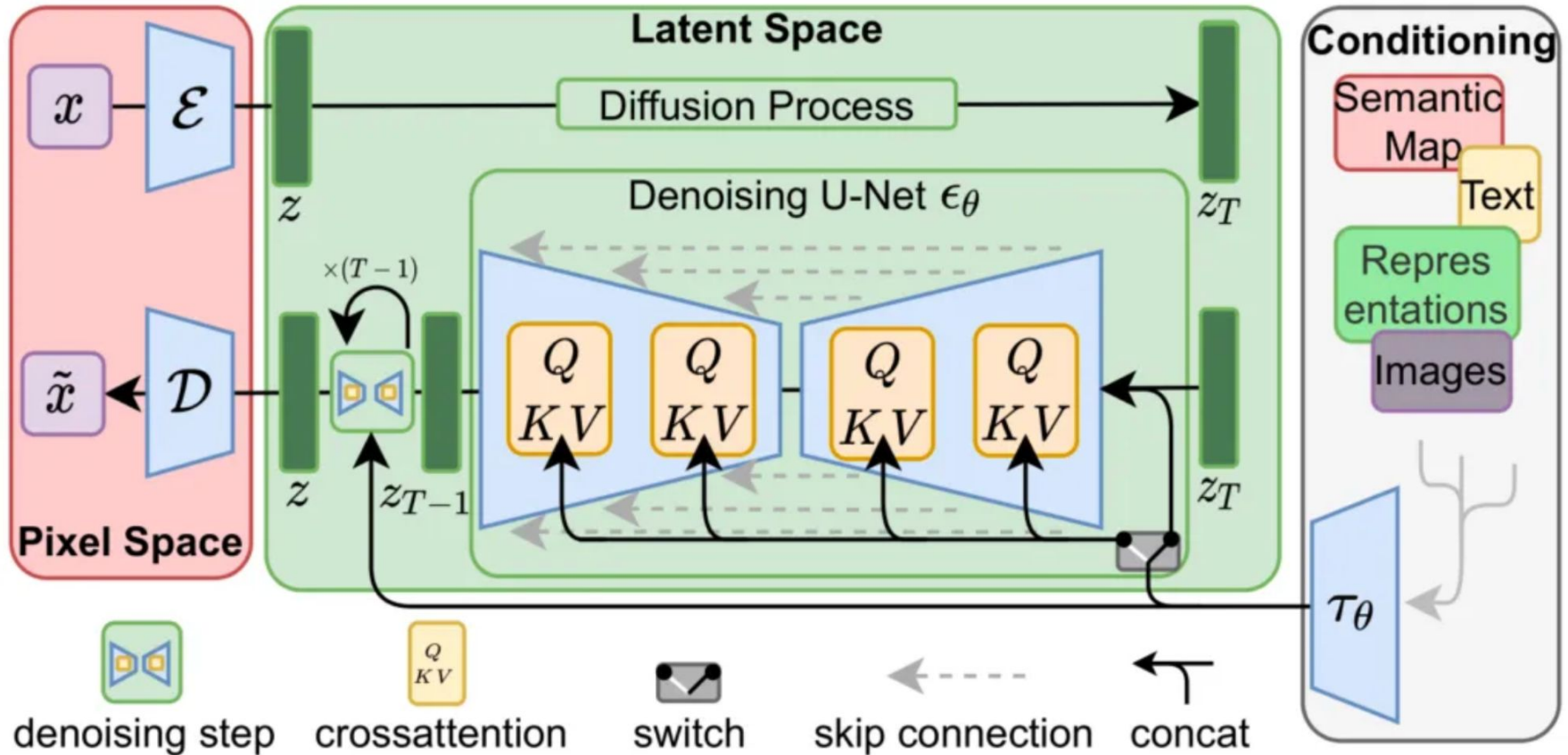


LEGEND

- \mathcal{L}_{GAN} : Adversarial loss for image synthesis.
- \mathcal{L}_{text} : Text rendering loss.
- \mathcal{L}_{style} : Style consistency loss.

How it works - Architecture

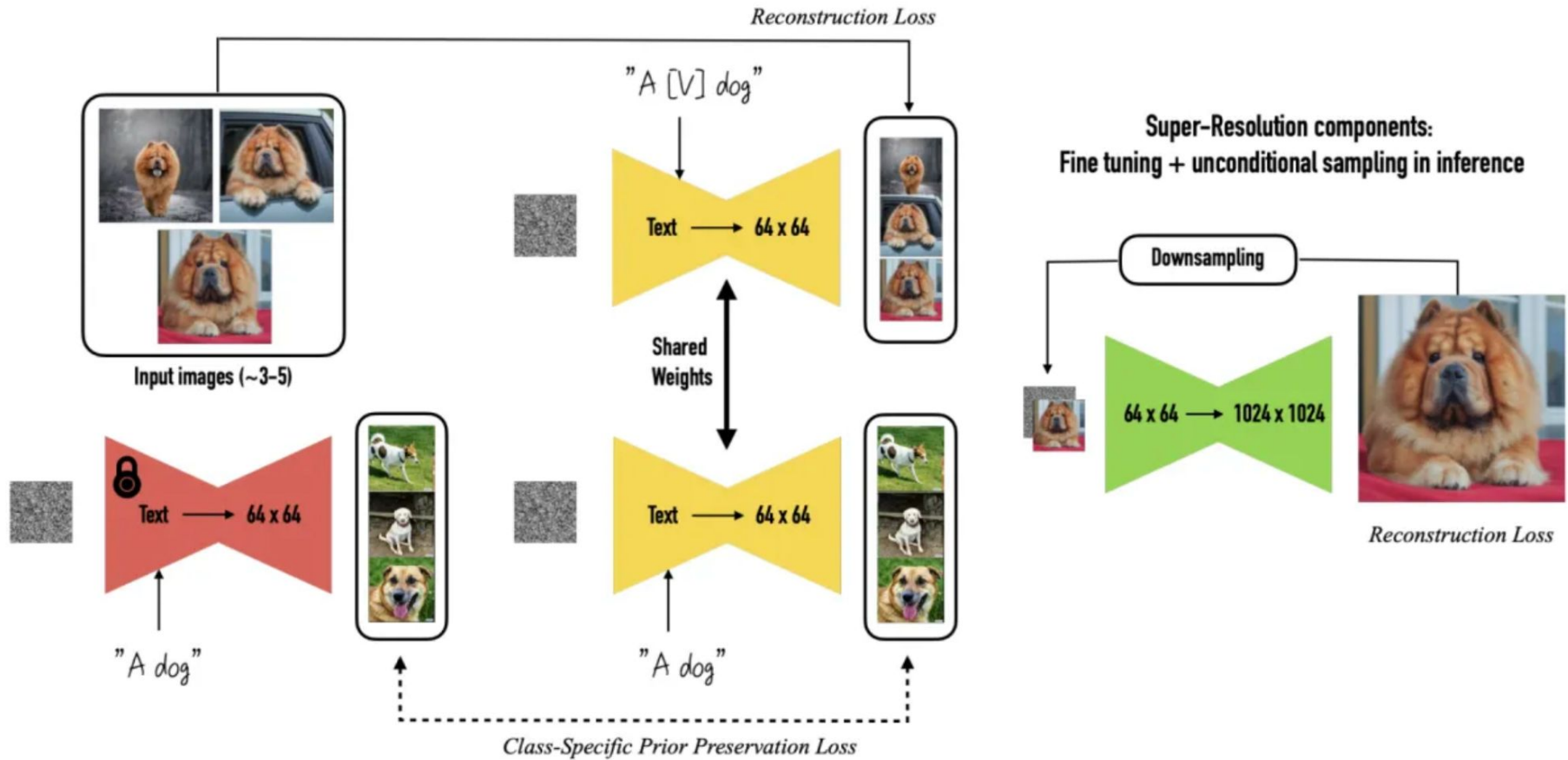
14



Architecture of Stable Diffusion Model

How it works - Architecture

15



Fine tuning a stable diffusion model

Our Methodology

16

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

WHY?

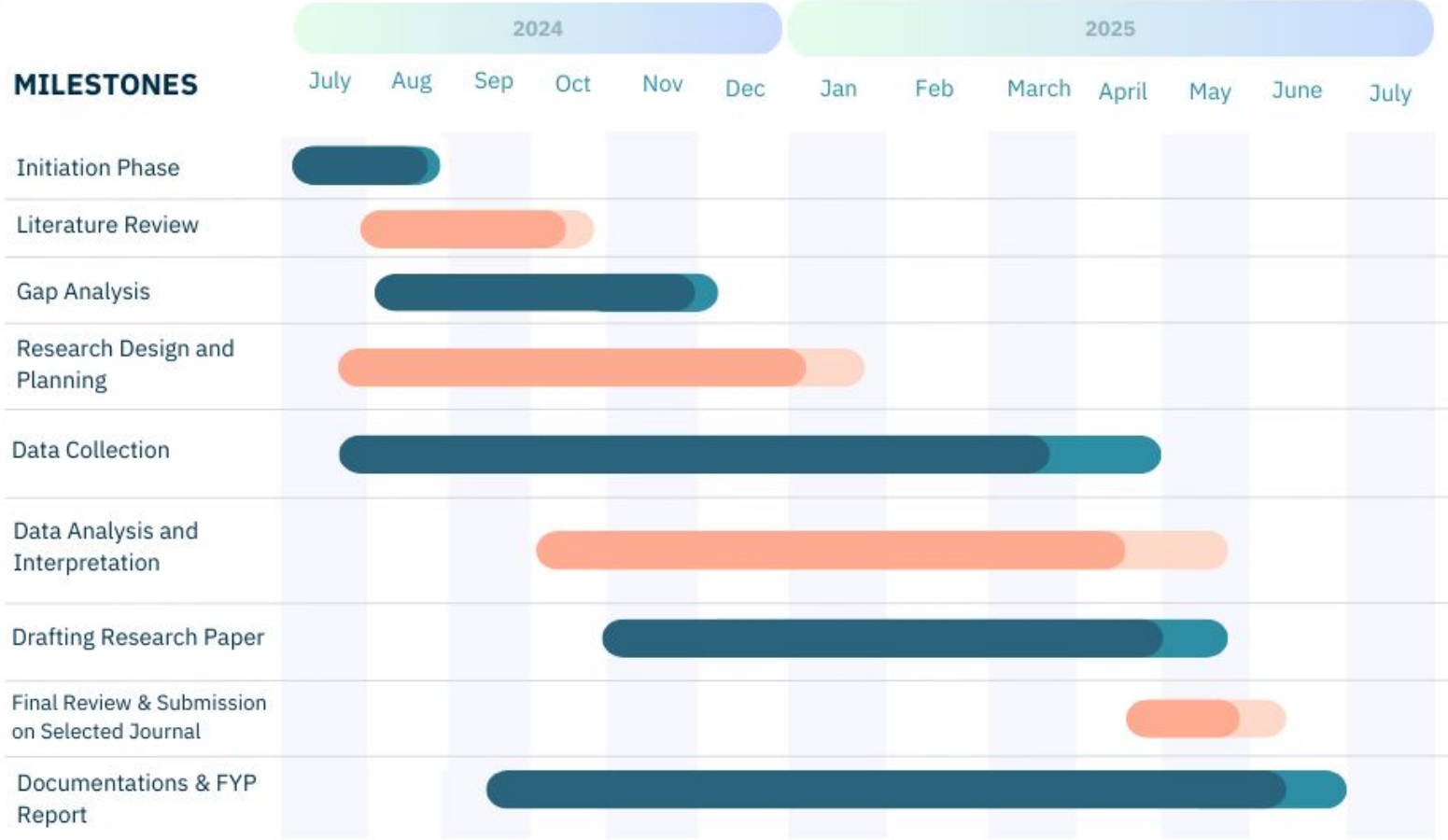
- Agile Scrum is chosen because it facilitates flexibility, iterative development, and continuous feedback loops, which are essential for refining the Recurrent Convolutional Generative Adversarial Network (RC-GAN). The project involves diverse tasks such as data preprocessing, model refinement, and 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

17

2-WEEK SPRINTS FOR ITERATIVE PROGRESS ON DELIVERABLES

MILESTONES



Selected Publication Venue - CVPR



18

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

CVPR CVPR My Stuff

Select Year: (2025) Dates Calls Author & Reviewer Guides Attend Expo Media

The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2025

The **IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)** is the premier annual computer vision event comprising the main conference and several co-located workshops and short courses. With its high quality and low cost, it provides an exceptional value for students, academics and industry researchers.

Register Invitation Letter (Coming soon)



Budget / Costing - Vast.ai



19

#GPUS: ANY 0X 1X 2X 4X 8X 9X+ On-Demand Any GPU Planet Earth Auto Sort									
m:31686	datacenter:68137	France, FR	2x H100 SXM	K14PN-D24 Series PCIe 5.0,16x 54.9 GB/s	↑6949 Mbps ↓8403 Mbps 2499 ports	verified Max Duration 12 days	\$4.271/hr		
vast.ai			107.1 TFLOPS Max CUDA: 12.2	80 GB 2504.4 GB/s 478.1 GB/s	AMD EPYC 9554 ... 64.0/256 cpu 387/1548 GB	SAMSUNG MZQL... 4852 MB/s 1041.5 GB	677.0 DLPerf 158.5 DLP/\$/hr	Reliability 99.04%	RENT
Type #15952754									
m:27587	host:145222	Thailand, TH	1x H100 SXM	SB27B24796 PCIe 5.0,16x 37.2 GB/s	↑2580 Mbps ↓6478 Mbps 691 ports	verified Max Duration 3 days	\$2.138/hr		
vast.ai			53.5 TFLOPS Max CUDA: 12.6	80 GB 2496.2 GB/s 478.1 GB/s	AMD EPYC 9124 ... 16.0/64 cpu 129/516 GB	nvme 2554 MB/s 379.4 GB	322.3 DLPerf 150.8 DLP/\$/hr	Reliability 99.84%	RENT
Type #13482713									
m:30032	datacenter:135125	California, US	1x H100 NVL	MZB3-G41-000 PCIe 5.0,16x 54.9 GB/s	↑7185 Mbps ↓7176 Mbps 249 ports	verified Max Duration 23 mon.	\$2.671/hr		
vast.ai			48.3 TFLOPS Max CUDA: 12.7	94 GB 3357.9 GB/s 318.7 GB/s	AMD EPYC 9124 ... 8.0/64 cpu 193/1548 GB	Micron_7450_MT... 2198 MB/s 2588.8 GB	310.1 DLPerf 116.1 DLP/\$/hr	Reliability 99.77%	RENT
Type #13868216									
m:32379	datacenter:135125	, US	4x H200	DGXH200 PCIe 5.0,16x 53.2 GB/s	↑933 Mbps ↓6316 Mbps 999 ports	verified Max Duration 27 days	\$12.804/hr		
vast.ai			214.1 TFLOPS Max CUDA: 12.7	140 GB 4046.7 GB/s 478.1 GB/s	Xeon® Platinum ... 112.0/224 cpu 1032/2064 GB	SAMSUNG MZW... 37899 MB/s 12875.9 GB	1974.6 DLPerf 154.2 DLP/\$/hr	Reliability 95.9%	RENT
Type #16580543									
m:32379	datacenter:135125	, US	8x H200	DGXH200 PCIe 5.0,16x 53.2 GB/s	↑933 Mbps ↓6316 Mbps 1999 ports	verified Max Duration 27 days	\$25.604/hr		
vast.ai			428.2 TFLOPS Max CUDA: 12.7	140 GB 4046.7 GB/s 478.1 GB/s	Xeon® Platinum ... 224.0/224 cpu 2064/2064 GB	SAMSUNG MZW... 37899 MB/s 25751.7 GB	3784.3 DLPerf 147.8 DLP/\$/hr	Reliability 95.9%	RENT
Type #16580545									
m:25258	datacenter:135125	California, US				verified			

Budget / Costing



20

1. GPU/Vast.ai

We initially added \$700 this semester in vast.ai

2. Colab Pro+

Subscribed to colab pro+ for daily heavy tasks

Credits

Remaining balance

\$201.24

Add Credit

Current Usage

Total:	2.94 \$/hr	70.56 \$/day
GPU:	2.933 \$/hr	70.392 \$/day
Disk:	0.007 \$/hr	0.168 \$/day

You will run out of credits in 2 days, 20 hours, 26 minutes

Colab

20 October 2024 at 22:13:42 GMT-7

ITEM	QUANTITY	PRICE
Colab Pro+	1	US\$49.99
Subscription monthly charge		

You will be charged the subscription cost automatically (currently US\$49.99 plus any applicable tax every month) until you cancel your subscription. Auto-renews on 20 November 2024.

VAT (5%) US\$2.50

Total US\$52.49

Payment method

Mastercard •••• 2195

Order number

COL.3172-9062-6618-86867

View Online

Budget / Costing

21

- **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
	Dataset Acquisition	\$100 - \$500
Data and Resources	Data Annotation	\$200 - \$600
	Research Materials	\$50 - \$150
Team and Development	Conference and Workshop Fees	\$100 - \$300
	Printing and Stationery	\$50 - \$100
Miscellaneous	Travel Expenses	\$100 - \$300
	Contingency	\$200 - \$400
Total		\$4,675 - \$11,700

FYP Deliverables



22

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



23

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!

