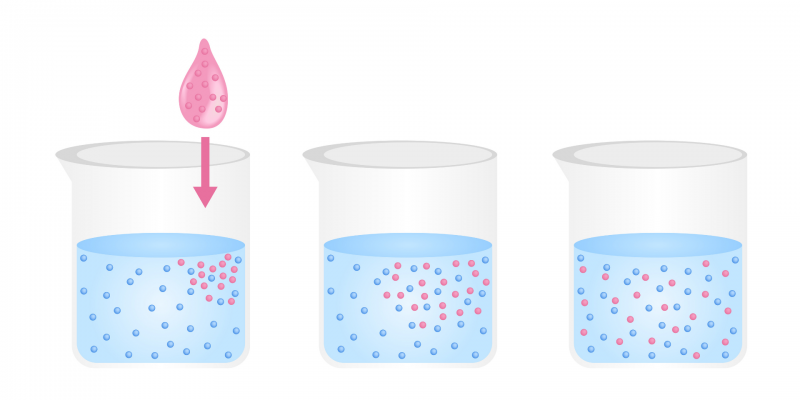
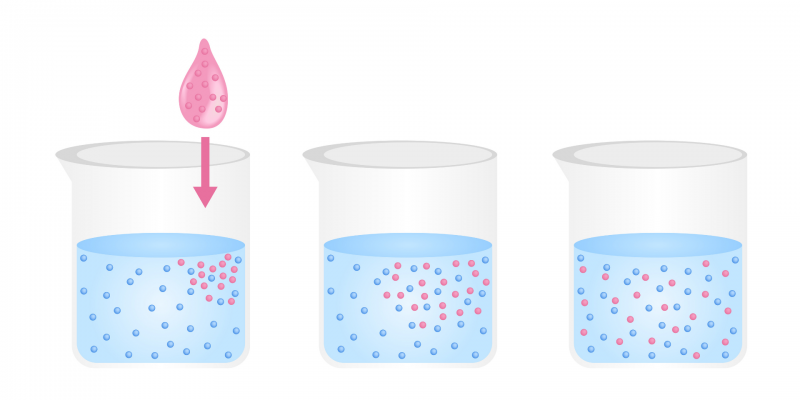
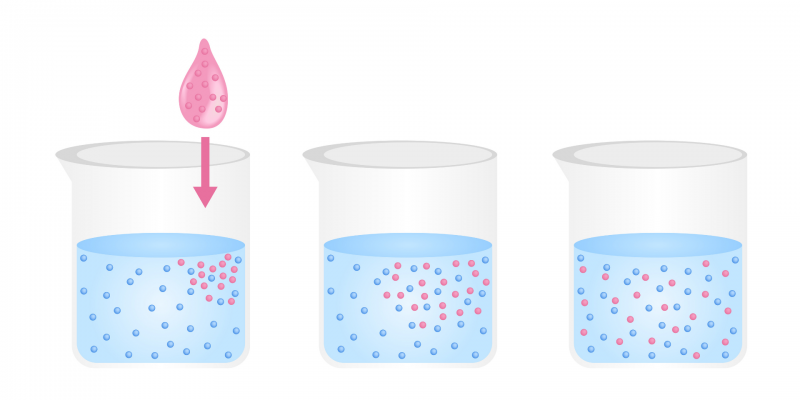
**MODELOS DE DIFUSIÓN**

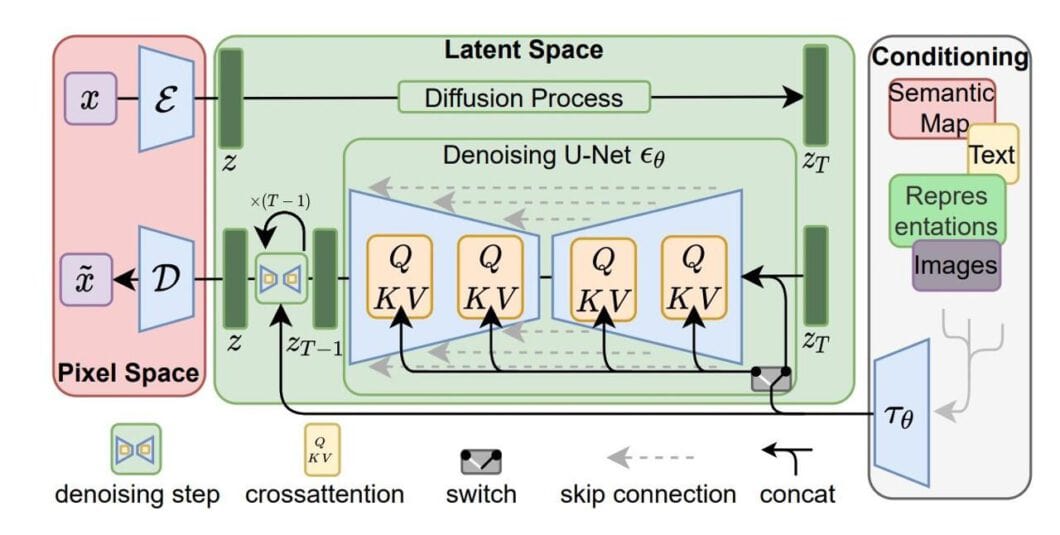
1. **Difusión:** La difusión es un proceso físico mediante el cual partículas, energía o información se dispersan de regiones de alta concentración a regiones de baja concentración hasta alcanzar un estado de equilibrio.

En términos matemáticos, suele modelarse con la ecuación de difusión (derivada de la segunda ley de Fick), que describe cómo evoluciona la distribución de una sustancia en el espacio y el tiempo.

En el contexto de modelos generativos, esta idea física se traslada a las imágenes: agregar ruido progresivamente a una imagen es equivalente a difundirla hasta perder su estructura original. Luego, entrenamos un modelo para invertir ese proceso, reconstruyendo la imagen a partir del ruido.



1. **Arquitectura de Stable Diffusion:** Stable Diffusion es un modelo generativo basado en difusión latente, diseñado para generar imágenes de alta calidad a partir de descripciones textuales. Su arquitectura se compone principalmente de tres bloques:



* **Autoencoder Variacional (VAE):**

Reduce las imágenes del **espacio de píxeles** al **espacio latente**, comprimiendo su información y permitiendo trabajar con representaciones más compactas y eficientes.

* **Modelo de Difusión Latente:**

Aprende a **eliminar progresivamente el ruido** dentro del espacio latente en lugar del espacio de píxeles. Esto hace el proceso más rápido y menos costoso computacionalmente, sin perder calidad.

* **Codificador de Texto (Text Encoder):**

Permite el **condicionamiento textual**, es decir, guía el proceso generativo a partir de **descripciones en lenguaje natural**, asegurando que la imagen generada corresponda al prompt del usuario.

1. **Modelo de Difusión:** El **modelo de difusión** es el núcleo de Stable Diffusion. Su funcionamiento se basa en dos procesos complementarios:
2. **Difusión hacia adelante (forward diffusion):** añadir ruido.
3. **Difusión inversa (reverse diffusion):** eliminar ruido.

En esencia, el modelo transforma progresivamente una imagen (o su representación latente 𝑧) en ruido puro y luego aprende a invertir este proceso para reconstruir la imagen original.

* 1. **Difusión (Forward Process):**
     + **Objetivo:** degradar gradualmente la imagen hasta convertirla en ruido gaussiano.
     + **Cómo se hace:** en cada paso t, se añade ruido siguiendo un **calendario de ruido** (*noise schedule*) que controla la intensidad en cada iteración.
     + **Resultado:** después de muchos pasos, la información original se vuelve irreconocible.

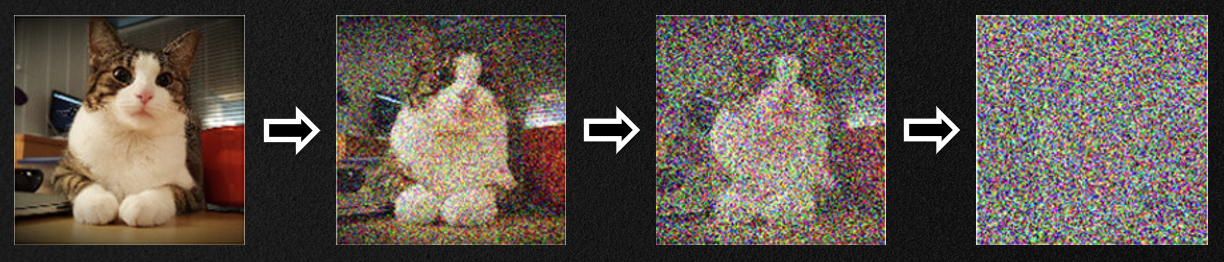
**Representación matemática:**





Donde:

* + - zt​: estado de la imagen en el paso t.
    - βt ​: parámetro del **noise schedule** que controla cuánta varianza de ruido se añade en cada paso.



* 1. **Difusión inversa(denoising):** La **difusión inversa** (reverse diffusion process) es la parte crucial de los modelos de difusión. Mientras que la difusión directa degrada la imagen añadiendo ruido, **la difusión inversa reconstruye una muestra coherente paso a paso a partir de ruido puro**.

**Proceso General**

* **Punto de partida:** ruido gaussiano puro zT​ en el último paso T.
* **Meta:** reconstruir progresivamente la imagen latente limpia z0.
* **Cómo se logra:** una red neuronal (en Stable Diffusion, una **UNet**) predice el ruido presente en cada paso y lo elimina iterativamente.

**Rol de la UNet**

En cada paso t, la UNet toma como entrada:

* El **latente con ruido** zt
* El **paso temporal** t
* (Opcional) **Condicionamiento textual** (el *prompt*).

Y produce como salida una predicción del **ruido** presente en esa representación:

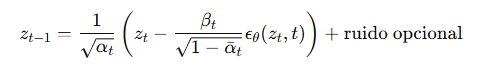
****

donde:

* c representa el condicionamiento textual.
* θ son los parámetros aprendidos por la UNet.

**Eliminación Progresiva del Ruido**

Con la predicción ϵθ(zt,t) se estima la versión menos ruidosa del latente:



Donde αt​ y αˉt son coeficientes derivados del calendario de ruido (*noise schedule*).

Así, **paso a paso**, el modelo:

1. Predice el ruido presente.
2. Lo sustrae del latente actual.
3. Obtiene un latente menos ruidoso para el siguiente paso.

Al final, z0​ es un **latente limpio** que luego el decodificador VAE transforma de nuevo al **espacio de píxeles** para producir la imagen final.

**Importancia del Proceso de Denoising**

* Es lo que **realmente hace posible generar imágenes desde ruido**.
* Permite controlar el estilo, calidad y fidelidad al prompt.
* Su diseño eficiente es lo que hace que **Stable Diffusion** pueda generar imágenes de alta calidad en relativamente pocos pasos (por ejemplo, 20–50 iteraciones).



* 1. **Papel de Noise Schedule**

El noise Schedule define cómo varía la intensidad del ruido βt​ en cada paso:

* Pasos iniciales: poco ruido.
* Pasos finales: ruido más fuerte.

Esto crea un degradado controlado que facilita el aprendizaje inverso.

1. **Denoising UNet:** La **UNet** es la **arquitectura central** de Stable Diffusion responsable de **predecir el ruido** presente en las representaciones latentes. Su tarea es recibir el latente ruidoso y devolver el **residuo de ruido estimado**, permitiendo así el proceso de denoising paso a paso.

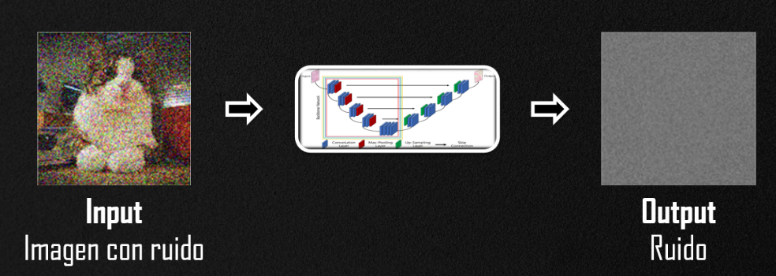
**Características Principales**

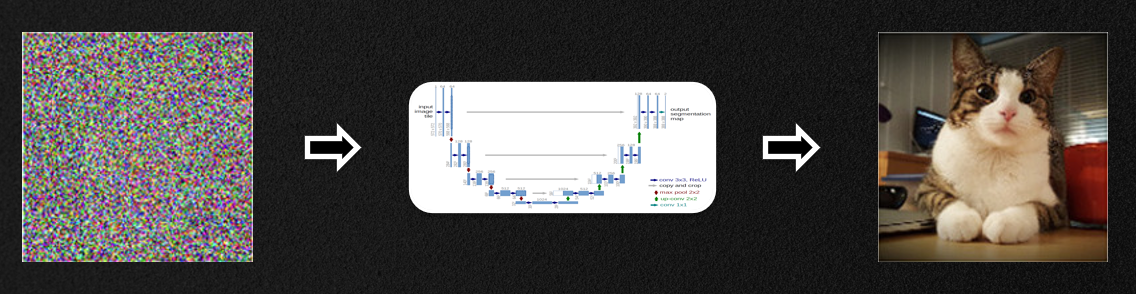
* **Estructura Encoder–Decoder con “Skip Connections”:** La UNet está formada por dos partes:
  + **Encoder (contracción):** extrae características y reduce progresivamente la resolución para capturar el contexto global.
  + **Decoder (expansión):** reconstruye la información con resolución creciente para recuperar detalles finos.
  + **Skip connections:** conectan capas del encoder con capas correspondientes del decoder, evitando la pérdida de información y preservando detalles locales.
* **Captura de Contexto Global y Detalles Locales:** Gracias a su diseño jerárquico, la UNet puede **comprender estructuras generales** (formas grandes) y **detalles finos** (texturas, bordes).
* **Predicción del Ruido:** La red se entrena para **minimizar la diferencia** entre el **ruido real** y el **ruido predicho** en cada paso de la difusión inversa:



donde:

* ϵ: ruido real aplicado.
* ϵθ: ruido predicho por la UNet.
* c: condicionamiento textual.





1. **UNet:** La **UNet** fue originalmente desarrollada para **segmentación de imágenes biomédicas** (Ronneberger et al., 2015).  
   En **Stable Diffusion**, esta arquitectura se adapta como **red central** para **predecir y eliminar el ruido** del latente ruidoso en cada paso del proceso de difusión inversa.

Se trata de una **arquitectura encoder–decoder con conexiones “skip”**, diseñada para trabajar a múltiples escalas y aquí extendida para **denoising y condicionamiento textual**.

### **5.1. Estructura Principal**

* **Encoder (contracción):** Serie de bloques convolucionales que reducen progresivamente la resolución, capturando **características globales** y contexto.
* **Bottleneck (capa intermedia):** Combina y procesa las características extraídas en un **espacio latente más compacto**, facilitando la integración de contexto y atención cruzada.
* **Decoder (expansión):** Reconstruye gradualmente la resolución original mediante **convoluciones traspuestas** (o upsampling), recuperando los detalles perdidos.
* **Skip Connections:** Conectan las capas equivalentes del encoder y del decoder, preservando **detalles finos** y mejorando el flujo de gradientes durante el entrenamiento.

### **5.2. Adaptación para Stable Diffusion**

* **Entrada:** el **latente ruidoso** zt​ junto con el **timestep** t.
* **Condicionamiento textual:** se añaden **bloques de cross-attention** para inyectar la información proveniente del **Text Encoder** en múltiples niveles de la red.
* **Salida:** en lugar de reconstruir la imagen completa, cada bloque predice directamente el **ruido ϵθ** presente en el latente.
* **Uso del resultado:** la predicción de ruido se utiliza para calcular el siguiente zt−1 en la difusión inversa.

### **5.3. Ventajas de la UNet en Stable Diffusion**

* **Escala multi-resolución:** captura simultáneamente **contexto global** (estructura de la imagen) y **detalle local** (texturas finas).
* **Flexibilidad para condicionamiento:** la **cross-attention** permite inyectar **embeddings textuales** en cualquier capa, mejorando el control sobre el estilo y el contenido generado.
* **Estabilidad y calidad:** bien entrenada, la UNet produce **eliminaciones de ruido precisas y consistentes** con el prompt textual, generando imágenes de alta calidad.



1. **Latent Space:** En Stable Diffusion, el modelo no trabaja directamente en el espacio de píxeles, sino en un espacio latente comprimido. Esto se logra gracias al autoencoder (VAE) que transforma la imagen 𝑥 en un vector latente 𝑧:



donde 𝐸 es el encoder y 𝐷 el decoder.

Características del Latent Space:

* **Dimensionalidad reducida:** los latentes suelen tener 1/4 o 1/8 del tamaño espacial original.
* **Eficiencia computacional:** menos parámetros y operaciones para la UNet.
* **Preservación de estructura:** aunque comprimido, conserva las características esenciales de la imagen (forma, composición, colores globales).
* **Proceso de Difusión en Latentes:**
* En lugar de aplicar ruido en píxeles (𝑥), se aplica sobre 𝑧.

El forward process (añadir ruido) y reverse process (denoising) se realizan enteramente en 𝑧.

Al final, el decoder del VAE 𝐷 reconstruye la imagen en píxeles.

1. **Condicionamiento Textual:** El condicionamiento textual es el mecanismo por el cual se controla la generación de imágenes con lenguaje natural.

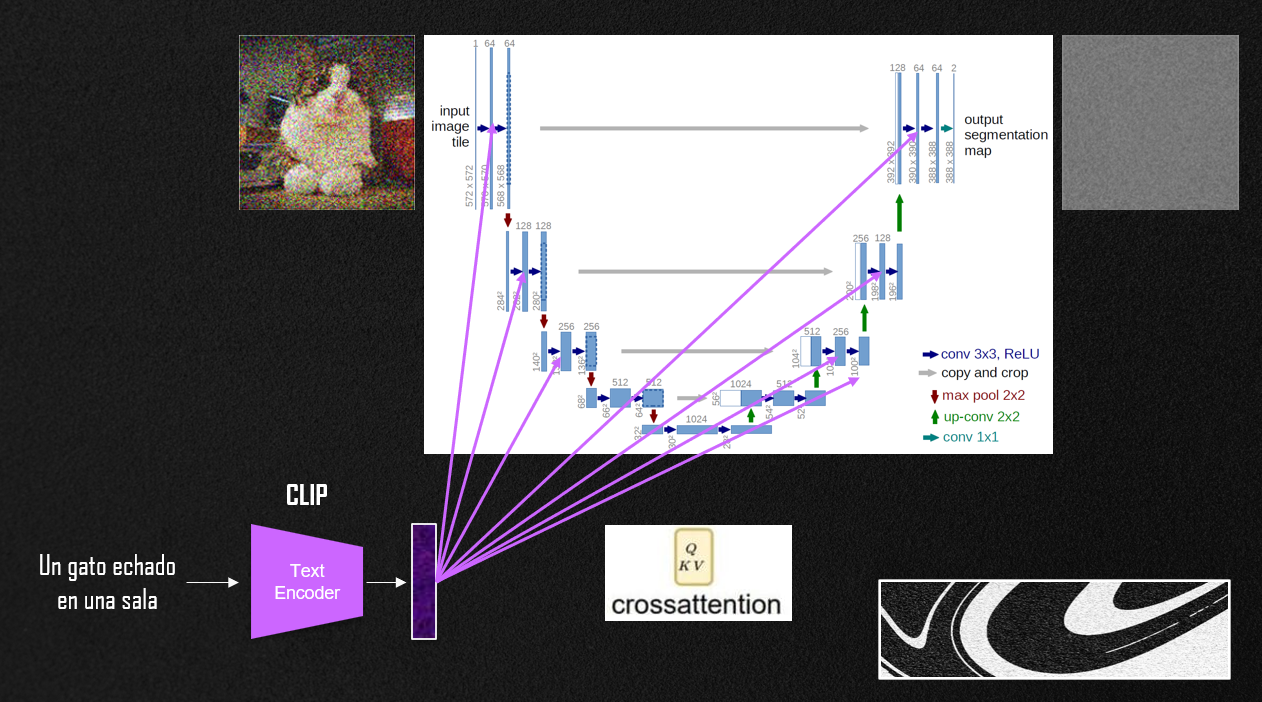
Se utiliza un Text Encoder (CLIP en el caso de Stable Diffusion).

Convierte las frases en embeddings numéricos.

Estos embeddings se inyectan en la UNet para guiar la eliminación de ruido hacia una imagen coherente con la descripción.

1. **Denoising UNet(Con condicionamiento textual):** Aquí la UNet no solo recibe la imagen latente ruidosa, sino también las representaciones del texto.

En Stable Diffusion, la UNet no solo recibe el latente ruidoso, sino también la información del texto para guiar la eliminación del ruido. Esto permite que el modelo genere imágenes coherentes con la descripción textual.



**Cómo funciona:**

**Entrada:** latente ruidoso 𝑧𝑡, timestep 𝑡 y embeddings del texto.

**Cross-Attention:** cada bloque de la UNet tiene capas de cross-attention que integran los embeddings del texto con las características visuales.

**Predicción del ruido:** la UNet estima 𝜖^𝜃(𝑧𝑡,𝑡 text), es decir, el ruido en cada paso condicionado por el texto.



1. **Secuencia de como entra el texto al Text Encoder(CLIP y Cross-Attention):**

El prompt se tokeniza (se convierte en palabras o sub-palabras).

El Text Encoder (CLIP) transforma los tokens en embeddings.

Estos embeddings pasan a capas de cross-attention dentro de la UNet.

Durante la denoising, la UNet consulta el embedding textual para guiar la reconstrucción.

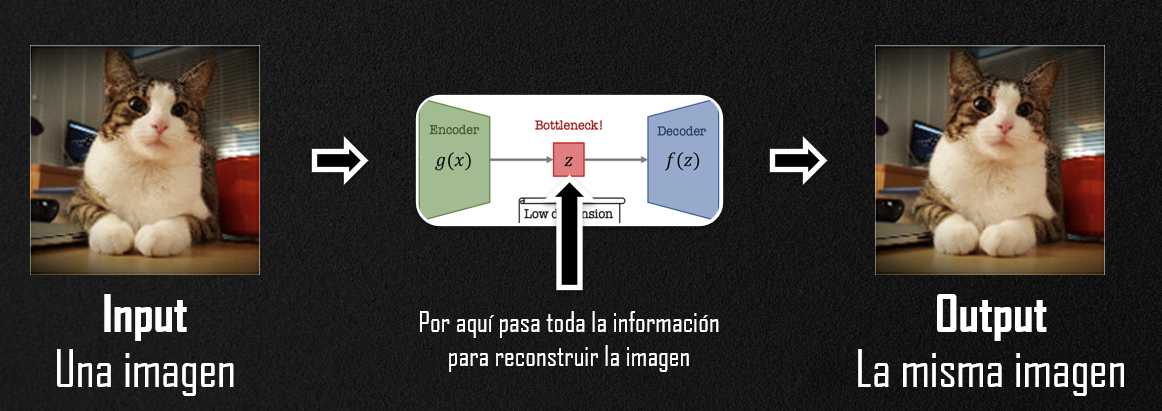
1. **Autoencoder:** El autoencoder usado en Stable Diffusion es un Variational Autoencoder (VAE) especializado:

**Encoder:** comprime la imagen al espacio latente (Z).

**Bottleneck:** espacio intermedio de baja dimensionalidad.

**Decoder:** reconstruye la imagen desde el espacio latente.

Esto reduce la dimensionalidad y permite que la UNet trabaje con menos datos sin perder calidad visual.



1. **Conexión completa del proceso:**

Texto → Embeddings (CLIP).

Imagen Latente + Ruido → UNet.

UNet elimina ruido paso a paso (condicionado por texto).

Decoder reconstruye imagen final desde espacio latente.

1. **Conclusiones:**

Los modelos de difusión se basan en principios físicos de ruido y denoising.

Stable Diffusion es eficiente gracias al espacio latente y al autoencoder.

La UNet con cross-attention permite al modelo relacionar texto e imagen con precisión.

**CREACIÓN DE UN MODELO DE DIFUSIÓN CONDICIONAL PARA GENERACIÓN DE DÍGITOS MNIST**

De Ruido Aleatorio a Arte Digital Controlado

**Resumen**

Este proyecto representa un viaje fascinante desde los fundamentos teóricos de los modelos de difusión hasta la implementación práctica de un sistema de generación de imágenes completamente controlable. Partiendo del concepto revolucionario de que "todo orden emerge del caos", desarrollamos un modelo capaz de transformar ruido gaussiano puro en dígitos manuscritos específicos mediante comandos textuales precisos.

**Innovación y Metodología**

El Arte de la Difusión Inversa

Mi enfoque se fundamenta en la elegante dualidad de los procesos de difusión: mientras que la naturaleza tiende al desorden (entropía), mi modelo neuronal aprende el camino inverso, guiando sistemáticamente el caos hacia estructuras reconocibles y significativas. Este paradigma representa una de las aproximaciones más sofisticadas en la generación sintética de contenido visual.

**Arquitectura**

El corazón del sistema es una arquitectura U-Net mejorada que incorpora:

**Time Embedding Posicional:** Inspirado en los transformers, permite al modelo comprender su posición exacta en el proceso temporal de 1000 pasos

**Skip Connections Inteligentes:** Preservan información crítica durante la compresión y reconstrucción

**Class Conditioning Dinámico:** Integra semánticamente el significado textual con la geometría visual

**Classifier-Free Guidance:** Técnica de vanguardia que permite control granular sobre la fidelidad al prompt

Transfer Learning Estratégico

Una característica destacada es la implementación de aprendizaje **progresivo:** comencé con un modelo incondicional robusto (base sólida) y evolucionamos hacia capacidades condicionales mediante transfer learning inteligente. Esta metodología no solo acelera la convergencia, sino que garantiza estabilidad durante todo el proceso.

**Logros Técnicos Destacados**

**1. Cronograma de Ruido Optimizado**

Tras experimentación empírica, refiné el cronograma de difusión, descubriendo que para el dominio MNIST, un enfoque lineal superaba las aproximaciones coseno más complejas. Esta decisión basada en evidencia demuestra que la simplicidad elegante a menudo triunfa sobre la complejidad teórica.

**2. Visualización del Proceso Creativo**

Implementé un sistema único de visualización temporal que revela el "nacimiento" de cada dígito a través de 5 snapshots críticos del proceso de denoising. Esto transforma la "caja negra" de la generación en un espectáculo visual comprensible, mostrando cómo el orden emerge gradualmente del caos.

**3. Control Semántico Preciso**

El modelo final no solo genera dígitos aleatorios, sino que responde a comandos específicos como "genera un siete" o "crea un cero", con capacidad de ajustar la fuerza del condicionamiento mediante el parámetro de Classifier-Free Guidance.

**Profundidad Conceptual**

La Filosofía del Denoising

Mi trabajo toca conceptos profundos sobre la naturaleza de la información y el reconocimiento de patrones. El modelo no simplemente "limpia" ruido; aprende representaciones latentes de lo que constituye un dígito válido en cada nivel de degradación. Es, en esencia, un filósofo digital que comprende la "digitalidad" en sus múltiples manifestaciones ruidosas.

**Validación y Resultados**

Métricas de Convergencia

El modelo demuestra convergencia estable con pérdidas decrecientes consistentes a lo largo de 15 épocas de entrenamiento condicional. La diferenciación entre predicciones condicionales e incondicionales confirma que el sistema ha internalizado exitosamente el concepto de condicionamiento semántico.

**Calidad Visual**

Las imágenes generadas exhiben:

Coherencia estructural: Dígitos bien formados y reconocibles

Variabilidad controlada: Diferentes estilos manuscritos para el mismo dígito

Fidelidad al prompt: Alta correspondencia entre comando textual y resultado visual

Robustez del Sistema

Pruebas con diferentes niveles de guidance demuestran la flexibilidad del modelo para balancear creatividad y precisión según los requerimientos específicos de la aplicación.

**Impacto e Implicaciones**

Contribución Académica

Este trabajo constituye una implementación educativa completa de técnicas de difusión estado-del-arte, proporcionando una base sólida para futuras investigaciones en generación condicional y control semántico.

**Escalabilidad**

La arquitectura modular permite extensión natural hacia:

Datasets más complejos (Fashion-MNIST, CIFAR-10)

Condicionamiento textual más rico

Generación multi-modal

Filosofía del Código Abierto

Cada componente está exhaustivamente documentado, permitiendo reproducibilidad total y facilitando la comprensión de conceptos avanzados de machine learning generativo.

**Conclusión: Arte y Ciencia Convergentes**

Este proyecto trasciende la mera implementación técnica para convertirse en una exploración de cómo las máquinas pueden aprender a crear. Hemos construido no solo un generador de dígitos, sino un sistema que entiende, interpreta y materializa intenciones semánticas.

En el fondo, hemos enseñado a una máquina el arte sublime de transformar el caos en orden, el ruido en belleza, y la incertidumbre en precisión controlada. Es, en su esencia más pura, la democratización de la creatividad artificial.

"En cada píxel generado reside la prueba de que la inteligencia artificial no solo computa, sino que también puede soñar con dígitos específicos."

**Repositorio:**

* <https://github.com/DgteMendez777/DiffusionModel.git>
* [git@github.com:DgteMendez777/DiffusionModel.git](mailto:git@github.com:DgteMendez777/DiffusionModel.git)

**Referencias bibliográficas:**

* https://www.iic.uam.es/innovacion/generar-imagenes-a-partir-de-textos-con-modelos-difusion/
* https://es.khanacademy.org/science/biology/membranes-and-transport/diffusion-and-osmosis/v/diffusion-video
* https://www.youtube.com/watch?v=RiQ4YcAm0tA
* https://www.youtube.com/watch?v=1Vi4woheMyE
* https://www.youtube.com/watch?v=-uH0ffkTbas
* https://www.vegaitglobal.com/media-center/knowledge-base/what-is-stable-diffusion-and-how-does-it-work
* https://proceedings.neurips.cc/paper\_files/paper/2019/file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf
* https://codoraven.com/blog/ai/stable-diffusion-clearly-explained/
* https://www.youtube.com/watch?v=NhdzGfB1q74