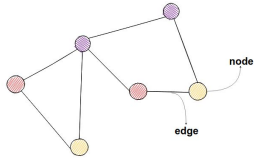
Automated Text Rectification in AI-Generated Visual Content



By
Prajit Sengupta
(CID:06021238)

Supervisor:
Dr Thomas Lancaster

Generated Source: Dall-E3

AI is redefining creativity, but can it master clarity??

Source: Imagen (Copilot)

"A battle between visuals and language"

Source: Emu (MetaAI)

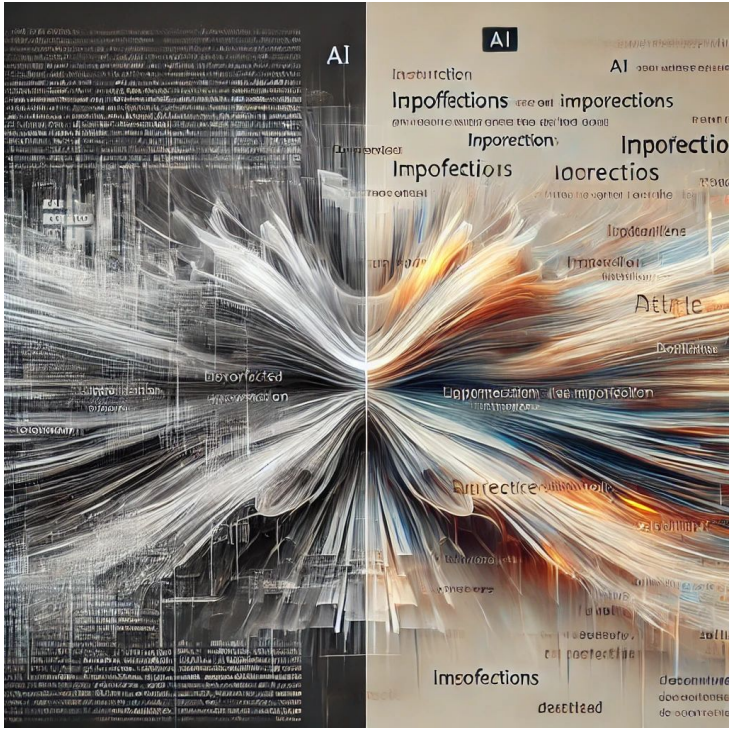
Source: Dall-E3 (ChatGPT)

Prompt: "Create an image of a highway board of the city of London."

Introduction

“A battle between visuals and language"

- AI-generated images (DALL-E, Imagen) often contain **incorrect, gibberish, or misaligned text.**
- Real World Image-Text Pairs can also contain textual errors.(Eg: Scanned Documents)
- **Solution:** GenFix - An automated image-text rectification model!



Source: Dall-E3

How AI Sees and Creates the World!

1. **Variational Autoencoders (VAEs)** - "Compress & Reconstruct"
2. **Autoregressive Models -** "Building Images Pixel by Pixel"
3. **Generative Adversarial Networks (GANs)** - "AI vs AI in a Creativity Battle"
4. **Diffusion Models** - "Start with Noise, End with Clarity"

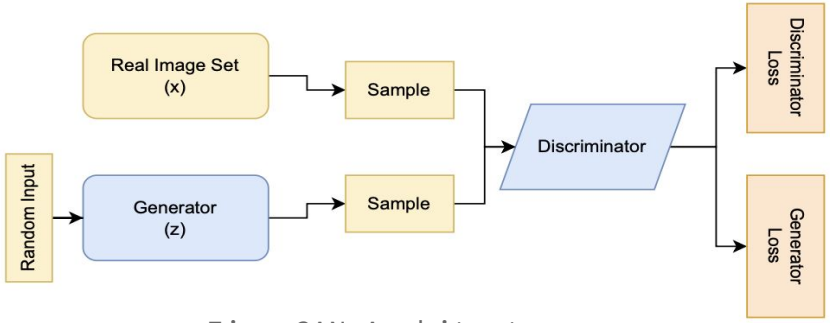


Fig: GAN Architecture

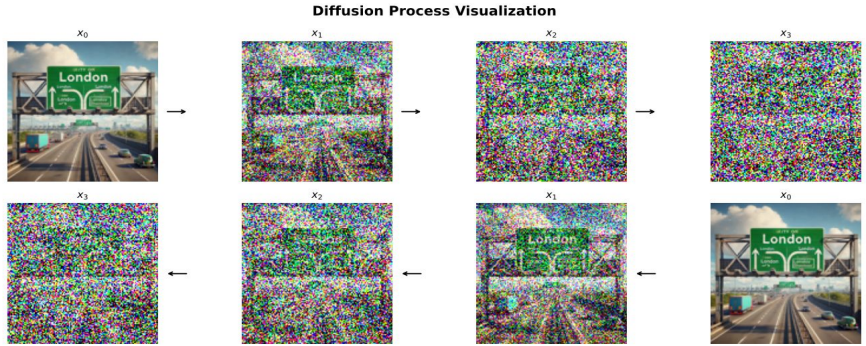


Fig: Diffusion Process

Challenges in AI Image-Text Generation

- **Focus on visual elements over textual content.** Text is often treated as a shape rather than meaningful characters, leading to gibberish outputs. [1]
- **Complexity of language and context.** Language is not just about words—it's about meaning, context, and relationships. [2]
- **Limited training data for text-in-images.** Scene text datasets (e.g., COCO-Text) are small compared to general image datasets, leading to poor generalization. [3]
- Difficult to **organize text boxes** around AI generated images compared to real word images

[1]- Lakhanpal et al, https://arxiv.org/abs/2403.16422
[2] - SM et al, https://aclanthology.org/2020.emnlp-demos.21/.
[3] - V et al, https://doi.org/10.1038/s41598-024-79705-4.

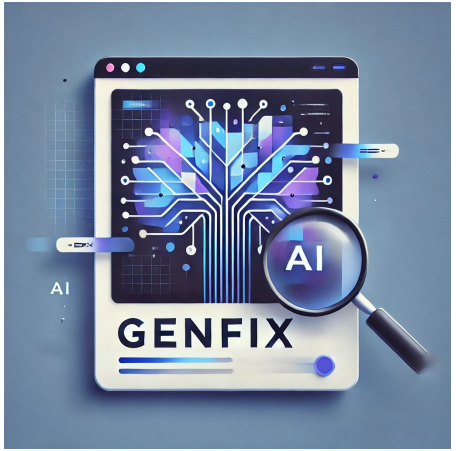
Building on Existing Research

Method	Approach	Limitations
STEFANN (2021)[1]	Font-Adaptive Neural Network (FANnet)	Fails on curved, artistic, and multi-lingual text
SRNet (2019)[2]	Style-aware GAN-based inpainting	Struggles with complex real-world images
AnyText (2024)[3]	Text-centric transformer model	Requires manual user input
TextDiffuser (2023)[4]	Diffusion model for text inpainting	Needs explicit text prompts, lacks automation
SwapText (2020)[5]	Text style transfer with deep features	Poor adaptation to real-world AI-generated text distortions

[1] *STEFANN:* https://openaccess.thecvf.com/content_CVPR_2020/papers/Roy_STEFANN_Scene_Text_Editor_Using_Font_Adaptive_Neural_Network_CVPR_2020_paper.pdf
[2] *SRNet: Style Retention Network for Text Inpainting* (2019) https://arxiv.org/abs/1908.03047
[3] *AnyText: A Transformer-Based Framework for AI-Assisted Text Editing* https://arxiv.org/abs/2311.03054
[4] *TextDiffuser: Diffusion Model for Text Image Restoration* https://jingyechen.github.io/textdiffuser/
[5] *SwapText: Image-based Text Style Transfer* https://arxiv.org/abs/2003.08152

"From Fragmented Tools to a Unified Brain."

- **GenFix** is an intelligent, **end-to-end pipeline** that understands, aligns, and corrects text within AI-generated images.
- GenFix is a ***context-aware***, ***font-adaptive***, and ***style-preserving*** image text rectification system.
- Think of GenFix as an **autonomous proofreader** for AI artists—it fixes text before it reaches the public eye.

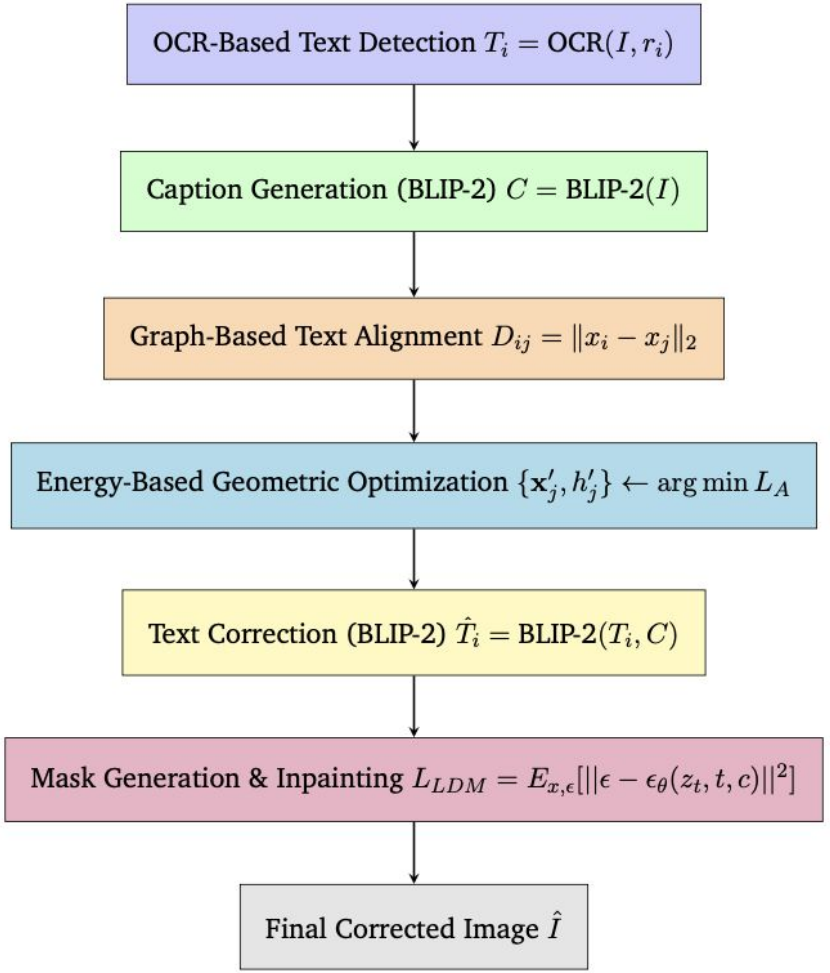


Source:Dall-E3

“Turning AI gibberish into human language.”

GenFix Pipeline: A Step-by-Step Approach

- **Step 1:** Detect & Extract text (OCR)
- **Step 2:** Generate context-aware captions (BLIP-2)
- **Step 3:** Align detected and corrected text (Graph-Based Matching)
- **Step 4:** Optimize text layout (Energy-Based Optimization)
- **Step 5:** Inpaint corrected text using Stable Diffusion



OCR- Text Extraction & Recognition

- Multiple experiments were performed with various OCR Models.
- The TrOCR model achieved the best performance in terms of Character Error Rate (CER), which was the lowest among all models.
- In the GenFix model **TrOCR**(for recognizing) + **EasyOCR** (bounding box detection) was used

Method	Dataset	Metric	Score	Remarks
Tesseract v4(31)	ICDAR 2015	CER	12.5	Struggles with noisy images
CRNN(32)	COCO-Text	CER	8.7	Handles curved text well
EAST(33)	ICDAR 2015	F1-Score	82.1%	Effective for scene text detection
TrOCR (9)	Synthetic + IAM	CER	4.22	Lightweight and Effective
PaddleOCR(34)	SynthText	CER	7.5	Lightweight and versatile
Google Vision OCR(35)	ICDAR 2015	WER	5.3	Best for multilingual scenarios

Table: Performance Comparison of OCR Methods on Benchmarks

Context & Caption Generation

- **BLIP-2 model** is used to generate a contextual description of the image.
- **Original image:** A stop sign with a spelling mistake ("SOTP")
  BLIP-2 caption: "A high quality photo of a Stop Sign"

Input Image


```
Processing image: /content/Incorrect_SOTP_sign.jpg
Running OCR on image...
OCR Results: [[[[1127, 907], [2303, 907], [2303, 1484], [1127, 1484]], 'SOTP']]
Hello
0.0
Detected 1 text regions
Image context: a high quality photo of a stop sign

/usr/local/lib/python3.11/dist-packages/transformers/generation/configuration_utils.py:628: UserWarning: `do_sa
  warnings.warn(
Region 1: 'SOTP' → 'Correct this text in image context: SOTP. Image shows: a high quality photo of a stop sign
. Correction must be: correct the spelling or, more likely to me; it should say STOP! and not S-O - T P . Corre
```

Source: Kernel Output (VSCode)

Graph-based Alignment

- AI-generated images often have non-linear text layouts (e.g., curved, scattered), leading to OCR misalignment.
- Cost matrix D where dij represents the edit distance between Ti and Sj.
- Hungarian algorithm to find the optimal assignment σ*

$$D = \begin{bmatrix} d_{11} & d_{12} & \dots & d_{1M} \\ d_{21} & d_{22} & \dots & d_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N1} & d_{N2} & \dots & d_{NM} \end{bmatrix}$$

$$\sigma^* = \arg\min_{\sigma} \sum_{i=1}^{N} D_{i,\sigma(i)}$$

Ti- Detected text regions
Sj- Corrected words

Energy-Based Geometric Optimization

- **Fidelity:** Penalizes deviations from original positions(λ_1) and heights(λ_2).
- **Spacing:** Enforces horizontal alignment with gap d(μ).
- **Uniformity**: Ensures consistent text height(ν).

$$L_A = \sum_{i=1}^{N} \left[\lambda_1 \|\mathbf{x}_i' - \mathbf{x}_i\|^2 + \lambda_2 (h_i' - h_i)^2\right] + \sum_{i=2}^{N} \left[\mu \|\mathbf{x}_i' - \mathbf{x}_{i-1}' - [d, 0]\|^2 + \nu (h_i' - h_{i-1}')^2\right]$$

Eqn: Minimize energy L_A

Mask generation & Inpainting

- **Stable Diffusion Inpainting Model** used (Pre-trained from Hugging-Face)
- Uses **latent diffusion** to intelligently predict the missing text.
- A **binary mask** is generated as per the previous pipeline output to isolate incorrect text, defining the inpainting area.

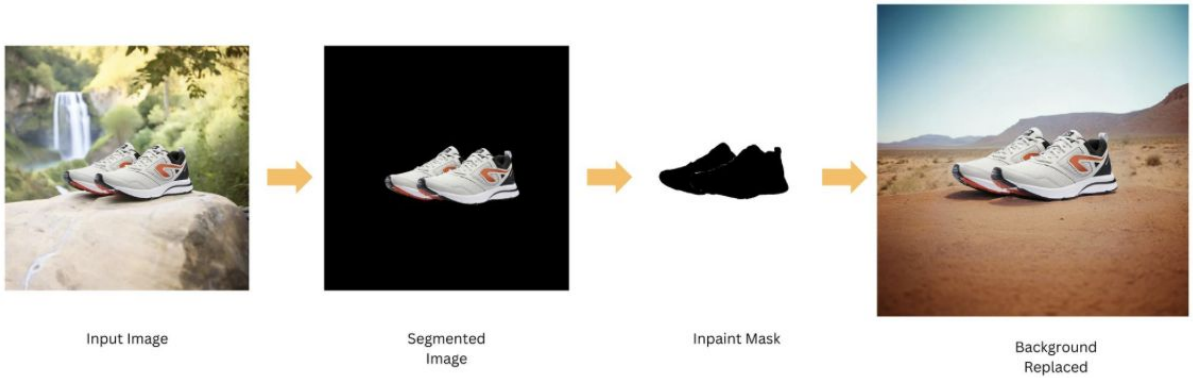


Image Source: Segmind Blog

Fig: Inpainting Process

TextSynth-100 Dataset

- **A Benchmark Dataset** – 100 image-text pairs designed to evaluate AI-generated text in images.
- **Diverse AI Model Coverage** – Includes outputs from DALL·E3 (ChatGPT), Copilot, Imagen (Gemini), and Emu (Meta).
- **Everyday Text, Everywhere** – Precise prompts – From street signs to café menus.

Note: The Dataset will be further expanded

Model evaluation

- High structural similarity **(SSIM: 0.9555)** preserves image integrity.
- GenFix achieves perfect text accuracy **(CER & WER = 0.0)**, ensuring precise corrections.
- GenFix performs well for shorter words(1-3) but starts introducing errors in longer word length (4-5).

Metric	Value	Interpretation
CER	‘0.0‘	Perfect character-level match between corrected and target texts.
WER	‘0.0‘	Perfect word-level match between corrected and target texts.
BLEU	‘0.1778‘	Misleading due to short texts; CER/WER are more reliable here (perfect).
SSIM	‘0.9555‘	Excellent structural similarity between original and corrected images.
PSNR	‘18.041 dB‘	Moderate pixel-level differences between original and corrected images.
FID	‘58.2983‘	Moderate feature-level differences; may indicate substantial corrections.

Table: Evaluation Metrics using the TextSynth Dataset

Text Length	F1 Score	Word Accuracy
1	1.0	1.0
2	1.0	1.0
3	1.0	1.0
4	0.8	0.8
5	0.6	0.8

Table: Performance of the text correction pipeline across different text lengths.

Experimental Results

- GenFix shows **superior** text accuracy and structural preservation.
- Highest structural preservation (SSIM: 0.9555) vs. STEFANN (0.7712), SRNet (0.79), Pix2Pix (0.63).
- Balanced correction with minimal distortion (PSNR: 18.041 dB).

Note: FID was evaluated on different datasets.

Method	SSIM / ASSIM	PSNR	FID
GenFix	**0.9555**	18.041	High
STEFANN (FANnet) (50)	0.7712	-	High
SRNet (24)	0.79	21.12	High
Pix2Pix (53)	0.63	16.54	Moderate

Table: Performance comparison of GenFix with existing methods

Ablation Study

"What If We Remove Key Parts? Let's Break GenFix!"

- **Graph-Based Text Alignment** Improves Accuracy - Without it, word placements were inconsistent. [Simulated Annealing]
- **OCR Inpainting** (fixing text regions using an OCR-based method) provides a bigger boost to 0.850.
- **BLIP-2 Context** Helps Avoid Over Corrections - Without it, the model sometimes changed correct words unnecessarily.

Method	Avg F1 Score
Baseline (No Correction)	0.750
With Simulated Annealing	0.800
With OCR In-painting	0.850
Full Model	0.950

Images with Spelling Errors!

Correcting spelling errors (e.g., “SOTP” → “STOP”). The contrast of the text font was also improved.

Images with Contextually Incorrect Text!

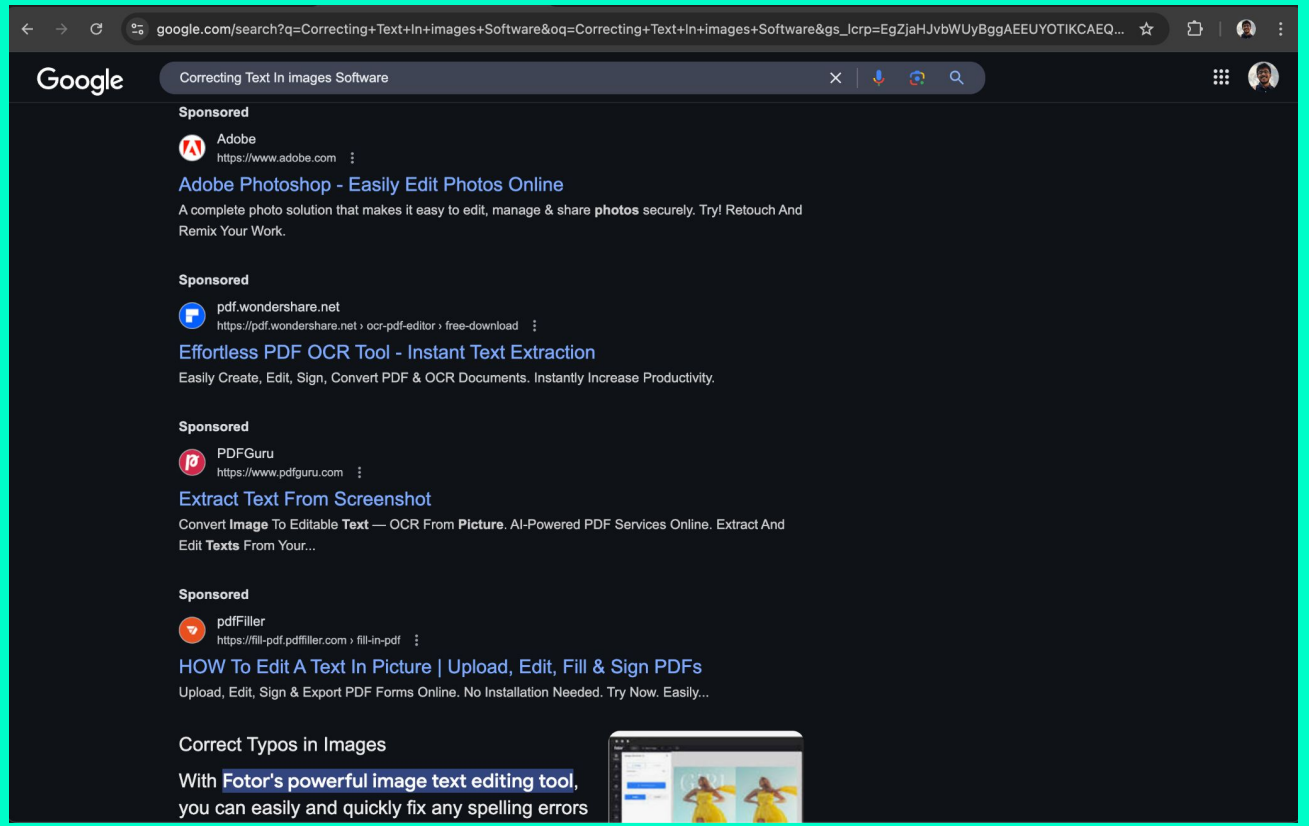
Removing redundant words (e.g., “Happy Happy Birthday” → “Happy Birthday”)

Scene Aware Text Generation!

The models tries to find words related to the scene such as ”Road” or ”Signboard” as shown in Figure.

Conclusion

- GenFix provides the first open-source automated text rectification system.
- Bridges the gap between AI-generated images and human-readable text.
- Unlike traditional tools, GenFix is an intelligent rectification system, not just an editor.



Search for any automated text correction in image tool!!!!

To be submitted to

ICCV Conference(CVF)

[https://iccv.thecvf.com/Conferences/2025/CallForPapers]

Scholar: https://scholar.google.com/citations?user=eQdWgWEAAAAJ&hl=en

Paper Link: https://arxiv.org/submit/6233005/view

Automated Text Rectification in AI-Generated Visual Content

Prajit Sengupta*
Department of Computing
Imperial College London
prajit.sengupta24@imperial.ac.uk

Thomas Lancaster
Department of Computing
Imperial College London
t.lancaster@imperial.ac.uk

Abstract

In a world where images with text are ubiquitous—on social media, advertisements, scanned documents, and AI-generated content—errors in textual content reduce credibility and usability. This paper introduces **GenFix**, the first open-source, fully automated model designed to detect, correct, and seamlessly integrate text in images while maintaining style consistency. GenFix combines vision-language models with inpainting techniques to achieve accurate, context-aware corrections. It ensures textual accuracy with a perfect Word Error Rate (WER) and Character Error Rate (CER) of 1, while preserving the visual integrity of images, achieving a Structural Similarity Index (SSIM) of 0.9555. To support evaluation, we introduce **TextSynth-100**, a benchmark dataset of AI-generated image-text pairs specifically designed for text correction models. This research aims to bridge the gap between what we see and what we understand, making text in images both accurate and visually consistent.[1]

1 Introduction

AI systems have demonstrated remarkable capabilities in a variety of fields, from acing competitive exams to generating code for complex workflows and even outperforming e-sports champions. Generative AI applications, such as ChatGPT [1], SORA [2], Dall-E [3] and Emu [4] have been making headlines for generating innovative text to images and video content. However, while these models do demonstrate great capabilities through their creativity and precision in generating varied visual and textual content, they do often fail when it comes to generating accurate meaningful text within images. [5]

Most of the times, the text generated within these images appears to be gibberish or some nonsensical sequence of characters. Such limitation arises due to the challenges of integrating LLMs (Large Language Models) with Image Generation Models, particularly in capturing the precise details of characters within an image [6]. Correcting such flawed text in generated images is essential for usability purposes and for the adoption of AI-generated visual content across various industries.

This problem has a significant impact in domains like advertising media, digital art, articles and automated content creation, where text within images often plays a crucial role in delivering clear and meaningful messages. Addressing this problem would require a multidisciplinary approach, combining concepts from computer vision, natural language processing, and certain AI techniques which we will discuss later in detail.

*Corresponding author: prajit.sengupta24@imperial.ac.uk

[1]This paper is based on research conducted as part of the Independent Study Option module for MSc Advanced Computing at Imperial College London

"Can we trust what we see if AI can't even trust what it writes?"

THANK YOU!!!!!