Image Edition Using Diffusion Models and Natural Language

María Graciel Cruz Cáceres
Departamento de Ciencia de la Computación
Universidad Católica San Pablo
Arequipa - Perú
Email: maria.cruz@ucsp.edu.pe

José Eduardo Ochoa Luna
Departamento de Ciencia de la Computación
Universidad Católica San Pablo
Arequipa - Perú
Email: jeochoa@ucsp.edu.pe

***Abstract*—La edición semántica basada en modelos de difusión es una técnica emergente para editar imágenes utilizando descripciones de texto, pero presenta importantes limitaciones como la calidad variable de las descripciones de texto y la necesidad de hardware especializado. En este trabajo se evalúan técnicas de edición estructural y semántica basadas en texto con modelos de difusión. Se identifican las limitaciones y se comparan las técnicas más relevantes con las métricas de *CLIP Score* y *Direction Similarity* para evaluar la calidad visual y coherencia de las imágenes. Se presentan la comparación entre dos técnicas del estado del arte qué presentan enfoques diferentes, *Plug and Play* e *Hiper*. Por un lado *Plug and Play* utiliza una guía de imagen y un mensaje de texto para inyectar gradualmente sus características en el modelo, mientras que *Hiper* utiliza la descomposición del texto de entrada para optimizar los resultados. Estos métodos tienen como objetivo mejorar la accesibilidad y utilidad de la técnica de edición semántica basada en texto con modelos de difusión en una variedad de aplicaciones prácticas en el futuro.**

***Index Terms*—Inteligencia artificial generativa, modelos generativos, modelos de dispersión, síntesis de imágenes, edición semántica de imágenes**

1. Introducción

La Síntesis de imágenes, que es el proceso de generar nuevas imágenes o manipular las existentes, es un campo que ha sido beneficiado recientemente gracias a los avances generados por el aprendizaje profundo [1]. Esta es una tarea importante ya que posee varias aplicaciones prácticas como diseño por computador asistido, generación de arte, edición de imágenes, realidad virtual, entre otros.

Actualmente, contamos con modelos generativos capaces de generar imágenes guiados con la interacción con el usuario a través de entradas limitadas no solo a imágenes, sino también incluyendo textos, sketches, trazos, grafos y disposiciones que son más intuitivos [2]. Además, hemos logrado avances enfoques más potentes y versátiles que no solo nos permiten generar imágenes, sino también editarlas.

Por supuesto aún existen varios desafíos con respecto a ello como obtener el significado sobre regiones semánticas, la necesidad de entrenar en conjuntos de datos muy grandes [3], preservar el contenido de la imagen original sin introducir cambios cambios en regiones no deseadas [4],inclusive la dificultad de los usuarios para encontrar un texto que describa con precisión cada detalle visual de la imagen [2], entre otros.

El presente trabajo se enfoca principalmente en los métodos de síntesis de imágenes basados en modelos de difusión para la edición de imágenes.

1.1. Objetivos

El objetivo del presente trabajo es hacer un levantamiento del estado del arte en la edición de imágenes a partir del lenguaje natural utilizando modelos de difusión identificando también las limitaciones presentes en las técnicas más relevantes. Además se pretende realizar una comparativa entre dos de ellas donde se analizarán y compararán los resultados obtenidos por ambas técnicas en diferentes métricas de evaluación, como la similitud entre las representaciones visual-textual y la similitud de la dirección de cambio imagen-texto entre la imagen original y la edición.

1.2. Definición del problema

La edición semántica basada en texto con modelos de difusión es una técnica que permite editar imágenes utilizando descripciones de texto. A pesar de que esta técnica ha avanzado significativamente en los últimos años, todavía presenta algunas limitaciones como:

Limitaciones de lenguaje natural ya que la calidad de las descripciones de texto utilizadas para editar imágenes puede variar significativamente, lo que puede afectar la calidad de la imagen generada; la necesidad de una gran cantidad de datos, y limitaciones de hardware ya que requieren hardware especializado para ejecutarse de manera efectiva y varias horas de entrenamiento; además de la dificultad para conservar las características originales de

una imágenes al realizar la edición.

El resto del trabajo se divide de la siguiente manera, en la sección 2 se encuentran los trabajos relacionados, posteriormente en nla sección 3 la descripción de dos de las técnicas mas recientes que no requieren afinamiento del modelo para cada imagen *Plug and Play* e *Hiper*. Finalmente en la sección 4 e muestran los experimentos realizados en ambas ttécnicas y finalmente en la sección 5 las conclusiones.

2. Trabajos relacionados

Todo comenzo con la generación de textos, a partir de allí la generación de imágenes surgió. El primer enfoque que se presento es la generación a partir de otras imágenes lo que permite la realización de algunas tareas como la super-resolution, desenfoque, traducción e inclusive la edición de imágenes. El problema con este enfoque es la falta de flexibilidad, en cambio, el control guiado por texto permite la generación de cualquier contenido de imagen con cualquier estilo a voluntad libre de los humanos.

2.1. Generación de imágenes a partir de texto

Gracias a la incorporación conocimientos de modelos de lenguaje grandes como (LLM) o modelos híbridos de visión y lenguaje los usuarios pueden generar imágenes realistas de alta resolución utilizando solo una descripción de texto que describe la escena deseada.

Los enfoques modernos de aprendizaje automático para la síntesis de texto a imagen comenzaron con el trabajo de Mansimov et al. (2015) llamado *AlignDRAW* [5]. Extendieron el modelo DRAW de Gregor et al. (2015) [6] para alinear las descripciones de texto con las regiones relevantes de la imagen, de esa genera imágenes que se ajustan a esas descripciones.

El trabajo de Zhang et al. [1] realiza un levantamiento del arte sobre los modelos de difusión en la generación de imágenes a partir de texto. En la figura 1 muestra una cronología sobre los trabajos más relevantes de generación de imágenes a partir de texto. Comienza por el trabajo *AlignDRAW* de Mansimov et al. (2015) [5], continua por los métodos basados en *GAN*, metodos de autregresion y finalmente los metodos basados en difusion como son Stable Diffusion [7] yDALL-E 2 [8].

2.2. Edición de imágenes a partir de texto

En la actualidad, existen numerosos trabajos del estado del arte en el campo de la edición de imágenes basados en textos. Estos utilizan técnicas avanzadas de inteligencia artificial para generar imágenes realistas y detalladas. Por un lado tenemos a la edición utilizando entradas auxiliares, siendo la más común el uso de máscaras. Por otro lado las que solamente utilizan un texto para indicar la edición deseada

Edición de imágenes con el uso de entradas auxiliares. Las entradas más comunes que encontramos para la síntesis de imágenes además del texto pueden clasificarse en entradas que son similares a imágenes y las que no. En las entradas similares a imágenes podemos encontrar el uso de máscaras, bocetos o dibujos lineales, mapas semánticos de etiqueas y poses representadas como puntos clave del cuerpo. Y con respecto a las entradas que no son similares a imágenes encontramos a etiquetas de clase, vectores de atributos y entradas similares a grafos y disposiciones. Un trabajo más exhaustivo sobre el uso de diferentes entradas para la síntesis de imagen es el trabajo de Xue et al. [2].

Un ejemplo sobre la edición con el uso de entradas auxiliares, en este caso de máscaras lo encontramos en el trabajo de Avrahami et al. (2022) [9] que combina modelos de imagen de lenguaje como CLIP y un modelo probabilístico de difusión de eliminación de ruido, *denoising diffusion probabilistic model* (DDPM), para generación de imágenes realistas. Algunos usos son añadir un nuevo objeto, remover/reemplazar/alterar los objetos existentes, edición del fondo y extrapolación de imágenes. A pesar de otorgar buenos resultados, el uso de una máscara trae algunas desventajas junto con ella como por ejemplo, dificulta la edición intuitiva y eliminar información estructural importante para el proceso de edición de la imágen.

Enfoque de aprendizaje *zero-shot*. La inteligencia artificial generativa usualmente requiere una gran cantidad de recursos, por un lado el acceso a una gran cantidad de datos y el acceso a recursos computacionales lo que puede llegar a ser muy desafiante en especial para tareas de síntesis de imágenes a texto [3]. Por ejemplo, se utilizaron 250M imágenes para entrenar a DALL-E [**?**] o 650 millones para entrenar DALL-E 2. En la contrucción de modelos de lenguaje de texto, ChatGPT3 [**?**] utilizo el conjunto de datos CommonCrawl que tiene 45TB de texto planos comprimidos. En el uso de recursos computacionales el uso de CPU no cumplen con las necesidades de entrenar grandes modelos de *deep learning*, por ejemplo el primer modelo entrenado con ImageNet, AlexNet [**?**] fue entrenado con *Graphics Processing Units (GPUs)*. Hoy en día se utilizan GPUs siendo la empresa que lidera el mercado NVIDIA, el uso de CUDA y uso de Tensor Processing Units (TPUs).

El enfoque de *zero-shot learning* también se encuentra presente en la edición de imágenes. Un ejemplo de este enfoque se presenta en el trabajo *"Paint by word"* de Andonian et al. [10] donde introducen el problema de Manipulación semántica de imágenes *zero-shot* donde utiliza redes GANs para realizar la edición a partir de 2 entradas: en primer lugar el texto y en segundo lugar la selección de una región de la imagen.

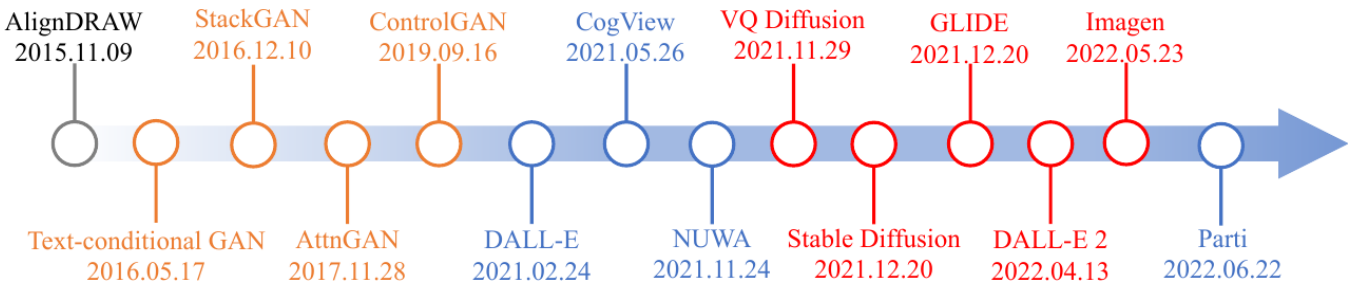

Figure 1. Trabajos representativos de generación de texto a imágenes [1]. Los métodos basados en GAN, autorregresivos y de difusión se representan con los colores amarillo, azul y rojo, respectivamente.

Edición de imágenes basados únicamente en lenguaje natural. Tenemos también los trabajos que no requieren de una entrada adicional. Estos se guian de la imágen de entrada, esta puede ser una imágen generada por medio de inteligencia artificial o una imagen del mundo real Los modelos que mayormente se utilizan son enfoques basados en *GANS* y enfoques basados en modelos de difusión.

Los enfoques de edición basados en GANs están limitadas a un dominio restringido y para su manipulación necesita realizar una inversión al espacio latente, se ha demostrado que existe un compromiso entre la precisión y la editabilidad en las imágenes invertidas. [11]. Por otro lado Los métodos que utilizan modelos de difusión son capaces de realizar síntesis de imágenes de alta calidad.

2.3. Enfoques basados en modelos de difusión

La edición de imágenes solía ser dominada por métodos de inversión de GAN combinados con CLIP. Sin embargo, estas técnicas a menudo presentaban limitaciones en su capacidad de inversión.

Actualmente los modelos de difusión han generado un interés creciente en la comunidad de modelos generativos desde la publicación de DDPM (Denoising Diffusion Probabilistic Model) en 2020 [12]. Esto se debe a su capacidad para mejorar la calidad y diversidad de la edición de imágenes.

Algunas de las consideraciones que podemos observar en los trabajos más recientes sobre la edición de imágenes a partir de texto son las siguientes: En primer lugar, en que medida se desea modificar la imágen, pueden realizar ediciones intentando conservar la mayor cantidad de información estructural como conservar el acercamiento, el ángulo de la cámara como en el trabajo de *Imagic* [13] o intentando conservar la información semántica, es decir conservar la mayor cantidad de características del sujeto como en la técnica Hiper [14]. Una segunda consideración sería si se reentrena o afina el modelo en cada iteración para cada conjunto diferente de entradas; se ofrece poder computacional a cambio de una mayor precisión u obtener imágenes parecidas en mayor medida a la original. Y finalmente sobre el enfoque de instrucción utiliza: Si utiliza una pequeña frase conteniendo la edición deseada (instrucción simple), si recibe solo una característica que desea cambiar (Característica simple) y si se requiere una instrucción que describa el conjunto de entrada además de la indicación de la edición (instrucción a instrucción).

En la Tabla 1 se presentan los trabajos recientes más relevantes en orden cronológico indicando que enfoque utilizan para la técnica que proponen. (i) Si se centran en conservar la información semántica o la información estructural, (ii) si se tiene que reentrenar el modelo en cada ocasión que se presenten entradas de un diferente contexto y finalmente (iii) si es modelado instrucción a intrucción o si la técnica interpreta información sobre el conjunto de entrada.

DiffusionCLIP. En primer lugar *DiffusionCLIP* [15] es un trabajo pionero que introduce el Modelado Difusivo (DM, por sus siglas en inglés) para solucionar este problema. Este adopta un DM pre-entrenado para convertir la imagen de entrada al espacio latente y luego ajusta finamente el DM en el proceso inverso con una función de pérdida compuesta por dos términos: la pérdida CLIP direccional local [21] y una pérdida de identidad. La primera se utiliza para guiar la imagen objetivo a alinearse con el texto y la segunda mitiga cambios no deseados. Para permitir una inversión completa, adopta un modelo de inversión determinista DDIM [22] en lugar del proceso inverso DDPM [12].

Prompt to prompt. El segundo trabajo es el trabajo de investigación de de Hertz et al. (2022) *"Promp-to-Prompt"* [16] que propone un marco de edición intuitivo de indicación a indicación. este trabajo con el uso de capas cruzadas de atención controla la relación entre el diseño espacial de la imagen y cada palabra en la indicación. De esta manera permite la edición localizada reemplazando una palabra, la edición global agregando una especificación e incluso controlar el grado en que una palabra se refleja en la imagen. La idea clave de la propuesta es utilizar *Cross-attention maps* (Mapas de atención cruzada) controlando qué píxeles atienden a qué tokens del texto de la consigna durante cuáles pasos de difusión.

Técnica	Enfoque	Información que conserva	Reentrenamiento y ajuste	Enfoque de Instrucción
Diffusion Clip [15]	Convierte la imagen al espacio latente con el modelo preentrenado. A continuación guiado por la pérdida CLIP ajusta el camino inverso del modelo. Se basa en DDMI y en caso sea un dominio no visto antes utiliza DDPM.	Semántica y estructural	Afinación del modelo	Instrucción a Instrucción
Prompt-to-Prompt [16]	A partir de imagenes generadas por un modelo de difusión, inyecta mapas de atención cruzada para realizar las ediciones.	Estructural	No requiere	Instrucción a Instrucción
Imagic [13]	Afina el modelo preentrenado de difusión a la imagen de entrada, luego interpola la representación de la imagen con la indicación objetivo creando asi la combinación de ambas indicaciones.	Semántica	Afinación del modelo	Instrucción simple
Text2live [4]	Utiliza un generador entrenado con los datos de entrada aprovechando un modelo CLIP pre-entrenado que se mantiene fijo. Genera una capa de edición que se sobrepone encima de la imagen original.	Estructural	Reentrenamiento	Característica simple
Plug and Play [17]	Al invertir la imagen de entrada, en el proceso de eliminación de ruido se extraen las características espaciales que son inyectadas en la generación de la nueva imagen.	Estructural	No requiere	Instrucción simple
Sine [18]	Realizan el ajuste fino del modelo basado en parches. Tratan la imagen de entrenamiento única como una función en coordenadas y la divide en subáreas.	Semántica y estructural	Afinación del modelo	Característica simple
Pix2pix Zero [19]	Edita una imagen en una dirección de edición, descubierta en el espacio incrustado del texto con la guía de atención cruzada.	Estructural	No requiere	Característica simple
HiPer [14]	Optimiza la incrustación de textos del espacio incrustado CLIP para representar al sujeto. En la inferencia aplica el incrustamiento personalizado y el texto objetivo personalizado.	Semántica	No requiere	Instrucción a Instrucción
MDP [20]	Propone un framework que permite manipular la trayectoria de difusión y realizar ediciones de imágenes guiadas por texto mediante diferentes tipos de manipulaciones.	Estructural	No requiere	Instrucción a Instrucción

TABLE 1. TÉCNICAS BASADAS EN MODELOS DE DIFSUIÓN PARA LA EDICIÓN DE IMÁGENES

Imagic. Siguiendo la línea de investigación podemos encontrar el trabajo de Kawar et al. *"Imagic"* [13] que nos presenta un método que permite la capacidad de realizar ediciones semánticas complejas basadas en texto utilizando una única imagen real preservando sus características originales. "Imagic" encontró que los diferentes métodos tenian varias desventajas: se encontraban limitados a un conjunto específico de ediciones como dibujar sobre la imagen, añadir un objeto o la transferencia de estilos; trabajaban sobre imágenes de un dominio específico o imágenes generadas sintéticamente o que requerían entradas adicionales auxiliares para indicar la edición deseada. Para mitigar estas limitaciones hace uso de un modelo pre-entrenado de difusión de texto a imagen para trabajar directamente con la imagen en vez de generar una nueva,produce un *embedding* de texto de manera que produce imágenes similares a la imagen de entrada, afina el modelo de difusión para capturar y reconstruir la apariencia de la imagen para finalmente realizar una interpolación entre el texto embebido y el optimizado lo que resulta en una combinación de la imagen de entrada y la edición deseada.

Text2Live. A continuación se encuentra el trabajo de Bartal et al. (2022) *"Text2LIVE"* [4] que presenta un método para la manipulación de la apariencia impulsada por texto de imágenes y videos naturales sin necesidad de entrenar con ejemplos previos utilizando el enfoque *zero-shot*. El objetivo del artículo es editar la imagen a partir del texto de entrada de una manera semánticamente significativa. *"Text2LIVE"* entrena un generador utilizando un conjunto de datos interno de ejemplos extraídos de una sola entrada (imagen o video y prompt de texto objetivo), mientras aprovechamos un modelo CLIP pre-entrenado para establecer las pérdidas. Generar una capa de edición (color + opacidad) que se compone sobre la entrada original para restringir el proceso de generación y mantener alta fidelidad a la entrada original.

SINE. Siguiendo con el enfoque de zero-shot tenemos el trabajo de Zahng et al. (2022) *"SINE"* [18]. Este trabajo aborda el problema de edición de imágenes reales en los casos donde para un token dado solo hay disponible una única imagen de referencia, un ejemplo de cuando estas situaciones ocurren sería la edición de una pintura o un cuadro. Intenta resolver lo problemas de desviación del lenguaje y el sobreajuste en contenido y geometría del modelo preentrenado a una sola imagen proponiendo nueva una guía basada en el modelo construida sobre la guía libre de clasificadores

Pix2pix Zero. También se tiene el trabajo de Parmar et al. [19] que propone el método llamado *"Pix2pix-zero"* que de forma similar al anterior trabajo trata de preservar el contenido original de la imagen mientras realiza una traducción de imagen a imagen con la guía de *cross-attention*

para mantener los mapas de *cross-attention* de la imagen original.

3. Técnicas a implementar

3.1. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation

Plug-and-Play presenta un marco de trabajo para la traducción de imagen a imagen. Recibe como entrada la imagen guía I^G y la indicación de texto P; como salida otorga I^* que es una nueva imagen que conserva la estructura original de la imagen guía. Utiliza la manipulación de características espaciales y autoatención dentro del modelo.

En la Figura 2 se ilustra el funcionamiento del método propuesto. Dada una imagen guía I^G se invierte al ruido inicial x_T^G. Progresivamente se elimina el ruido de la inversión mediante muestreo *DDIM*. En el proceso de eliminación de ruido se extraen las características espaciales (f_t^l, q_t^l, k_t^l); donde f_t^l son las características intermedias, q_t^l son las indicaciones y k_t^l son las llaves. Para generar la traducción de imagen asigna a x_T^* el valor de x_T^G y se inyectan las características guías (f_t^l, q_t^l, k_t^l) en ciertas capas.

Información preliminar. Este método se basa en el modelo de difusión *Latent Diffusion Model - LDM* conocido como *Stable diffusion* [7]. El proceso de difusión se aplica al espacio latente de un *autoencoder* de imágenes previamente entrenado. El modelo esta basado en una arquitectura *U-net* [23] la cual es la estructura base para la eliminación de ruido.

Una arquitectura *U-net* consta de 3 bloques: un bloque residual, un bloque de auto atención y un bloque de atención cruzada. En el bloque residual, se combinan las características de las imágenes ϕ_t^{l-1} de la capa anterior $l-1$ a través de la convolución lo que produce las características inmediatas f_t^l. En el bloque de atención proyecta las características en consultas q_t^l, llaves k_t^l y valores v_t^l. la salida de este bloque esta dado por $\hat{f}_t^l$ descrito en la ecuación 1

$$\hat{f}_t^l = A_t^l v_t^l, donde = Softmax(q_t^l {k_t^l}^T) \tag{1}$$

Finalmente en el ultimo bloque se calcula la atención cruzada entre las características espaciales de la imagen y la incrustación de tokens de la indicación de texto P.

3.1.1. Método de Plug-and-Play. A partir de dos observaciones inyectan las características directamente sin reentrenar el modelo por cada imagen. La primera observación es: las características espaciales extraídas de las capas decodificadoras intermedias codifican información semántica localizada, estas capas se ven afectadas en menor medida. La segunda es: la auto-atención representa las similitudes entre las características espaciales, permiten retener detalles finos de diseño y forma.

La información espacial de una imagen esta codificada en ϵ_θ, realizaron un análisis *PCA* sobre un conjunto de imágenes y observaron que en las capas intermedias se encuentra información semántica localizada con variaciones de apariencias entre imagen e imagen.

Inyección de características. El proceso de inyección de características se da de la siguiente manera:

Sea x_T^G el ruido inicial obtenido de invertir I^G utilizando *DDIM* [7] y sea la indicación objetivo P para la generación de la imagen I^*. Utiliza el mismo ruido de x_T^G (el ruido inicial) en x_T^*. En cada paso t del proceso regresivo extrae las características guia f_t^l del paso de eliminación de ruido: $z_{t-1}^G = \epsilon_\theta(x_t^G, \varnothing, t)$. Estas características son inyectadas en la generación de I^*, (en el paso de reducción de rúido x_t^*) y sobreescribe las características resultantes f_t^{*l} en f_t^l, esta operación se expresa como:

$$z_{t-1}^* = \hat{(\epsilon)}_\theta(x_t^*, P, t; f_t^l) \tag{2}$$

Para conservar las características espaciales no las modifica en niveles profundos.

Auto atención. Los módulos de auto atención calculan la similitud A_t^l entre las características espaciales después de proyectarlas linealmente en consultas y llaves. La atención está alineada en las primeras capas con el diseño semántico de la imagen, se agrupa así regiones de acuerdo a partes semánticas. La inyección de la matriz de auto atención se realiza remplazando la matriz A_t^l en la ecuación 2

$$z_{t-1}^* = \hat{(\epsilon)}_\theta(x_t, P, t; f_t^4, A_t^l) \tag{3}$$

La capa máxima de inyección A_t^l controla el nivel de fidelidad a la estructura original, al mismo tiempo que mitiga el problema de fuga de apariencia.

Todo se resume en el Alg.1. El algoritmo es controlado por los parámetros τ_f que define el paso de muestreo t hasta que f_t^4 sea inyectado y τ_A el paso de muestreo hasta que que A_t^l sea inyectado.

3.2. Highly Personalized Text Embedding for Image Manipulation by Stable Diffusion (Hiper)

Hiper descompone el espacio de incustración CLIP, de esa manera permite mantener la identidad de las imágenes al personalizarlas y manipularlas. El objetivo del método propuesto es optimizar la incrustación de textos en el espacio incrustado clip para representar al sujeto. Utiliza las partes finales del espacio incrustado ya que en esa parte encuentra la identidad del sujeto. El método a comparación de trabajos anteriores no necesita afinar el modelo para crear ediciones semánticas complejas. Las ediciones que permite son altamente personalizadas que incluyen manipulaciones como el fondo, textura y movimiento utilizando una sola imagen y el texto objetivo como entrada.

En la Figura 3 podemos observar a grandes rasgos el funcionamiento. Esta compuesto por dos partes, primero el entrenamiento donde se descompone tanto el texto como la

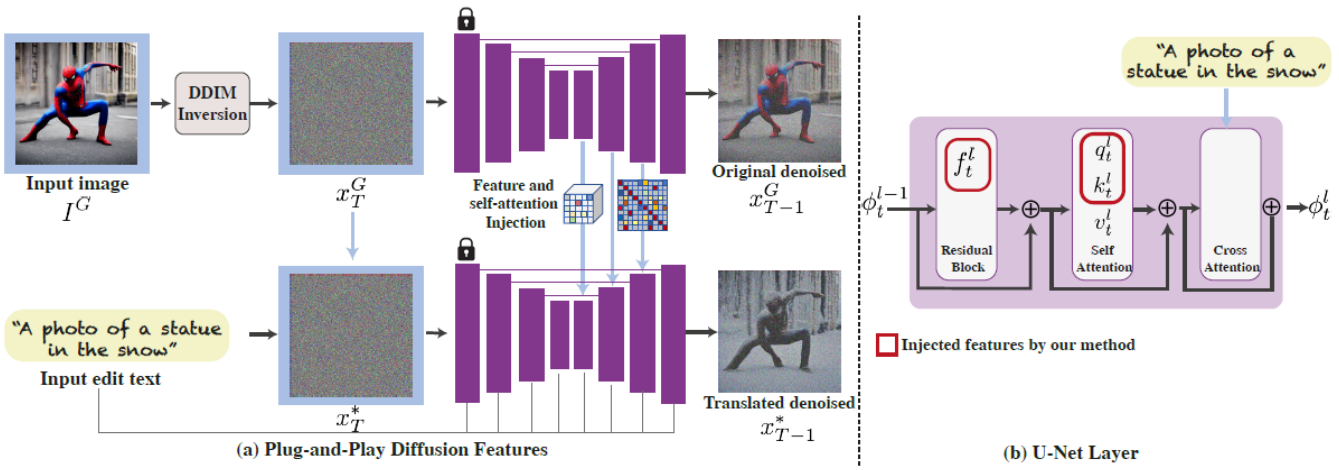

Figure 2. Características de difusión de Plug-and-Play. (a) El marco de trabajo toma como entrada una imagen de guía y una frase de texto que describe la traducción deseada. La imagen de guía se invierte para obtener un ruido inicial, que luego se va depurando progresivamente utilizando muestreo *DDIM*. Durante este proceso, extrae las características espaciales de las capas decodificadoras y su auto-atención, como se ilustra en (b). Para generar nuestra imagen traducida guiada por texto inyecta las características de guía (f_t^l, q_t^l, k_t^l) en ciertas capas

Algorithm 1 Plug-and-Play Difussion Features

Input:
I^G ▷ Imagen real guía
P ▷ Indicación objetiva
τ_f, τ_A ▷ Thresholds de injección

$x_T^G \leftarrow$ DDIM-inv(I^G)
$x_T^* \leftarrow x_T^G$ ▷ Comienza de la misma semilla
for $t = T...1$ **do**
  $Z_{t-1}^G, f_t^4, A_t^l = \epsilon_\theta(x_t^G, \varnothing, t)$
  $x_{t-1}^G = DDIM - samp(x_t^G, z_{t-1}^G)$
  if $t > \tau_f$ **then** $f_t^{*4} = f_t^4$
  else $f_t^{*4} = \varnothing$
  end if
  if $t > \tau_A$ **then** $A_l^{*t} = A_t^l$
  else $A_t^{*l} = \varnothing$
  end if
  $z_{t-1}^* = \hat{\epsilon}_\theta(x_T^*, P, t; f_t^{*4}, A_t^{*l})$
  $x_{t-1}^* = DDIMsamp(x_t^*, z_{t-1}^*)$
end for
Output: $I^* = x_0^*$

imagen de entrada, se obtiene el incrustamiento personalizado e_{hper} y después la inferencia donde aplica el texto objetivo optimizado e'_{tgt} junto con el incrustamiento personalizado e_{hyper} en el modelo de difusión ya pre-entrenado para obtener la imagen final.

3.2.1. Entrenamiento. En primer lugar tenemos una una indicación de texto y que nos sirve para denominar a la imagen. Calculamos la incrustación de texto que denominaremos e_{src}, e_{src} se encuentra en el espacio $\mathbb{R}^{CM}$ que se obtiene al utilizar el tokenizador propio de CLIP [24] τ_ϕ en nuestra indicación y, $e_{src} = \tau_\phi(y)$.

A continuación selecciona los últimos N tokens al final del texto a lo cual denomina incrustación altamente personalizable *(Hiper embedding)*, $e_{hyper} \in \mathbb{R}^{CN}$, los $M - N$ tokens restantes son denotados como $e'_{sem} \in \mathbb{R}^{C(M-N)}$. La descomposición del texto puede expresarse como:

$$e_{src} = [e'_{src}, e_{hyper}] \tag{4}$$

La parte que no contiene información relevantes es eliminada, e'_{src} aun contiene información similar a e_{src}. A continuación, para una imagen dada p_0 y su espacio latente $x_0 = E(p_0)$ el *(Hiper embedding)* e_{hyper} es optimizado donde el e'_{src} generado del texto original se mantiene:

$$e_{hyper} = \arg\min_{e_h \in \mathbb{R}^{C \times M}} \mathbb{E}_{x_t, \epsilon \sim \mathcal{N}(O,I)}[||\epsilon - \epsilon_\theta(x_t, t, [e'_{src}, e_h])||^2] \tag{5}$$

3.2.2. Inferencia. En el paso de inferencia e'_{src} se remplaza por el texto incrustado recortado e'_{tgt} de la indicación de texto objetivo, la cola de N tokens del final son reemplazados por el *(Hiper embedding)* e_{hyper}.

Posteriormente, la difusión inversa genera muestras latentes mediante la siguiente ecuación:

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t, e_{cmp})) + \sigma_t \epsilon \tag{6}$$

donde el espacio compuesto e_{cmp} esta dado por:

$$e_{cmp} = [e'_{tgt}, e_{hyper}] \tag{7}$$

La imagen final se obtiene al aplicar el decodificador a la muestra del espacio final.

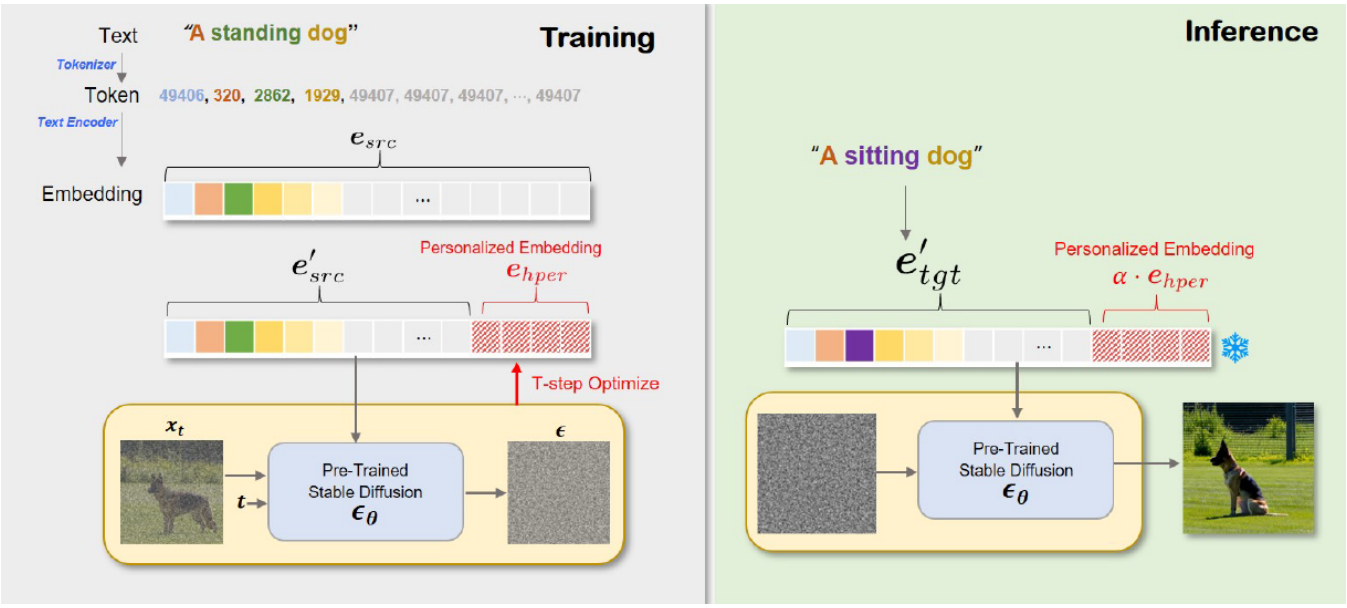

Figure 3. El método propuesto. (Entrenamiento) Primero, se convierte la frase de texto de origen, que tiene el significado de la imagen, en incrustación de texto. Se eliminan algunas partes de la incrustación de texto que no tienen información. La parte informativa de la incrustación de destino y la incrustación personalizada se concatenan y se utilizan como entrada de la red neuronal pre-entrenada *U-net*. Durante el entrenamiento, solo se optimiza la incrustación personalizada. Aunque esta Figura lo representa como un aprendizaje en el espacio de imagen, en realidad la incrustación se optimiza en el espacio latente. (Inferencia) La incrustación de destino también se recorta y se concatena con la incrustación personalizada. El vector de incrustación personalizada se calibra multiplicándolo por = 0:8. El modelo pre-entrenado de texto a imagen, que está condicionado por esa incrustación, genera una imagen que tiene el significado del texto de destino y el sujeto de la imagen de origen.

Cuando se condiciona la incrustación compuesta e_{cmp} al modelo pre-entrenado de difusión estable, la imagen de salida tiene el significado del texto objetivo y la personalización del sujeto de la imagen de origen. Un enfoque simple pero con impresionantes resultados de rendimiento.

4. Experimentos y resultados

4.1. *Benchmark*

Para la evaluación de ambas técnicas se considero el *benchmark TEdBench (Textual Editing Benchmark)* propuesto por Kawar et al. [13]. Fue utilizado originalmente por la técnica de *Imagic*, y es utilizado también por la técnica *Hiper* [14] para la realización de los experimentos.

TEdBench contiene 100 pares de entradas de texto-imagen. Por un lado contiene un total de 100 indicaciones de texto que describen una edición simple que se aplicará a una imagen específica. Estas ediciones pueden ser cambio de posturas y edición de múltiples objetos. Por otro lado las imágenes son fotografías reales, lo que quiere decir que no son generadas por otra inteligencia artificial. *TEDbecnh* contiene en total 50 fotografías de las cuales 21 corresponden a animales y 29 a objetos. Las imágenes se encuentran a color y su tamaño es de $1024 \times 1024px$.

4.2. Métricas

Se utilizaron dos métricas para evaluar el desempeño de ambas técnicas, *CLIP Score* y Similaridad de Dirección Texto-Imagen basado en CLIP.

CLIP Score es una métrica que se utiliza para evaluar la correlación entre una leyenda generada para una imagen y el contenido real de la imagen. Es una medida de similitud entre las representaciones visual y textual de un par imagen-texto.

Primero se transforma la imagen y el texto en un espacio vectorial común utilizando el modelo CLIP. El modelo CLIP ha sido entrenado en un gran conjunto de datos de imágenes y texto. *CLIP Score* se calcula tomando la similitud del coseno entre las transformaciones conocidas como incrustación de la imagen y del texto. La similitud del coseno es una medida de similitud entre dos vectores y se calcula tomando el producto punto de los dos vectores y dividiéndolo por el producto de sus longitudes. El puntaje está limitado entre 0 y 1, siendo 1 el mejor puntaje indicando que existe una gran correlación entre la imagen y el texto dado.

CLIP score se encuentra descrito en la siguiente ecuación:

$$CLIPScore(I, C) = max(cos(E_I, E_C), 0) \qquad (8)$$

Donde:

- E_I es la incrustación de una imagen I
- E_C es la incrustación de un texto C

La Similaridad de Dirección Texto-Imagen de CLIP evalúa la calidad de los modelos de edición de imágenes. Mide la similitud entre la dirección del cambio entre dos leyendas imágenes. Se calcula tomando la similitud del coseno entre las incrustaciones de la imagen original y la imagen resultante después de la edición junto con las dos leyendas. El puntaje está limitado entre 0 y 1, siendo 1 el mejor puntaje. Un puntaje cercano a 1 indica que el modelo de edición de imágenes preserva el contenido semántico de la imágen.

Para el cálculo de la Similaridad de dirección de texto-imagen denotada por $L_{\text{direction}}$ se necesita calcular la diferencia del espacio entre las incrustaciones de imágenes denotada con ΔI y la diferencia del espacio entre las incrustaciones de las leyendas denotada con ΔT. La fórmula de similairdad quedaría de la siguiente forma:

$$L_{direction} = 1 - \frac{\Delta I \cdot \Delta T}{|\Delta I| \cdot |\Delta T|} \tag{9}$$

4.3. Experimentos

Se ejecutaron ambas técnicas sobre el *benchmark* seleccionado. Por cada imagen resultante de la edición se calcularon las métricas de *Clip Score* y *Clip text similarity*. Finalmente se extrajo el promedio de los puntajes obtenidos por todas las imágenes en cada técnica. Los resultados obtenidos se presentan expuestos en la Tabla 2.

Técnica	*CLIP Score*	*CLIP Similarity*
Plug and Play [17]	0.2245	0.2430
Hiper [14]	0.2173	0.1547
Imagic [13]	0.2148	0.1021

TABLE 2. RESULTADOS DE LOS PUNTAJES OBTENIDOS AL EDITAR IMÁGENES CON LAS TÉCNICAS PLUG AND PLAY E HIPER UTILIZANDO LAS MÉTRICAS DE CLIP SCORE Y CLIP TEXT SIMILARITY. ADICIONALMENTE SE PRESENTA JUNTO CON LAS MÉTRICAS OBTENIDAS CON LA TÉCNICA IMAGIC

Se obtuvo qué el puntaje de *CLIP Score* y *Clip Direction similarity* de la técnica de *Plug and Play* es mayor con respecto a la técnica de *Hiper*. Con respecto a la técnica de *Plug and Play* obtuvo un puntaje de *Clip Score* de 0.2245 , este puntaje es menor al indicado al reportado en el trabajo de [17] 0.2820, y en cuanto a *Hiper* obtenemos un puntaje de 0.2173, el cual es ligeramente mayor al reportado en el trabajo de [14] 0.2047. A pesar de aparentar ser puntajes relativamente bajos, según el trabajo de *Plug and Play*, el estado del arte en el uso de modelos de difusión para la edición de imágenes se encuentra en el rango de 0.20 a 0.2820 cuyo puntaje corresponde a la misma técnica de *Plug and Play*.

5. Conclusiones

Con base en lo expuesto anteriormente, se puede llegar a la conclusión de que existe un interés considerable en el campo de la edición de imágenes a partir de texto. Existe una gran cantidad de trabajos en los últimos dos años y continuamente se publican estudios relacionados con el tema.

Se encuentra qué actualmente los resultados de las métricas seleccionadas es bajo. Esto podría deberse a dos motivos. En primer lugar es posible que no se hayan encontrado las métricas apropiadas para la evaluación de estos modelos y en segundo lugar que existe un gran espacio de mejora.

Se encuentra que la técnica de *Plug and Play* obtuvo mejores resultados en cuanto a ambas métricas. La diferencia entre ambas técnicas es de 0.01 con respecto al CLIP Score y 0.09 con respecto a la métrica de Clip Similarity. La diferencia entre los puntajes de ambas técnicas es muy baja por lo que no se puede afirmar que una técnica es mejor que la otra. Además se encuentra que la técnica de *Plug and Play* necesita menor cantidad de recursos a comparación de *Hiper*, esto puede deberse al paso previo en la optimización de la incrustación antes de realizar la edición.

La diferencia de ambas técnicas radica en que la optimización del trabajo de *Hiper* se centra en la incrustación de texto. Por otro lado *Plug and Play* se centra en el proceso de edición

Entre las limitaciones qué se encontraron al realizar este trabajo la principal fue la limitación de recursos. Originalmente los experimentos realizados en la propuesta de la técnica *Hiper* [14] fueron ejecutados en la tarjeta gráfica *RTX 3090* y la cantidad de memoria *CUDA* es superior a la ofrecida por *Colab*.

Se identificaron algunos espacios de mejora los cuáles incluyen la preservación de la estructura general de las imágenes, la composición espacial. De igual forma es complicado conservar y modificar los puntos de vista en las imágenes y la iluminación.

Como posible trabajo futuro, se plantea abordar la dificultad de especificar con lenguaje natural ediciones espaciales. Esto implica desarrollar técnicas que permitan realizar modificaciones como cambiar la posición de un objeto específico, alterar la disposición entre objetos o colocar un objeto en una posición precisa, entre otras posibilidades.

References

[1] C. Zhang, C. Zhang, M. Zhang, and I. S. Kweon, "Text-to-image diffusion model in generative ai: A survey," *arXiv preprint arXiv:2303.07909*, 2023.

[2] Y. Xue, Y.-C. Guo, H. Zhang, T. Xu, S.-H. Zhang, and X. Huang, "Deep image synthesis from intuitive user input: A review and perspectives," *Computational Visual Media*, vol. 8, pp. 3–31, 2022.

[3] C. Zhang, C. Zhang, S. Zheng, Y. Qiao, C. Li, M. Zhang, S. K. Dam, C. M. Thwal, Y. L. Tun, L. L. Huy, D. kim, S.-H. Bae, L.-H. Lee, Y. Yang, H. T. Shen, I. S. Kweon, and C. S. Hong, "A complete survey on generative ai (aigc): Is chatgpt from gpt-4 to gpt-5 all+ you need?" 2023.

[4] O. Bar-Tal, D. Ofri-Amar, R. Fridman, Y. Kasten, and T. Dekel, "Text2live: Text-driven layered image and video editing," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*. Springer, 2022, pp. 707–723.

[5] E. Mansimov, E. Parisotto, J. L. Ba, and R. Salakhutdinov, "Generating images from captions with attention," 2016.

[6] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra, "Draw: A recurrent neural network for image generation," in *International conference on machine learning*. PMLR, 2015, pp. 1462–1471.

[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[8] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[9] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 208–18 218.

[10] A. Andonian, S. Osmany, A. Cui, Y. Park, A. Jahanian, A. Torralba, and D. Bau, "Paint by word," *arXiv e-prints*, pp. arXiv–2103, 2021.

[11] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or, "Designing an encoder for stylegan image manipulation," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–14, 2021.

[12] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[13] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, "Imagic: Text-based real image editing with diffusion models," in *Conference on Computer Vision and Pattern Recognition 2023*, 2023.

[14] I. Han, S. Yang, T. Kwon, and J. C. Ye, "Highly personalized text embedding for image manipulation by stable diffusion," *arXiv preprint arXiv:2303.08767*, 2023.

[15] G. Kim, T. Kwon, and J. C. Ye, "Diffusionclip: Text-guided diffusion models for robust image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2426–2435.

[16] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," 2022.

[17] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel, "Plug-and-play diffusion features for text-driven image-to-image translation," *arXiv preprint arXiv:2211.12572*, 2022.

[18] Z. Zhang, L. Han, A. Ghosh, D. Metaxas, and J. Ren, "Sine: Single image editing with text-to-image diffusion models," 2022.

[19] G. Parmar, K. K. Singh, R. Zhang, Y. Li, J. Lu, and J.-Y. Zhu, "Zero-shot image-to-image translation," 2023.

[20] Q. Wang, B. Zhang, M. Birsak, and P. Wonka, "Mdp: A generalized framework for text-guided image editing by manipulating the diffusion path," *arXiv preprint arXiv:2303.16765*, 2023.

[21] R. Gal, O. Patashnik, H. Maron, A. H. Bermano, G. Chechik, and D. Cohen-Or, "Stylegan-nada: Clip-guided domain adaptation of image generators," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–13, 2022.

[22] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.