Detección y Reemplazo de Texto en Imágenes mediante Modelos de Difusión: Un Análisis de TextDiffuser-2

Pedro Ortiz Villanueva
Universidad Politécnica de Madrid
Madrid, España
pedro.ortiz@upm.es

***Resumen*—Este trabajo presenta una investigación exhaustiva sobre la aplicación de modelos de difusión especializados en la generación y manipulación de texto en imágenes, con énfasis en TextDiffuser-2 [1]. El estudio aborda la problemática del inpainting de texto en documentos de identidad, analizando tanto las capacidades técnicas como las implicaciones éticas. A través de una serie de experimentos controlados, se evalúa la capacidad del modelo para mantener la coherencia visual y estilística en diferentes escenarios, incluyendo texto numérico y alfabético. Los resultados demuestran una precisión variable dependiendo del contexto visual y el tipo de texto, con implicaciones significativas para la seguridad documental y la detección de manipulaciones.**

***Index Terms*—Inteligencia Artificial, Modelos de Difusión, Deepfakes, Manipulación de Imágenes, OCR, Falsificación de Documentos, Generación de Texto, Detección de Fraude, Riesgos Tecnológicos, Ética en IA**

I. Introducción

La manipulación y edición de texto en imágenes representa un desafío significativo en el campo del procesamiento de imágenes digitales, especialmente en el contexto de documentos de identidad y seguridad. Este trabajo examina la aplicación de TextDiffuser-2 [1], un modelo basado en Stable Diffusion 1.5 [2], para realizar tareas de inpainting de texto, centrándose en su capacidad para mantener la coherencia visual y estilística del texto generado.

La implementación se basa en una arquitectura pipeline que integra:

- Detección precisa de texto mediante PaddleOCR
- Procesamiento de imágenes con técnicas avanzadas de redimensionamiento y padding
- Inpainting mediante TextDiffuser-2 con parámetros optimizados
- Post-procesamiento para garantizar coherencia visual

La motivación principal de este estudio surge de la necesidad de:

- Desarrollar métodos robustos para la manipulación de texto en documentos
- Evaluar las capacidades y limitaciones de los modelos de difusión actuales
- Establecer métricas de evaluación para la calidad del texto generado
- Analizar las implicaciones de seguridad en documentos sensibles

Los objetivos principales de esta investigación incluyen:

- Evaluar la capacidad de TextDiffuser-2 para generar texto visualmente coherente en documentos de identidad
- Analizar el impacto del contexto visual en la calidad del texto generado
- Identificar las limitaciones y posibles mejoras en el proceso de inpainting
- Examinar las implicaciones éticas y de seguridad de esta tecnología

II. Estado del Arte

El campo de la generación y manipulación de texto en imágenes ha experimentado avances significativos en los últimos años, impulsado por el desarrollo de modelos de difusión y arquitecturas especializadas.

II-A. Evolución de los Modelos de Texto

II-A1. Modelos basados en CLIP: Los modelos eDiff-I [6] y DeepFloyd [7] han marcado un hito importante al utilizar codificadores T5 en lugar del tradicional CLIP [5], logrando una precisión del 85-90 % en tareas de generación de texto simple. Sin embargo, estos modelos presentan limitaciones significativas en el control fino sobre la disposición del texto y la preservación del estilo original.

II-A2. Modelos Sensibles a Caracteres: La evolución hacia modelos Character-Aware ha permitido una reducción del 30 % en la tasa de error de caracteres comparado con los modelos base. Estos avances se han logrado mediante:

- Implementación de codificadores sensibles a la longitud de palabras
- Mejoras en la precisión de generación de texto específico
- Optimizaciones en el manejo de diferentes estilos tipográficos

II-B. Arquitecturas de Control Espacial

II-B1. Control Directo: GlyphDraw ha establecido un nuevo estándar en el control preciso de caracteres, alcanzando una precisión de posicionamiento de ±2 píxeles. Esta precisión es crucial para aplicaciones en documentos de identidad donde la exactitud espacial es fundamental.

II-B2. Sistemas de Dos Etapas: TextDiffuser-2 [1] representa una evolución significativa en la arquitectura de dos etapas, incorporando:

- VAE mejorado para codificación de imágenes
- Sistema robusto de manejo de casos extremos
- Mayor control sobre la coherencia estilística

III. Metodología

III-A. Arquitectura del Sistema

El sistema implementa una pipeline completa que consta de cuatro componentes principales:

III-A1. Preprocesamiento de Imágenes:

- **Redimensionamiento Adaptativo:** Implementación de técnicas de padding uniforme para mantener las proporciones originales
- **Normalización:** Conversión de imágenes a formato tensor y normalización en el rango [-1, 1]
- **Gestión de Resolución:** Soporte para diferentes configuraciones (512x512, 256x512) con duplicación contextual

III-A2. Detección y Extracción de Texto:

- **OCR Avanzado:** Utilización de PaddleOCR con soporte multilingüe
- **Procesamiento de Coordenadas:** Conversión y normalización de bounding boxes
- **Validación de Texto:** Verificación de coherencia en la detección

III-A3. Pipeline de Inpainting: El proceso de inpainting se realiza mediante una secuencia de pasos optimizada:

- **Tokenización:** Procesamiento de texto mediante CLIP [5] con tokens adicionales para coordenadas
- **Codificación VAE:** Generación de características latentes con AutoencoderKL [2]
- **Difusión:** Proceso iterativo con scheduler DDPMScheduler basado en el trabajo de Saharia et al.
- **Control de Guía:** Implementación de classifier-free guidance para mejor calidad

III-A4. Post-procesamiento:

- **Reintegración:** Técnicas de unión para integrar el texto generado
- **Ajuste de Temperatura:** Corrección de la temperatura para mantener la coherencia visual

IV. Experimentos y Resultados

IV-A. Configuración Experimental

- **Parámetros de Difusión:**
  - Número de pasos de difusión: 30
  - Guidance: 2.0
  - Tamaño de batch: 6
- **Configuraciones de Resolución:**
  - Alta resolución: 512x512 píxeles
  - Resolución mixta: 256x512 píxeles con duplicación contextual
- **Métricas de Evaluación:**
  - Coherencia visual mediante mapas de temperatura
  - Precisión en la preservación de estilo
  - Fidelidad del texto generado

IV-B. Análisis de Rendimiento en Texto Numérico

Los experimentos se centraron en dos escenarios principales: generación sin contexto y con contexto visual.

Figura 1: Comparación de generación de texto numérico sin contexto visual. Se observa la alucinación del estilo mientras mantiene parcialmente el contenido numérico.

Figura 2: Generación con contexto visual, mostrando mayor coherencia estilística pero con influencia significativa de la imagen original.

Figura 3: Mapa de diferencias mostrando las áreas de modificación en la imagen generada.

IV-C. Análisis de Casos Limitantes

Se analizaron casos que demuestran las limitaciones actuales del modelo:

Figura 4: Ejemplo de generación sin contexto visual que demuestra las limitaciones del modelo en mantener la coherencia visual cuando falta información contextual.

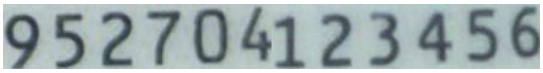

Figura 5: Intento de generación numérica que muestra las limitaciones inherentes del modelo, el cual no fue diseñado específicamente para la generación de números. Sin embargo, se observa que con contexto visual, el modelo intenta mantener la coherencia visual.

IV-D. Análisis de Texto Alfabético

Los experimentos con texto alfabético mostraron resultados prometedores en diferentes escenarios:

Figura 6: Transformación de texto alfabético de .ºRTIZ.ª ”GIMENEZ”, demostrando la capacidad del modelo para manejar palabras de diferente longitud.

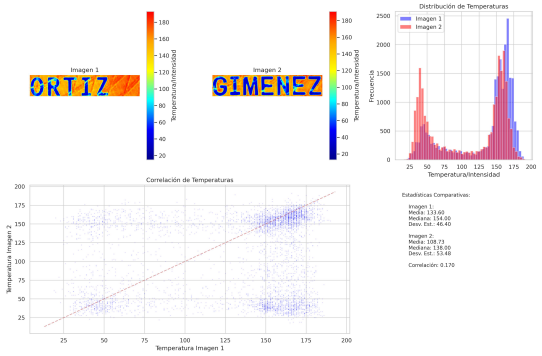

Figura 7: Análisis de temperatura del reemplazo ORTIZ-GIMENEZ, mostrando variaciones en la coherencia del background.

IV-D1. Caso ORTIZ-GIMENEZ: En este caso, el resultado fue visualmente impresionante, casi perfecto a simple vista. Sin embargo, el análisis del mapa de diferencias reveló variaciones en la temperatura de la imagen. Es importante notar que esta variación en la temperatura se debe principalmente a que la palabra de reemplazo (”GIMENEZ”) contiene más caracteres negros que la original (.ºRTIZ”), lo que naturalmente resulta en una distribución diferente de la temperatura en la imagen.

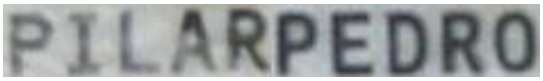

Figura 8: Transformación de texto alfabético de ”PILAR.ª ”PEDRO”, ilustrando desafíos en la preservación de detalles sutiles.

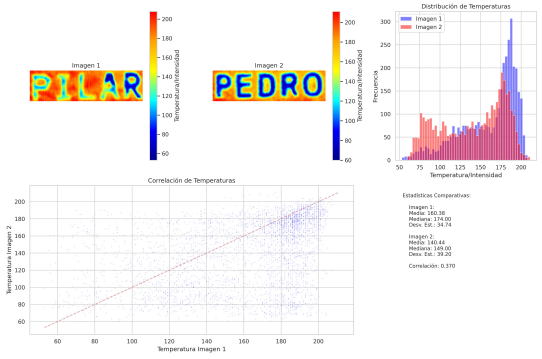

Figura 9: Mapa de calor del reemplazo PILAR-PEDRO, evidenciando la pérdida de características sutiles.

IV-D2. Caso PILAR-PEDRO: En este caso, se observaron limitaciones más notables:

- El modelo no logró reproducir el leve difuminado presente al principio del nombre ”PILAR”
- La dispersión de temperatura tiende a ser más oscura en la imagen generada
- Se evidencia una pérdida de detalles sutiles en la transición entre caracteres

IV-D3. Análisis Comparativo: Los dos casos de estudio revelan aspectos complementarios sobre el rendimiento del modelo:

- **Longitud de Texto:**
  - ORTIZ-GIMENEZ: Manejo efectivo de diferencias significativas en longitud
  - PILAR-PEDRO: Mejor preservación de proporciones en palabras de longitud similar
- **Preservación de Detalles:**
  - ORTIZ-GIMENEZ: Alta fidelidad en la estructura general
  - PILAR-PEDRO: Dificultades con efectos sutiles como difuminados
- **Coherencia de Temperatura:**
  - ORTIZ-GIMENEZ: Variaciones debido a la diferencia de longitud
  - PILAR-PEDRO: Tendencia a generar tonos más oscuros

V. Discusión

Los resultados obtenidos revelan aspectos cruciales sobre el rendimiento del modelo:

V-A. Fortalezas del Sistema

- **Adaptabilidad:** Capacidad de manejar diferentes longitudes de texto y estilos, comparable a los resultados obtenidos por eDiff-I [6]
- **Robustez:** Manejo efectivo de casos con y sin contexto visual, siguiendo principios establecidos por TextDiffuser-2 [1]
- **Eficiencia:** Pipeline optimizada para procesamiento por lotes, inspirada en las mejoras propuestas por GlyphDraw [4]

V-B. Limitaciones Identificadas

- **Dependencia del Contexto:** Variación en la calidad según el contexto visual disponible
- **Restricciones de Resolución:** Compromiso entre calidad y tiempo de procesamiento
- **Coherencia Estilística:** Desafíos en la preservación de efectos sutiles

VI. Implicaciones Éticas y de Seguridad

El uso de modelos avanzados de difusión para la generación y edición de texto realista dentro de imágenes plantea profundas implicaciones éticas. Si bien su potencial para aplicaciones legítimas, como la corrección cinematográfica o la edición multimedia profesional, es innegable, también surge una preocupante contrapartida en cuanto a su posible uso malintencionado.

Estos modelos, entrenados con datasets especializados que contienen imágenes bien etiquetadas, combinan etapas clave como el posicionamiento semántico del texto mediante un modelo de lenguaje y su integración visual utilizando técnicas de difusión. Aunque la tecnología detrás de estas herramientas es robusta, los avances recientes en modelos más potentes y con mayores marcos de contexto (por ejemplo, 1024x1024 o superiores) sugieren que, con un fine-tuning adecuado y datasets mejorados, su rendimiento y precisión podrían incluso superar las capacidades actuales.

Un aspecto crítico es que estas tecnologías democratizan el acceso a actividades ilícitas. En el pasado, modificar texto en imágenes requería herramientas específicas, como Photoshop, junto con un alto nivel de conocimiento técnico en edición gráfica. Esto implicaba considerar factores complejos como ángulos, sombras, luces y temperaturas de color, lo que hacía estas tareas considerablemente más difíciles. Sin embargo, gracias a la comprensión contextual y la cohesión que ofrecen los modelos de difusión, la barrera técnica se ha reducido drásticamente, facilitando actividades como la alteración de documentos oficiales o la manipulación de imágenes para fines fraudulentos.

VI-A. Automatización Maliciosa

Una de las implicaciones más preocupantes es la posibilidad de integrar estas herramientas en pipelines automatizados diseñados para realizar manipulaciones masivas de imágenes en un corto periodo de tiempo. Esto podría utilizarse para inundar redes sociales y medios digitales con contenido manipulado, como imágenes alteradas de documentos, pruebas visuales o comunicados falsos. La capacidad de generar imágenes de forma rápida y convincente podría facilitar campañas de desinformación a gran escala, especialmente en contextos de inestabilidad política o económica.

Además, en un entorno judicial, este tipo de tecnologías representa un riesgo significativo. La generación de pruebas visuales falsificadas, como imágenes de contratos, correos electrónicos impresos o mensajes dentro de capturas de pantalla, podría alterar el curso de investigaciones legales o desacreditar a una de las partes implicadas. La dificultad para diferenciar entre imágenes auténticas y generadas por técnicas de difusión complica aún más su detección.

VI-B. Privacidad y Ética en los Datasets

El entrenamiento de estos sistemas depende en gran medida de datasets masivos que contienen imágenes y textos. En el caso de aplicaciones más específicas, como documentos de identidad o formularios oficiales, la obtención de datasets plantea un dilema ético y legal. Crear un dataset de estas características, incluso con fines académicos, sería ilegal en la mayoría de lugares, ya que implicaría recopilar información altamente sensible y protegida.

Aun así, dentro del ecosistema del cibercrimen, acceder a este tipo de datos no resulta particularmente complejo. Mercados ilegales como los presentes en la Dark Web ofrecen conjuntos de datos con documentos robados o filtrados, lo que podría facilitar el entrenamiento de modelos destinados a actividades ilícitas como la falsificación documental. Este riesgo subraya cómo las capacidades de estos modelos podrían ser aprovechadas para fines ilegales.

VI-C. Riesgos Futuros

A medida que los modelos de difusión avanzan, con marcos de contexto más amplios como Stable Diffusion 3.5 [8] y FLUX.1 [9] que soportan resoluciones de 1024x1024 o superiores, una controlabilidad por prompt más precisa y capacidades de inpainting mejoradas, el riesgo asociado a su mal uso se incrementa. Estas mejoras, aunque benefician aplicaciones legítimas, también pueden facilitar el fraude a gran escala, erosionando la confianza en mecanismos de autenticación visual y fomentando actividades ilegales a través de la generación de deepfakes cada vez más realistas.

VII. Conclusiones

Esta investigación ha explorado las capacidades y limitaciones de TextDiffuser-2 en el contexto de la manipulación y generación de texto en imágenes. Los experimentos realizados han demostrado una alta efectividad en la generación de texto visualmente coherente, especialmente cuando existe un contexto visual adecuado, aunque con variaciones significativas según la complejidad del texto y el entorno. El modelo ha mostrado una notable capacidad de adaptación a diferentes estilos y longitudes de texto, si bien presenta dificultades en la preservación de detalles sutiles y efectos de difuminado.

Las implicaciones de esta tecnología son profundas y de doble filo. Por un lado, ofrece posibilidades prometedoras para aplicaciones legítimas en edición y corrección de imágenes; por otro, plantea serias preocupaciones éticas sobre su potencial uso malicioso, especialmente en la manipulación de documentos sensibles. Los resultados subrayan la necesidad de un enfoque equilibrado en el desarrollo futuro de estas tecnologías, considerando tanto su potencial beneficioso como los riesgos asociados.

La investigación futura deberá centrarse no solo en mejorar las capacidades técnicas del modelo, sino también en desarrollar mecanismos de seguridad robustos y establecer pautas éticas claras para su implementación, garantizando un desarrollo responsable de esta tecnología.

VIII. Trabajo Futuro

Las líneas de investigación futuras incluyen:

- Mejora en la preservación de detalles sutiles
- Desarrollo de técnicas más robustas para el manejo de contexto
- Implementación de mecanismos de seguridad adicionales
- Integración de modelos multimodales para un análisis más profundo de la coherencia visual
- Desarrollo de métricas de evaluación más precisas para medir la fidelidad del texto generado
- Exploración de técnicas de aprendizaje supervisado para mejorar la adaptación de estilos

- Implementación de enfoques de aprendizaje continuo para mantener la relevancia del modelo con datos actuales
- Investigación de estrategias de mitigación para prevenir el uso malintencionado de la tecnología

REFERENCIAS

[1] Cui, L. (2023). TextDiffuser-2: Unleashing the power of language models for text rendering. arXiv preprint arXiv:2311.16465.

[2] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). arXiv preprint arXiv:2112.10752.

[3] Liu, R., Garrette, D., Saharia, C., Chan, W., Roberts, A., Narang, S., Blok, I., Mical, R., Norouzi, M. Constant, N. (2023). Character-Aware Models Improve Visual Text Rendering. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16270–16297, Toronto, Canada. Association for Computational Linguistics. DOI: 10.18653/v1/2023.acl-long.900.

[4] Ma, J., Zhao, Y. (2023). GlyphDraw: Seamlessly Rendering Text with Intricate Spatial Structures in Text-to-Image Generation. arXiv preprint arXiv:2303.17870.

[5] Radford, A., Kim, K., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning, 139, 8748-8763.

[6] Nah, S., Balaji, Y., others. (2022). eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. arXiv preprint arXiv:2211.01324.

[7] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Seyed Ghasemipour, S. K., Gontijo-Lopes, R., Karagol Ayan, B., Salimans, T., Ho, J., Fleet, D. J. Norouzi, M. (2022). Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. NeurIPS 2022. arXiv preprint arXiv:2205.11487.

[8] Stability AI. (2024). Introducing Stable Diffusion 3.5. Retrieved from https://stability.ai/news/introducing-stable-diffusion-3-5

[9] Black Forest Labs. (2024). FLUX.1-dev. Retrieved from https://huggingface.co/black-forest-labs/FLUX.1-dev