Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Evaluación de rendimiento de humanos y modelos generativos en tareas de generación de imágenes a partir de texto

Tesis de Licenciatura en Ciencias de la Computación

Paula Pérez Bianchi

Director: Diego Fernandez Slezak
Codirector: Pablo Riera
Buenos Aires, 2024

EVALUACIÓN DE RENDIMIENTO DE HUMANOS Y MODELOS GENERATIVOS EN TAREAS DE GENERACIÓN DE IMÁGENES A PARTIR DE TEXTO

Frente a la evolución de los modelos generativos, persisten preguntas sobre sus capacidades en comparación con las de los humanos. En esta tesis se propone una tarea innovadora para abordar este problema, evaluar modelos de texto a imagen mediante un experimento de neurociencia inspirado en el juego del "teléfono descompuesto". Este experimento, realizado con humanos, se planteó en el contexto de las ciencias cognitivas con el objetivo de identificar los elementos descriptivos que influyen en la comunicación del conocimiento. Dado que los modelos generativos pueden estar involucrados en múltiples fases dentro del experimento, inicialmente se concentró el análisis en la fase que incluye la generación de imágenes a partir de descripciones textuales. Esto permitió comparar directamente el rendimiento de los modelos de texto a imagen con la capacidad humana. Se exploraron dos modelos Stable Diffusion v1.5 y DALL·E 3 y varias técnicas de *alignment* para generar imágenes de composiciones geométricas utilizando las descripciones humanas recolectadas durante el experimento anterior. Finalmente se realizo una evaluación *crowdsourced* de las imágenes generadas, encontrándose que los humanos superan tanto a DALL·E 3 como a Stable Diffusion v1.5 en esta tarea.

Palabras claves: IA Generativa, Human-machine interaction, Evaluación de modelos, fine-tuning, Stable Diffusion.

EVALUACIÓN DE RENDIMIENTO DE HUMANOS Y MODELOS GENERATIVOS EN TAREAS DE GENERACIÓN DE IMÁGENES A PARTIR DE TEXTO

In light of the evolving capabilities of generative models, questions persist regarding their performance compared to humans. In this thesis, an innovative task is proposed to address this issue, evaluating text-to-image models through a neuroscience experiment inspired by the game "Chinese whispers". This experiment, conducted with humans, was designed within the context of cognitive sciences to identify the descriptive elements that influence knowledge communication. Since generative models can participate in multiple phases of the experiment, the initial analysis focused on the generation of images from textual descriptions. This allowed for a direct comparison between the performance of text-to-image models and human capabilities. Two main models, Stable Diffusion v1.5 and DALL·E 3, were explored, along with various alignment techniques to generate images of geometric compositions using human-generated textual descriptions collected during a prior experiment. Finally, a crowdsourced evaluation of the generated images was conducted, revealing that humans outperformed both DALL·E 3 and Stable Diffusion v1.5 in this task.

Keywords: Generative AI, Human-machine interaction, Model Evaluation, fine-tunning, Stable Diffusion.

AGRADECIMIENTOS

Nada de esto habría sido posible sin el apoyo incondicional de mis padres. Gracias, mamá, por confiar siempre en mí y apoyarme, incluso en mis ideas más locas. Gracias, papá, por tener siempre las palabras justas para motivarme. A ambos, gracias por confiar en mi criterio y por escucharme hablar interminablemente sobre las cosas locas que aprendo en la facultad. Prometo siempre estar ahí para ayudarlos a instalar la impresora.

Mis primeros pasos en Exactas fueron extraños. Era plena pandemia y nadie sabía muy bien qué hacer. Todo el contacto con otros estudiantes era a través de grupos online. Así, entre seudónimos y chats de Discord, fui conociendo a diversas personas (muchos, integrantes de La Plebe) que hicieron este proceso mucho más llevadero y que hoy son mis amigos. Gracias a mis compañeros de TP por las risas y la buena onda, incluso en los momentos de mayor tensión. Espero que hayan disfrutado de mis chistes, aunque fueran producto de estar totalmente quemada. Mención especial a Lucy, Vale, Sinno, Simón y Echu, con quienes compartí trabajos prácticos, docencia o ambos.

A veces pienso que haber terminado estudiando en Exactas es una de las mejores cosas que me pasó y, a la vez, una de las más inexplicables. Gracias a todas las personas que, de alguna manera, influyeron en que descubriera la computación. Un agradecimiento especial a Vaio, por guiarme en mis primeros pasos en la ciencia y la tecnología. Gracias por darme la oportunidad de descubrir este mundo y por despertar mi amor por la investigación.

Gracias también a todos los profes y ayudantes que lograron transmitir su pasión por la computación, incluso a través de Zoom, y más tarde de forma presencial. Muchos de ustedes son la razón por la cual soy docente hoy. Gracias a la educación pública por darme las herramientas para desarrollarme y aprender tanto.

Un agradecimiento especial a mis alumnos de Laboratorio de Datos e Introducción a la Programación, quienes completaron mi experimento después de mis innumerables súplicas. También gracias a mis amigos, que tuvieron que testear la plataforma para tomar los datos. Aún hoy sueñan con molinos y porciones de pizza.

Esta tesis se desarrolló en el contexto de la beca BIICC. Quiero agradecer especialmente a mis directores, Pablo y Diego, quienes me acompañaron a lo largo de este proceso con su guía y apoyo. Además, extiendo mi agradecimiento a toda la gente del LIAA, quienes siempre me hicieron sentir parte de la vida en el laboratorio, algo que disfruté muchísimo.

Finalmente, quiero agradecer a mis amigas de toda la vida, las de Carmen. Gracias por bancarme siempre y por compartir no sólo las alegrías, sino también las frustraciones.

Este es el cierre de uno de los procesos más transformadores de mi vida. Mi paso por la universidad me dio una nueva manera de ver el mundo y me hizo crecer de forma exponencial.

A los de siempre y los de ahora.
En especial, a mi abuela que nunca entendió que estudio pero estaba orgullosa igual.

Índice general

1. INTRODUCCIÓN

1.1. Motivación y objetivos

En los últimos años, los modelos generativos de imágenes han experimentado una evolución significativa, impulsada principalmente por los avances en el área del Deep Learning y, más recientemente, por los modelos de difusión. Estas técnicas han permitido la generación de imágenes de alta calidad, con aplicaciones que van desde la creación artística hasta la síntesis de datos para tareas de aprendizaje. Sin embargo, persisten preguntas fundamentales sobre la precisión, las limitaciones y el rendimiento de estos modelos en comparación con la percepción y la capacidad creativa humana.

Uno de los objetivos de esta tesis es evaluar modelos de texto a imagen, como Stable Diffusion y DALL·E, en un entorno controlado para determinar con precisión su rendimiento frente al criterio humano en una tarea especifica. El experimento que se va a utilizar fue presentado en "Quantitative Pedagogy: A Digital Two Player Game to Examine Communicative Competence " [8], donde se evaluó qué tan efectivos son los humanos para describir y reproducir una composición geométrica.

La evaluación de modelos generativos representa un desafío significativo debido a la naturaleza subjetiva de las tareas que estos modelos realizan, como la generación de imágenes, texto o datos sintéticos. A diferencia de otros ámbitos del aprendizaje automático, donde las métricas objetivas y bien definidas son suficientes para medir el desempeño, en los modelos generativos es necesario evaluar tanto la calidad como la diversidad de las salidas generadas. Esto ha llevado al desarrollo de diversas estrategias y enfoques para realizar evaluaciones exhaustivas y confiables. En modelos de texto a imagen, se emplean métricas como el CLIP Score [9], que mide la alineación semántica entre una imagen generada y su descripción textual correspondiente. En el caso de modelos de texto, métricas como *perplexity* [1] permiten evaluar la fluidez y coherencia del texto generado. Estas herramientas ofrecen una primera aproximación al desempeño de los modelos, pero no capturan completamente aspectos subjetivos o contextuales. Por esta razón, la evaluación humana sigue siendo fundamental. En este caso, los evaluadores califican aspectos como la calidad, coherencia y fidelidad de las salidas generadas, a menudo utilizando plataformas de *crowdsourcing*. Además, los experimentos comparativos entre humanos y modelos generativos permiten analizar si estos son capaces de igualar o superar el desempeño humano en tareas específicas, proporcionando información clave sobre sus limitaciones.

Ante esta situación, es fundamental desarrollar un método que permita evaluar estos modelos de manera objetiva y analizar sus limitaciones de forma rigurosa. En esta tesis se propone explorar un enfoque innovador basado en reemplazar a los participantes humanos por distintos modelos en un experimento de neurociencia.

Otro aspecto clave de esta tesis será investigar el impacto de distintos métodos de *alignment* en el rendimiento de los modelos generativos con el propósito de entender cuáles son las mejores técnicas para mejorar la performance de los modelos generativos en esta tarea específica. Asimismo, se pretende determinar si, incluso con estas adaptaciones, los modelos pueden alcanzar un rendimiento comparable al de los humanos.

Finalmente, un objetivo esencial es comprender qué componentes de las descripciones son particularmente útiles para los humanos en la tarea de generación de imágenes, pero

no logran el mismo efecto en los modelos, y viceversa. Identificar estas diferencias en el procesamiento y la percepción entre los humanos y los modelos permitirá identificar puntos de divergencia que podrían ser claves para mejorar las capacidades de los sistemas generativos. De este modo, se busca determinar cuál de los métodos analizados demuestra el mejor rendimiento en la generación de imágenes para esta tarea específica.

2. EXPERIMENTO ANTERIOR

2.1. Quantitative Pedagogy: A Digital Two Player Game to Examine Communicative Competence

En 2013, un grupo de investigadores desarrolló un experimento, presentado en [8], con el objetivo de investigar el método óptimo para comunicar conocimiento conceptual entre humanos. Este experimento plantea una dinámica de comunicación entre personas inspirada en el juego " Teléfono Descompuesto" donde la comunicación verbal está sujeta a diversas fuentes de ruido. El experimento fue formulado y llevado a cabo en el ámbito de las ciencias cognitivas antes de la aparición de la inteligencia artificial generativa, por lo que no fue diseñado para integrar dicha tecnología. Su objetivo principal era explorar la competencia comunicativa, especialmente en el contexto de las limitaciones pedagógicas.

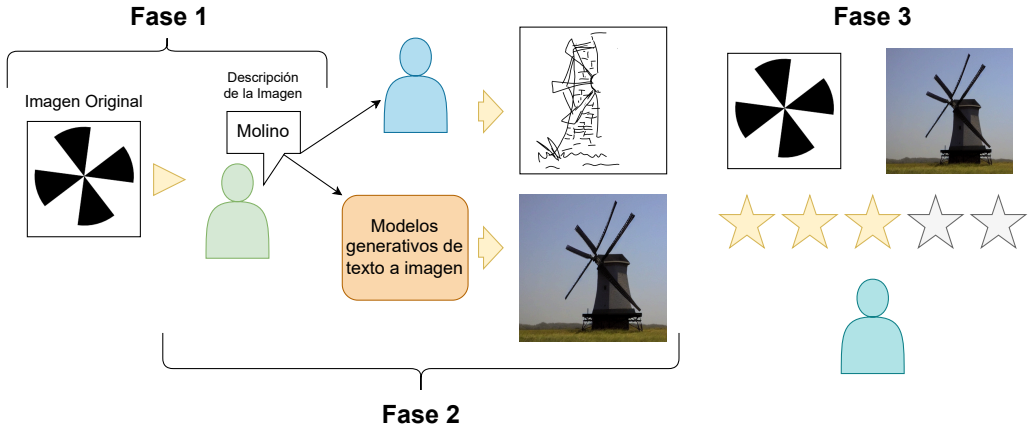


Fig. 2.1: Diagrama de las tres fases del experimento: (1) creación de descripciones textuales basadas en composiciones geométricas, (2) generación de dibujos a partir de estas descripciones, ya sea por participantes humanos o mediante modelos generativos, y (3) evaluación de las imágenes generadas mediante un proceso *crowdsourced*. Este diagrama representa el flujo del experimento original, así como la incorporación de modelos generativos en las fases correspondientes.

El experimento consistió en un juego de dos jugadores donde uno cumple el rol de emisor y el otro de receptor. El juego se desarrolló en tres fases (ver Figura 2.1). En la primera fase, el emisor observaba una composición geométrica básica, como las que se encuentran en la grilla ilustrada en la Figura 2.2 y se le pedía que la describiera en palabras teniendo en cuenta que esa descripción seria utilizada por otro participante para recrear la figura original. Posteriormente en la fase 2, se le asignaba una descripción a un receptor quien tenia que realizar un dibujo basado en la descripción proporcionada. La fase 3 consistía en evaluar las descripciones y los dibujos en una plataforma de *crowdsourcing*, motivada por GWAP (Game with a Purpose, [15]).

El experimento se realizo con un set de composiciones geométricas mostrado en la Figura 2.2. Las imágenes utilizadas fueron diseñadas específicamente para este experimento, y su contenido geométrico junto con su estilo simple responden a la intención de generar

imágenes con una carga emocional neutral. Las imágenes con alto contenido emocional tienden a captar mayor atención, lo que podría introducir un sesgo en este experimento. Por ello, se utilizaron imágenes diseñadas con contenido emocional neutral. Estas se organizaron en cuatro familias, cada una de las cuales incluye cuatro variantes de composiciones que utilizan las mismas figuras geométricas, pero con variaciones en posición, orientación y cantidad.

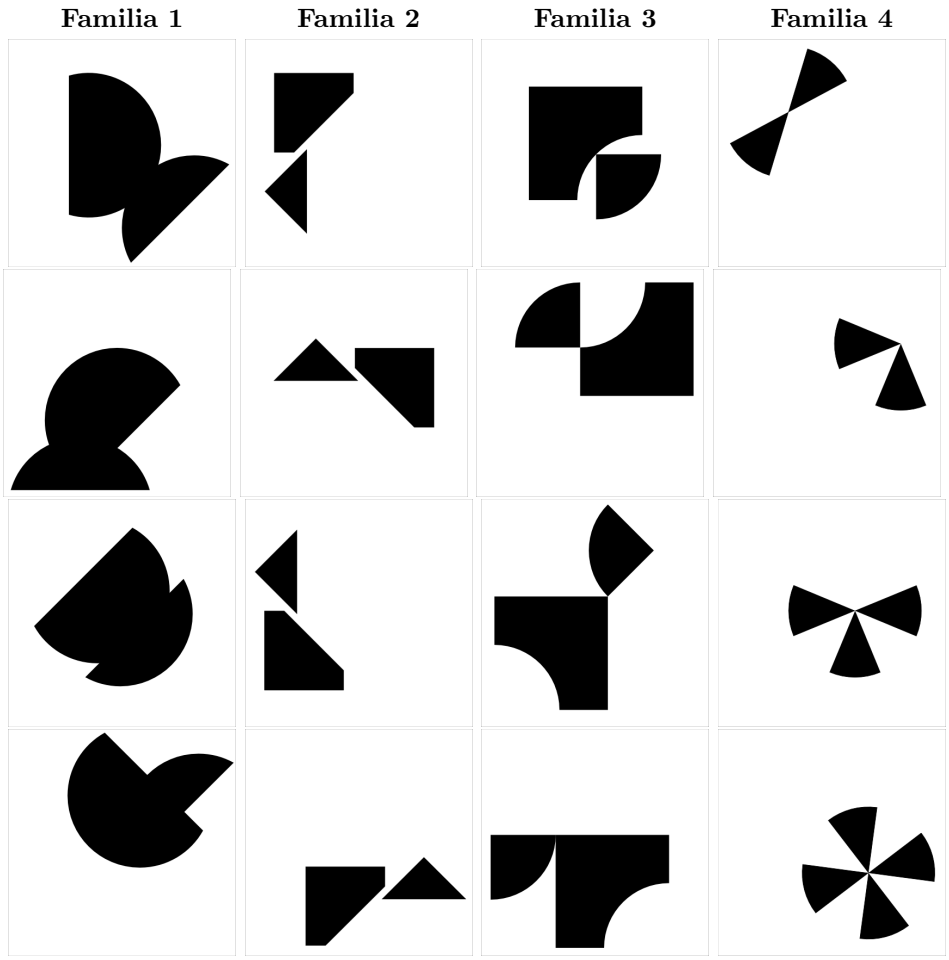
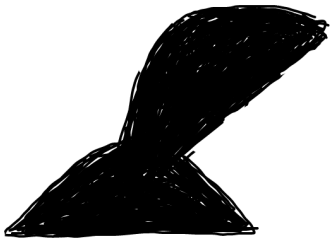


Fig. 2.2: Grilla con las imágenes originales utilizadas durante la Fase 1 del experimento [8]. Cada familia consta de 4 composiciones geométricas que representan variaciones creadas con las mismas figuras.

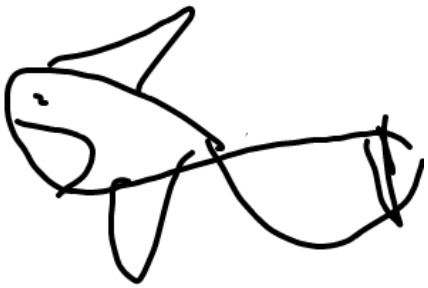
Para recolectar los distintos tipos de datos provenientes de las diferentes fases del experimento se utilizo una aplicación *Android* desarrollada especialmente para esta tarea. Las vistas de la aplicación se pueden ver en la Figura 2.3. En primer lugar se recolectaron descripciones de las imágenes originales (Ver Figura 2.2). La aplicación asignaba de forma aleatoria una de las 16 imágenes originales y solicitaba al usuario que escribiera un texto destinado a ayudar a otra persona a reproducirla. En la segunda fase, se mostraba a la persona una de las descripciones generadas en la fase anterior y se le solicitaba que realizara un dibujo sobre un lienzo blanco, utilizando un trazo negro y siguiendo la descripción como referencia. Las únicas acciones permitidas por la aplicación eran dibujar con negro o borrar.

Finalmente para la ultima fase se le mostraban a los participantes pares de imagen original y dibujo generado por un humano durante la fase 2 y se les pedía que le asignaran una calificación utilizando un sistema de estrellas donde 1 estrella era la peor calificación y 5 estrellas la mejor.

"en la parte inferior izquierda de la pantalla hay un medio circulo con la parte chata hacia abajo ocupando dos tercios del lado sur de la pantalla. esta figura es rellena de color negro. del centro de la figura sale otro medio circulo a 45 grados con la parte chata hacia el sur-este. este medio circulo es levemente mas grande., aprox. 20 porciento mas. nuevamente este circulo esta relleno en color negro. la parte mas alta de este semi circulo apenas supera la mitad de la pantalla."	*"Es un cuadrado al que le quitaron el cuarto inferior derecho. El corte lo hicieron en forma de semicirculo/arco. Entonces queda el lado izquierdo del cuadrado recto, el superior tambien recto y el de la derecha hasta la mitad tambien recto. Despues hay que unir el punto de abajo del lado izquierdo con el punto de abajo del lado derecho con una linea curva (semicirculo, arco). De ahi la porcion que le falta al cuadrado, es como una porcion de pizza (un triangulo que tiene una base curva, en semicirculo), se pone pegada al cuadrado, en ese espacio en blanco que le quedo con la punta apuntando hacia el centro. Todas las figuras estan pintadas de color oscuro."*
	

Tab. 2.1: Pares de dibujos generados por los participantes junto con las descripciones textuales utilizadas para crearlos.

Tanto los dibujos generados como las descripciones fueron muy interesantes. Hubo descripciones que eras demasiado metafóricas, lo que provocó que los dibujos generados no se alinearan con las figuras originales. Dos ejemplos de estos casos se muestran en los pares de dibujos y descripciones en Tabla 2.2. Por otro lado, también se observaron casos en los que la creatividad de los participantes produjo buenos resultados. Algunas descripciones, de carácter más procedimental, utilizaron metáforas para enriquecer las instrucciones, generando dibujos de alta calidad, como se puede apreciar en Tabla 2.1.

“hoguera”	*“pez”*
	

Tab. 2.2: Pares de dibujos generados por los participantes junto con las descripciones textuales utilizadas para crearlos.

La recolección de datos se realizo en el evento TEDx 2013 en Buenos Aires. Aproximadamente 700 personas participaron. Se recolectaron 689 descripciones y 621 dibujos. Ademas 4000 calificaciones de las distintas partes del experimento.

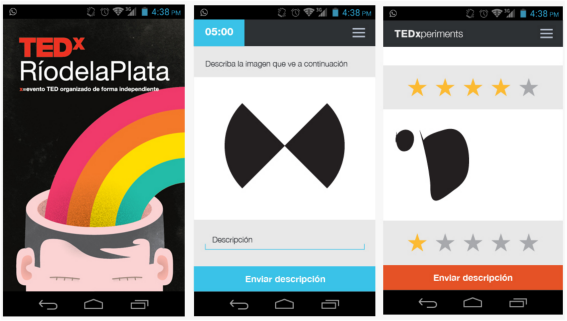


Fig. 2.3: Vistas de la Aplicación Android utilizada para la evaluación correspondiente a la fase 3 del experimento.

Como resultado de este experimento se encontró que las descripciones que producen mejores dibujos son coherentes y procedurales. En contraste las descripciones mas creativas, metafóricas o que hacen uso de conceptos matemáticos no mejoran la reproducción de las figuras.

2.1.1. Reformulacion del experimento en el contexto de AI Generativa

El diseño original del experimento puede incorporar modelos de inteligencia artificial generativa en varias fases. En la primer fase, los modelos de imagen a texto (como BLIP [7]) ofrecen la capacidad de convertir imágenes en descripciones textuales. En la segunda fase, los modelos de lenguaje de gran escala (LLMs) poseen el potencial de reescribir y mejorar las descripciones, mientras que los modelos de texto a imagen (como Stable Diffusion [12]) pueden generar replicas de los dibujos originales a partir de las descripciones. Por último,

en la ultima fase, los Visual Language Models (VLMs, como CLIP [9]) pueden utilizarse para evaluar la similitud entre las imágenes originales y las evaluadas.

Ante esta situación, se decidió incorporar inteligencia artificial generativa en una de las fases del experimento, manteniendo intacto el diseño original en las demás etapas. Dado que el principal interés radicaba en el uso de modelos de texto a imagen, se optó por utilizarlos para reemplazar a los humanos en la fase 2 como se ilustra en la Figura 2.1. Para ello, se alimentaron diversos modelos de texto a imagen con las descripciones generadas por los humanos en el experimento original, con el objetivo de producir imágenes que se asemejaran a las composiciones geométricas originales. Posteriormente, se comparará el rendimiento de estos modelos con el desempeño de los humanos.

2.1.2. Datos encontrados del experimento anterior

Inicialmente, se pensaba que los datos recolectados durante este experimento no estaban disponibles. Sin embargo, finalmente se localizaron los dibujos junto con las descripciones que los generaron, apareados con la figura original de la cual surgieron. Del conjunto total de datos, se identificaron 595 descripciones de imágenes originales y 621 dibujos. La distribución de las descripciones y los dibujos correspondientes a cada familia e imagen se encontraba bastante equilibrada, como se muestra en Tabla 2.3 y Tabla 2.4.

Familia \ Imagen	**1**	**2**	**3**	**4**
1	40	41	39	40
2	38	38	32	32
3	36	38	34	29
4	40	40	43	35

Tab. 2.3: Cantidad de descripciones del experimento anterior por familia e imagen.

Familia \ Imagen	**1**	**2**	**3**	**4**
1	41	42	42	40
2	39	38	33	35
3	39	39	37	31
4	40	44	43	38

Tab. 2.4: Cantidad de dibujos del experimento anterior por familia e imagen.

2.1.3. Datos utilizados

Para los experimentos de generación de imágenes, se utilizó solo una porción de las descripciones originales con el objetivo de reservar datos no utilizados para posibles necesidades futuras. Las descripciones se seleccionaron mediante un muestreo balanceado, considerando la familia y el número de imagen. De las 595 descripciones disponibles, se emplearon 474.

3. MODELOS GENERATIVOS DE TEXTO A IMAGEN

3.1. Evolución de los Modelos Generativos de Texto a Imagen

Si bien el público general descubrió los modelos generativos en estos últimos dos años con la explosión de ChatGPT, DALL·E y Stable Diffusion. Llegar a los modelos generativos de texto a imagen que están en boga hoy en día fue un proceso impulsado por distintos avances en redes neuronales y procesamiento del lenguaje natural. Estos modelos se basan en una combinación de redes neuronales profundas y técnicas de aprendizaje automático que, en conjunto, permiten la generación de imágenes complejas y de alta calidad a partir de *prompts*, es decir, descripciones textuales de las imágenes deseadas.

Antes de hablar de los modelos generativos es necesario mencionar los avances en redes neuronales que permitieron tener representaciones de imágenes que hoy usan los modelos generativos. En 2017 el campo de la inteligencia artificial experimentó una revolución fundamental gracias al desarrollo de los Transformers, introducidos en "Attention is All You Need" [14]. Este avance marcó un cambio de paradigma en el procesamiento de lenguaje natural (NLP) y posteriormente en el aprendizaje multimodal. Los Transformers, con su mecanismo de *self-attention*, permitieron a los modelos procesar relaciones contextuales en secuencias largas de datos, algo que las redes neuronales recurrentes (RNNs) y las convolucionales (CNNs) luchaban por manejar eficientemente.

La capacidad de los Transformers para capturar dependencias complejas en los datos multimodales fue esencial para integrar texto e imágenes en un marco cohesivo. Por ejemplo, CLIP [9] utilizó Transformers para aprender un espacio latente común para texto e imágenes, un paso crucial que guió el desarrollo de modelos generativos de imágenes.

Con esta base sólida en la representación del texto y las relaciones semánticas, se pudo avanzar hacia la creación de imágenes a partir de descripciones textuales, utilizando modelos generativos como las GANs y, más adelante, los modelos de difusión.

El desarrollo de modelos generativos comenzó con redes como los Variational Autoencoder (VAEs [6]) y los Generative Adversarial Networks (GANs). Los GANs, introducidos en "Generative Adversarial Networks" [3], abrieron un campo completamente nuevo al enfrentar dos redes (generadora y discriminadora) en un juego competitivo. La red generadora intenta crear imágenes convincentes y la red discriminadora evalúa si las imágenes generadas son realistas o no. En principio estas redes no era guiadas usando contenido textual, de todas formas los primeros intentos de generar imágenes basadas en texto se apoyaron en esta arquitectura, pero a menudo enfrentaban problemas de coherencia y calidad en la interpretación del texto. Un ejemplo de esto fueron los StackGANs [16] que introdujeron un enfoque en dos etapas. En la primera etapa, se generaba una imagen de baja resolución a partir del texto, y en la segunda, se refinaba para producir detalles más realistas. Este enfoque marcó un punto de inflexión al lograr imágenes visualmente atractivas y más coherentes con las descripciones textuales.

Más recientemente, la introducción de los modelos de difusión como DALL·E [11], DALL·E 2 [10] y Stable Diffusion [12] ha permitido una generación de imágenes mucho más detallada y de alta calidad. Los modelos de difusión introducidos en "Denoising diffusion probabilistic models" [4] son una clase de modelos probabilísticos que trabajan con variables latentes y se fundamentan en la idea de invertir un proceso de degrada-

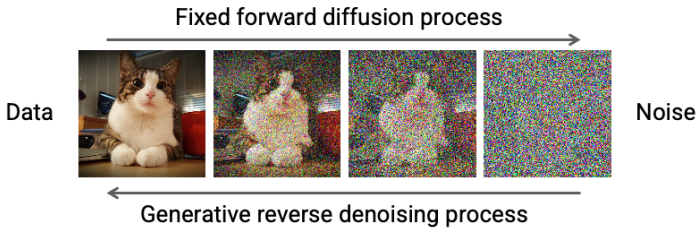


Fig. 3.1: Diagrama del proceso de difusión hacia adelante (*forward diffusion process*) y el proceso inverso de eliminación de ruido (*reverse denoising process*), ilustrando cómo las imágenes originales se degradan progresivamente con ruido y luego se reconstruyen eliminándolo paso a paso.

ción para reconstruir o generar datos. El proceso de difusión fue inspirado de ideas de la termodinámica fuera del equilibrio y se da en dos fases (ver Figura 3.1):

- **Proceso de Difusión (Forward Process):** Se añade ruido gaussiano incrementalmente a los datos originales (por ejemplo, imágenes) a lo largo de un número determinado de pasos T hasta que los datos se transforman en ruido puro (ver Figura 3.2). Podemos modelar este proceso como:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I})$$

  donde:

  - $\mathbf{x}_t$ es el estado de los datos en el paso t.
  - β_t controla la cantidad de ruido añadido.
  - $\mathcal{N}$ representa una distribución gaussiana.

- **Proceso de Denoising (Reverse Process):** Se aprende un proceso inverso para eliminar el ruido de manera iterativa, generando datos a partir del ruido inicial (Ver Figura 3.2). Este proceso se modela con una red neuronal que aprende a predecir la distribución condicional $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t))$$

  Aquí:

  - μ_θ es la media predicha por la red neuronal para reconstruir los datos en el paso $t-1$.
  - Σ_θ puede ser fijada o aprendida como la varianza condicional.

Este enfoque ha demostrado ser particularmente efectivo en la generación de imágenes a partir de texto debido a su capacidad para generar detalles finos y complejos, un aspecto en el que los modelos basados en GANs a menudo enfrentaban dificultades. Modelos

como Stable Diffusion emplean esta técnica para mejorar la resolución y coherencia de las imágenes generadas, adaptándose bien a descripciones detalladas o complejas.

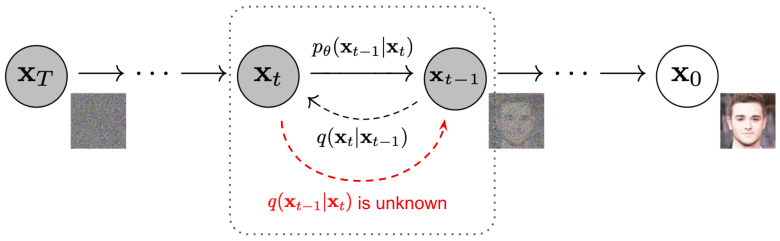


Fig. 3.2: Diagrama que muestra las probabilidades de transición en el proceso de difusión hacia adelante y las probabilidades inversas utilizadas en el proceso de denoising, tomado del trabajo presentado en [4].

Una vez introducido el proceso de difusión surge la posibilidad de guiar este proceso hacia resultados específicos incorporando información adicional, en particular descripciones textuales llamados *prompts*, que actúan como contexto para influir en la generación. Este enfoque, conocido como *text conditioning*, combina las capacidades del modelo de difusión con representaciones semánticas del texto para alinear las imágenes generadas con las instrucciones dadas. Esta capacidad resulta especialmente interesante en el contexto de esta tesis ya que es exactamente lo mismo que hacen los humanos durante la fase 2 del experimento.

Un mecanismo de *text conditioning* que ha demostrado buenos resultados es el que utiliza Stable Diffusion. Sin embargo, antes de abordar este mecanismo, es necesario introducir la arquitectura de este modelo, la cual se ilustra en Figura 3.3. Este modelo fue seleccionado entre las opciones disponibles debido a que es de código abierto y cuenta con una amplia comunidad que desarrolla herramientas que facilitan su implementación y personalización. Además, Stable Diffusion se distingue por su capacidad para generar imágenes de alta resolución con un nivel notable de detalle, lo que lo hace ideal para una amplia gama de aplicaciones. Estas incluyen desde la creación artística y el diseño gráfico hasta la generación de contenido visual para medios interactivos. Su diseño eficiente y flexible lo posiciona como un avance significativo en el ámbito de los modelos generativos, al combinar innovación técnica con practicidad para abordar los desafíos más complejos en la generación de imágenes.

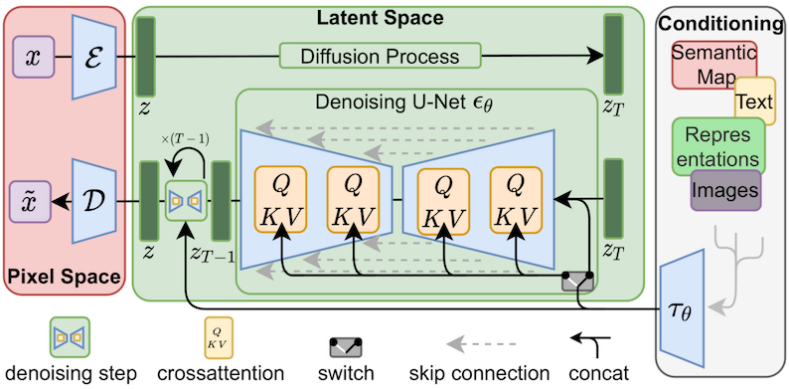


Fig. 3.3: Diagrama de la arquitectura de Stable Diffusion. Este diagrama está tomado de [12].

El modelo comienza utilizando un Autoencoder Variacional (VAE) para transformar las imágenes del espacio de píxeles al espacio latente. El VAE consta de dos partes principales: un *encoder*, que comprime las imágenes en una representación latente más compacta conservando sus características esenciales, y un *decoder*, que reconstruye las imágenes en el espacio de píxeles a partir de las representaciones latentes generadas. Este enfoque no solo reduce la dimensionalidad de los datos, sino que también permite realizar el proceso de difusión en un espacio más manejable.

La parte fundamental de Stable Diffusion es la red neuronal UNet, encargada del proceso de denoising en el espacio latente. Esta UNet utiliza una arquitectura en forma de "U", con conexiones de salto que vinculan las capas de codificación y decodificación, asegurando la preservación de los detalles durante el proceso de refinamiento. Además, el UNet incorpora módulos de *cross-attention*, que permiten integrar información textual y visual, alineando las características del espacio latente con las instrucciones textuales proporcionadas.

El entrenamiento del modelo sigue el enfoque estándar de los modelos de difusión. En el proceso *forward*, se añade ruido gaussiano progresivamente a las representaciones latentes, transformándolas en ruido puro. En el proceso *reverse*, el UNet aprende a eliminar este ruido de forma iterativa, refinando las representaciones latentes hasta reconstruir imágenes coherentes.

El flujo completo de generación de imágenes con Stable Diffusion incluye varias etapas. Primero, se ingresa una descripción textual que se convierte en un embedding mediante CLIP. Luego, el modelo genera una muestra de ruido aleatorio en el espacio latente, que se refina iterativamente utilizando el embedding textual como guía. Finalmente, el *decoder* del VAE convierte la representación latente refinada en una imagen en el espacio de píxeles.

El mecanismo de *text conditioning* es una parte clave del modelo. En primer lugar se utilizan *embeddings textuales*, como los generados por CLIP, que nos permiten representar al texto como un vector numérico que captura las relaciones semánticas del texto en un espacio latente compartido entre texto e imagen. El proceso de obtener los embeddings requiere tokenizar el texto, es decir, dividir el texto en tokens que pueden ser palabras o subpalabras y luego generar los embeddings correspondientes para cada token. Luego, estos embeddings se integran en el proceso de denoising a través de las capas de *cross-attention* de la UNet, lo que permite generar imágenes consistentes con las descripciones proporcionadas. Este mecanismo es fundamental para aplicaciones como la generación de imágenes a partir de texto, donde resulta crucial que las imágenes estén alineadas con las descripciones textuales.

Este modelo incorpora el mecanismo de *cross-attention* para integrar descripciones textuales en el proceso de generación de imágenes. *Cross-attention* es una técnica que permite al modelo alinear directamente características del texto con las características visuales latentes, asegurando que las imágenes generadas reflejen con precisión la descripción textual. En el contexto de NLP, *Cross-attention* es un mecanismo que permite al modelo encontrar la relación que hay entre los tokens de dos secuencias distintas. En los modelos de texto a imagen este mecanismo se usa para calcular una ponderación entre las características latentes de la imagen y las representaciones textuales para ajustar las características generadas en función del contexto textual.

Ademas Stable Diffusion tiene el hiperparametro *Classifier-Free Guidance* que permite mejorar la calidad de las imágenes generadas al proporcionar un mayor control sobre la alineación entre las descripciones textuales y las imágenes resultantes. Esta técnica

equilibra dos objetivos principales: la fidelidad al texto de entrada y la diversidad visual. Durante el proceso de generación, el modelo realiza dos predicciones: una condicionada al texto de entrada y otra no condicionada. Ambas predicciones se combinan mediante un factor de guidance, que ajusta el nivel de influencia del texto sobre la generación de imágenes. Un valor de guidance más alto refuerza la fidelidad al texto, aunque puede sacrificar la naturalidad y diversidad de las imágenes. Esta técnica no solo simplifica el sistema al eliminar la necesidad de un clasificador externo, sino que también permite ajustar dinámicamente el balance entre calidad visual y alineación semántica.

La evolución hacia los modelos actuales de texto a imagen ha sido un proceso de constante innovación, marcado por distintos avances de las redes neuronales y de modelos específicos para esta tarea. Modelos como DALL-E, desarrollado por OpenAI, y Stable Diffusion, desarrollado por Stability AI, representan la cúspide de estos avances. Ambos modelos se entrenan en conjuntos masivos de datos que combinan imágenes con descripciones textuales, lo que permite una mejor comprensión y generación de imágenes a partir de una amplia variedad de contextos y estilos. Esta combinación de grandes conjuntos de datos, arquitecturas avanzadas de Transformers y modelos de difusión ha permitido a DALL-E y Stable Diffusion capturar la esencia del texto e interpretarla en representaciones visuales con un nivel de detalle y coherencia sin precedentes. La capacidad de estos modelos para interpretar y generar contenido visual a partir de lenguaje natural abre nuevas posibilidades en el ámbito de la creación visual y plantea preguntas importantes sobre su capacidad y limitaciones, aspectos que serán explorados a fondo en esta tesis.

3.2. Experimento 1: Descripciones de humanos en español como entrada para Stable Diffusion

El primer experimento realizado consistió en utilizar las descripciones producidas por humanos en el experimento original [8] directamente como *prompts* para Stable Diffusion v1.5. Aunque no se esperaba que este modelo comprendiera completamente los *prompts* en español, se buscaba obtener una primera aproximación al experimento y poder explorar la metodología para los experimentos de generación de imágenes. Además, aunque se asumía que el modelo no estaba específicamente entrenado para funcionar de manera óptima con descripciones en español, no se disponía de certezas al respecto.

Inicialmente se utilizo scripts de Python usando el paquete Diffusers de HuggingFace para realizar inferencia y así generar las imágenes. Pero este método tenia un problema ya que algunas descripciones eran demasiado extensas y se truncaban al tokenizarse como entrada al modelo ya que CLIP solo soporta 77 tokens. Frente a esto se empleó AUTOMATIC1111 que es una UI que facilita el uso de estos modelos y en particular resuelve el manejo de estas descripciones para incorporarlas al modelo. Una vista de esta UI se puede ver en Figura 3.4. El manejo de *prompts* largos se hace cortando el *prompt* en fragmentos de 75 tokens. Para cada fragmento, se genera un *embedding* utilizando CLIP , tras lo cual se concatenan todos los *embeddings* y el resultado se pasa a la UNet. Para poder realizar el experimento de generación de imágenes se desarrolló una modificación en AUTOMATIC1111 que permite generar imágenes utilizando un conjunto de datos que contiene las descripciones y otros datos relacionados con la imagen original correspondiente. Las imágenes generadas se obtuvieron configurando un valor de `cfg=7` [1] y un *seed*

[1] Este valor fue seleccionado en base a una exploración inicial viendo su efecto al generar un set de imágenes y también respetando lo que usa la comunidad como valor default.

aleatorio, el cual se incluyó en el nombre final de cada imagen.

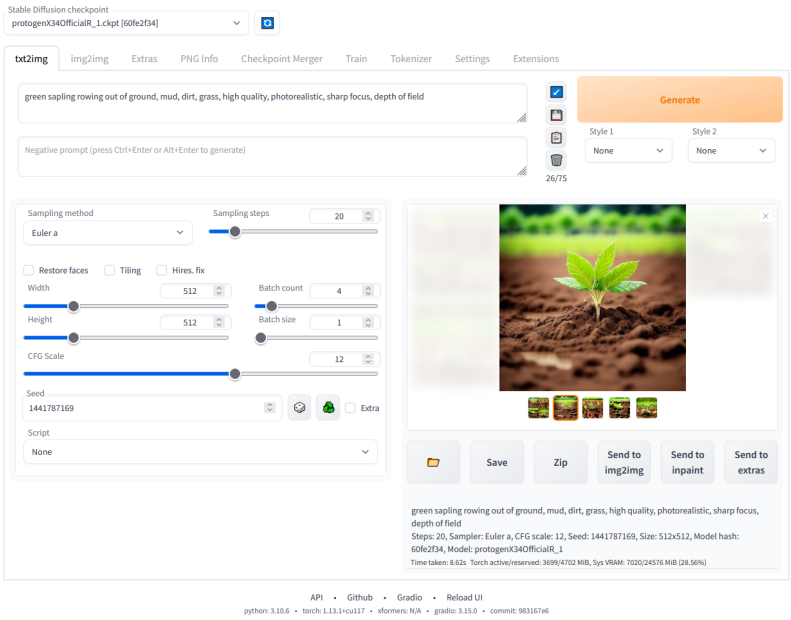


Fig. 3.4: Interfaz de usuario de Automatic1111 utilizada para generar imágenes con Stable Diffusion v1.5.

Las imágenes generadas no mostraron en su mayoría similitudes con las originales (Ver Tabla 3.1 y Tabla 3.2). Sin embargo, debido a la gran cantidad de imágenes producidas, resulta complejo evaluar el experimento en su totalidad. Además, al analizar las descripciones, se observó que muchas de ellas no eran de buena calidad, y que los resultados generados por Stable Diffusion no eran incorrectos en relación con dichas descripciones.

"es un circulo que esta dividido en 8 triangulos, como una pizza, pero una porcion esta pintada de negro y la otra no, una de negro y la otra no, etc. a las que quedan sin pintar hay que borrarles la linea. y listo!"	*"es una pizza mirada desde arriba con solo 3 porciones (color negro). La pizza esta dividida en 8. Las 3 porciones son las que en un reloj estaran posicionadas a las 3, 6 y 9 horas."*
	

Tab. 3.1: Pares de imágenes generadas por Stable Diffusion en Sección 3.2 junto con las descripciones textuales utilizadas para crearlos.

Dado que la generación de imágenes en modelos como Stable Diffusion está directamente relacionada con el proceso de tokenización, que descompone el texto en unidades más pequeñas llamadas *subwords*. Este proceso es esencial porque los modelos no trabajan con palabras completas, sino con tokens que representan fragmentos de palabras, raíces, sufijos o incluso palabras enteras, dependiendo de su frecuencia en el vocabulario del modelo. Cuando se introducen palabras en español, el modelo intenta reconocerlas en su vocabulario. Si una palabra existe tal cual, como "pizza", el tokenizador la reconoce como un único token, lo que facilita que el modelo asocie esta entrada con imágenes precisas. Esto ocurre porque "pizza" es igual tanto en español como en inglés, y el modelo ya tiene datos relacionados con esta palabra. Sin embargo, si una palabra no está en el vocabulario, el modelo la divide en *subwords* conocidas. Por ejemplo, la palabra "mariposa"podría dividirse en "mari" y "posa". Estas divisiones a veces permiten una interpretación adecuada, pero si las subwords no corresponden bien al concepto global, el modelo podría generar algo incorrecto o fuera de contexto. Stable Diffusion esta entrenado principalmente con datos en inglés, lo que significa que las palabras y frases en este idioma están mejor representadas y asociadas a conceptos visuales claros. Esto explica por qué los prompts en inglés tienden a generar resultados más precisos. Al usar español, si una palabra es similar a su equivalente en inglés o comparte tokens comunes, como "artista" y "artist", el modelo puede interpretarla correctamente. Sin embargo, palabras específicas del español o conceptos menos frecuentes podrían no tener una representación clara, lo que requiere descripciones más detalladas para evitar confusiones en la generación.

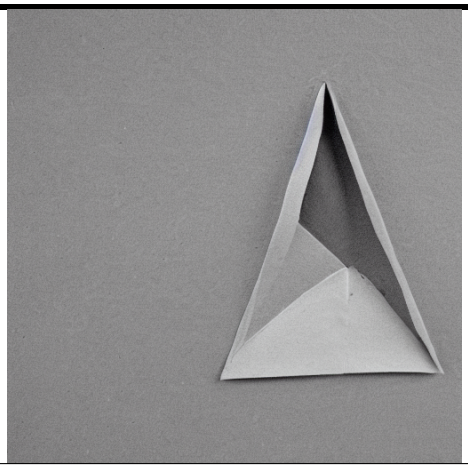
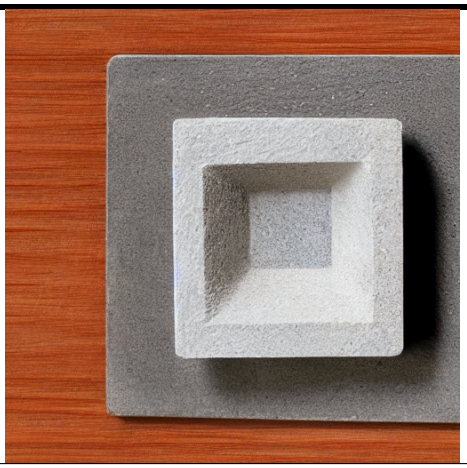
"en la hoja en blanco, sobre la izquierda, centrado, con margen, se encuentra un triangulo con la punta indicando a la izquierda y debajo de este triangulo se encuentra un cuadrado que no esta completo. esta parte incompleta del cuadrado es el mismo triangulo solo que esta parte incompleta esta arriba a la derecha del cuadrado."	*"cuadrado"*
	

Tab. 3.2: Pares de imágenes generadas por Stable Diffusion en Sección 3.2 junto con las descripciones textuales utilizadas para crearlos.

Un caso interesante se observó en las imágenes de la familia 3, donde aparecieron numerosas imágenes de jugadores de fútbol y, en particular, de un jugador específico (ver Tabla 3.2). Esto ocurrió porque las descripciones utilizadas suelen incluir frases como "dibuja un cuadrado", y el modelo interpretó "cuadrado" como una referencia al apellido del futbolista Juan Cuadrado, quien forma parte de la selección de Colombia.

3.3. Experimento 2: Descripciones de humanos en inglés como entrada para Stable Diffusion

Ante los resultados obtenidos en el experimento anterior (Sección 3.2), se decidió traducir las descripciones al inglés con el objetivo de evaluar si esto generaba imágenes que se aproximaran más a las composiciones originales. La traducción de las descripciones producidas por humanos se realizó utilizando chatGPT. Posteriormente, al igual que en el experimento Sección 3.2, se generaron las imágenes utilizando la misma metodología y el mismo conjunto de hiperparámetros. Aunque muchas de las imágenes generadas no mostraron altos niveles de geometría y ninguna fue significativamente similar a las originales, se observó una mejora en los resultados. Por ejemplo, al comparar las mismas descripciones utilizadas en las imágenes de Tabla 3.2 con los resultados obtenidos en Tabla 3.3, se aprecia una mayor relación con las composiciones originales. A pesar de esto, las imágenes generadas siguen estando lejos de replicar el estilo de las originales.

"On a blank sheet, on the left side, centered with margins, there is a triangle with its upper vertex pointing to the left. Below this triangle, there is an incomplete square. The incomplete part of the square corresponds to the same triangle, positioned in the upper right corner of the square. "	*"Scuare"*
	

Tab. 3.3: Pares de imágenes generadas por Stable Diffusion en Sección 3.3 junto con las descripciones textuales utilizadas para crearlos.

Dado que muchas descripciones no resultan adecuadas para generar las imágenes deseadas y que el número de imágenes a evaluar es demasiado grande, se desarrolló un método para cuantificar el nivel de "geometría"presente en las imágenes. Este enfoque permite comparar los distintos experimentos de generación utilizando una métrica concreta, proporcionando un criterio objetivo para evaluar los resultados.

3.3.1. Cuantificar geometría

Para cuantificar la geometría de las imágenes, se utilizaron sus *embeddings*. El método empleado consistió en comparar los *embeddings* de las imágenes mediante la distancia coseno para determinar el nivel de geometría. Además, se utilizó un modelo de KNN para predecir este nivel en las imágenes generadas. El objetivo de este enfoque es medir cuán geométrica es una imagen y, de este modo, permitir la comparación entre diferentes experimentos para identificar cuál generó mayor cantidad de geometría. Si bien esta métrica no es perfecta, ofrece una aproximación que facilita evaluar la generación de un experimento sin necesidad de analizar todas las imágenes en detalle.

Para entrenar el modelo de KNN se utilizo un dataset que consto de:

- Imágenes en blanco y negro de geometría generadas automáticamente con scripts en Python.
- Imágenes generadas durante las batallas de prompts con colores.
- Imágenes en blanco y negro generadas con Stable Diffusion utilizando un datasets de prompts variados.

Se dividieron las imágenes de forma balanceada para obtener un conjunto de desarrollo y otro *hold-out* para la evaluación final. Se exploro 3 redes distintas para generar los embeddings: CLIP, VGG16 y EfficientNet. Se utilizo *5-fold cross-validation* para determinar el mejor modelo. Los mejores resultados se obtuvieron con CLIP y `K=3`.

Para cuantificar el nivel de geometría en un experimento de generación, se midió la geometría de cada imagen generada de manera individual y, posteriormente, se calculó el promedio de estas mediciones. Los resultados para los experimentos realizados se encuentra en Tabla 3.4.

Experimento	**Métrica de geometría**
Stable Diffusion con descripciones en español	0.09
Stable Diffusion con descripciones en ingles	0.64

Tab. 3.4: Métrica de geometría de los distintos experimentos de generación de imágenes.

4. MÉTODOS DE ALIGMENT

4.1. Introducción a métodos de alignment

En el uso de modelos de generación de imágenes una tarea esencial es alinear el contenido generado a objetivos específicos o personalizaciones deseadas. Este proceso, conocido como alignment, abarca una variedad de técnicas que buscan ajustar el comportamiento del modelo para incorporar nuevos conceptos, estilos visuales o comportamientos específicos. Entre los métodos más utilizados para este fin se encuentran *fine-tuning*,*LoRA fine-tuning* y técnicas como *Textual Inversion*.

El *fine-tuning* completo implica ajustar todos los pesos del modelo entrenándolo nuevamente con datos adicionales. Si bien es una solución poderosa, también es extremadamente costosa en términos computacionales y de almacenamiento, ya que requiere acceso a infraestructura de alto rendimiento, grandes cantidades de datos, y puede resultar en sobreajustes si los nuevos datos no están cuidadosamente equilibrados. Además, también es posible "romper" el modelo, es decir, hacerlo incapaz de generar imágenes realistas, un fenómeno conocido como *catastrophic forgetting*.

Para visualizar de forma más clara los recursos necesarios para realizar el fine-tuning de un modelo de texto a imagen, se puede tomar como ejemplo un caso concreto. Para ajustar Stable Diffusion v1.5 utilizando los scripts de Diffusers, es necesario contar con una GPU con al menos 24 GB de memoria VRAM.

Alternativamente, técnicas más ligeras como *LoRA fine-tuning* y *Textual Inversion* han ganado popularidad debido a su eficiencia. Estas metodologías permiten personalizar el modelo sin modificar todos sus pesos, lo que las hace más accesibles para investigadores y creadores con recursos limitados. Por ejemplo, LoRA introduce modificaciones únicamente en matrices específicas del modelo, reduciendo significativamente la complejidad computacional, mientras que *Textual Inversion* crea embeddings personalizados que capturan conceptos o estilos visuales únicos, sin necesidad de alterar el modelo base.

Estas técnicas permiten que los modelos de difusión sean más versátiles y personalizables, equilibrando costos computacionales y calidad de los resultados, lo que ha democratizado su uso en aplicaciones creativas y comerciales. Sin embargo, la elección del método depende de los recursos disponibles y de los objetivos específicos del usuario, con el *fine-tuning* siendo preferido para ajustes amplios y las técnicas ligeras para adaptaciones rápidas y específicas.

4.2. Textual Inversion

Textual Inversion es una técnica que permite a los modelos generativos de texto a imagen asociar nuevos conceptos o estilos visuales con tokens específicos de texto, sin necesidad de modificar los pesos de la red base. Fue introducida en "An image is worth one word: Personalizing text-to-image generation using textual inversion" [2]. Esta técnica de alingment ofrece una forma eficiente de personalizar la generación de imágenes, ampliando las capacidades del modelo al integrar conceptos visuales altamente específicos.

El mecanismo de *Textual Inversion* se basa en entrenar *embeddings* textuales personalizados que capturen las características visuales de un concepto deseado. Durante el

proceso de entrenamiento, el modelo utiliza un conjunto de imágenes representativas del concepto objetivo, como un objeto único, una persona o un estilo artístico. Este conjunto de imágenes debe ser variado, pero consistente en términos de estilo o características visuales. El concepto puede ser, por ejemplo, un estilo artístico único, un objeto que no esté en el conjunto de datos original, o un atributo característico de un sujeto. A partir de estas imágenes, se ajusta iterativamente un vector de *embeddings* que actúa como una "palabra clave" para el concepto. Este vector reemplaza una cantidad predefinida de tokens en el vocabulario del modelo y se entrena para maximizar la coherencia entre las imágenes generadas y las características visuales del concepto proporcionado. Por ejemplo, si se desea representar un estilo de pintura específico, se puede crear un token como "[miEstilo]" que el modelo podrá reconocer después del entrenamiento.

El proceso de entrenamiento se enfoca en ajustar el *embedding* de texto personalizado sin alterar los pesos del modelo base, lo que hace que esta técnica sea modular y eficiente. Se quiere que el modelo incorpore la correspondencia entre el nuevo token y el concepto visual proporcionado. La función de pérdida utilizada compara las imágenes generadas con las imágenes originales, asegurando que el *embedding* capture de manera efectiva el concepto deseado (ver Figura 4.1). Una vez entrenado, el token personalizado puede incluirse en cualquier prompt para influir en el resultado de la generación. Por ejemplo, al incluir "[miEstilo]" en la descripción textual, el modelo generará una imagen que refleje el estilo visual específico que el token representa. Este proceso permite al usuario controlar y personalizar el contenido visual sin necesidad de añadir nuevas capas o redes.

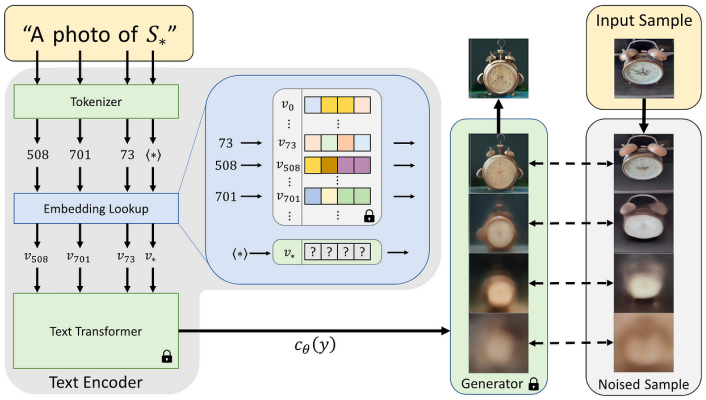


Fig. 4.1: Diagrama del proceso de transformar el texto en *embeddings* y como estos son utilizados en el método de *textual inversion* para obtener el *embedding* asociado a la palabra clave. Durante este proceso, el *embedding* se optimiza mediante un objetivo de reconstrucción.

La implementación en Stable Diffusion aprovecha el mecanismo de Text Conditioning basado en *embeddings* preentrenados de CLIP [9]. En lugar de utilizar únicamente los *embeddings* predefinidos para las palabras existentes, *Textual Inversion* crea un nuevo vector en el espacio latente textual. Este vector se utiliza durante el proceso de difusión, guiando el refinamiento iterativo de las características visuales para que reflejen el concepto asociado al token.

Una de las principales ventajas de Textual Inversion es su capacidad para extender el vocabulario del modelo sin necesidad de reentrenarlo completamente. Esto facilita la personalización, permitiendo a los usuarios incorporar conceptos únicos, como personas,

lugares o estilos específicos, en las imágenes generadas. Esto es importante ya que se reduce fuertemente los recursos de computo necesarios en comparación con otras técnicas de *alignment*. Además, los *embeddings* entrenados son reutilizables y modulares, lo que permite compartirlos o utilizarlos en diferentes contextos.

En términos prácticos, el método se utiliza ampliamente en la creación de retratos o generación de imágenes con nuevos estilos artísticos. Sin embargo, *Textual Inversion* presenta ciertas limitaciones. Por ejemplo, la calidad del concepto aprendido depende en gran medida del conjunto de datos proporcionado para el entrenamiento y de si el modelo conoce ese concepto. Esta técnica puede ser menos eficaz para capturar conceptos que requieren alta variabilidad, como expresiones faciales o múltiples poses de una persona. Además, aunque el proceso es más rápido que reentrenar un modelo completo, aún puede ser computacionalmente costoso. Finalmente, los conceptos aprendidos pueden estar limitados al contexto de las imágenes utilizadas durante el entrenamiento, lo que afecta su capacidad de generalización y versatilidad a la hora de generar imágenes.

En conclusión, *Textual Inversion* es una herramienta poderosa y simple para personalizar la generación de imágenes en modelos como Stable Diffusion, ampliando sus capacidades mediante el aprendizaje de nuevos conceptos visuales sin alterar la arquitectura base. Este enfoque se destaca por su flexibilidad, modularidad y capacidad para generar imágenes fieles a conceptos específicos y se implementa en el ecosistema de modelos basados en difusión.

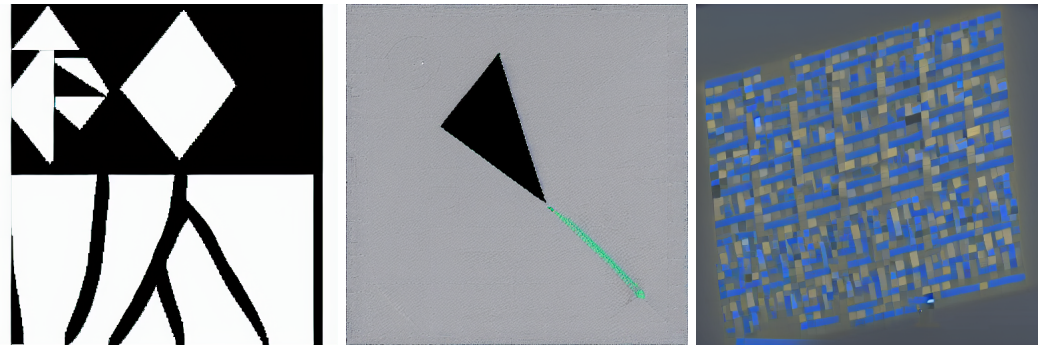
4.3. Experimento 3: Descripciones de humanos en en inglés como entrada para Stable Diffusion con Textual Inversion generado con figuras geométricas

Como siguiente paso, dado que Stable Diffusion con su preentrenamiento base no parece generar imágenes que contengan geometría ni se alinean con el estilo de dibujos 2D buscado, se exploró el método de *Textual Inversion*. Este enfoque permite incorporar a cada *prompt* nuevos tokens que posicionan la generación en un sector del espacio latente más cercano a las imágenes deseadas. Cabe destacar que este método no implica agregar información nueva al modelo, ya que no modifica sus pesos, sino que trabaja dentro del espacio latente preexistente tal como se explico anteriormente.

Para implementar esta técnica, se utilizó un conjunto de 300 imágenes geométricas generadas de manera sintética con scripts en Python. Usando el método de *Textual Inversion*, se aprendió un nuevo embedding que reemplaza 10 tokens asociados a la palabra "Geometría" para reforzar el alineamiento con las características deseadas. Este procedimiento busca orientar al modelo hacia la creación de imágenes más cercanas al estilo geométrico y 2D requerido. [1]

Si bien esta técnica generó imágenes con un contenido más geométrico y alineadas en estilo a las originales, como se observa en Figura 4.2a y Figura 4.2b, no logró producir ninguna imagen que fuera verdaderamente similar a las originales. Además, muchas de las imágenes generadas resultaron ser abstractas, careciendo de coherencia y sin respetar el estilo deseado, como se aprecia en Figura 4.2c.

[1] Usamos `lr=0,005` que es lo recomendado en [2]



(a) (b) (c)

Fig. 4.2: Imágenes generadas utilizando las descripciones de humanos en inglés, agregando la palabra "Geometría".

4.4. LORA

LoRA (Low-Rank Adaptation) [5] es una técnica utilizada para optimizar y adaptar grandes modelos de lenguaje y de generación de imágenes a tareas específicas. Esta técnica fue introducida en 2021 en el contexto de modelos de lenguaje pero ha sido adaptada con éxito a modelos de visión y difusión para personalizar su comportamiento sin necesidad de reentrenar el modelo completo. En lugar de ajustar todos los parámetros del modelo, LoRA introduce matrices de bajo rango en ciertas capas de la red para adaptar el modelo de manera eficiente sin alterar su estructura completa. Esta técnica reduce considerablemente el costo computacional y de almacenamiento del proceso de fine-tuning, lo cual es clave al trabajar con modelos extremadamente grandes ya que los recursos son en general limitados.

El enfoque de LoRA se basa en la idea de que las actualizaciones necesarias para adaptar un modelo preentrenado a una nueva tarea suelen residir en un subespacio de baja dimensionalidad por lo que lo podemos generar con una matriz de pequeño rango. En lugar de modificar directamente los pesos del modelo se introducen matrices de bajo rango que capturan estas actualizaciones. Estas matrices componen lo que se denomina *adapter* y son entrenadas mientras los pesos originales permanecen congelados, lo que reduce significativamente el costo computacional del ajuste.

En términos concretos, durante el fine-tuning con LoRA por cada capa de las que se van a modificar se introducen dos matrices de bajo rango A y B. Donde A es de $r \times d$ y B es de $d \times r$ con r el rango que es un parámetro de esta técnica y d que es la dimensión de la salida de la capa a modificar. (ver Figura 4.3)

Luego, durante el entrenamiento, la salida de la capa se ajusta mediante una actualización de la forma:

$$W_{adaptado} = W_{original} + \alpha \cdot B \cdot A$$

donde $W_{original}$ son los pesos originales y α es un factor de escalado.

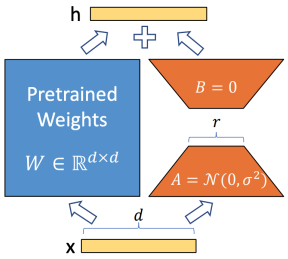


Fig. 4.3: Diagrama de la reparametrización utilizada en LoRA. Se muestra cómo las matrices A y B, de menor rango, se combinan en paralelo con los pesos originales del modelo, permitiendo ajustar únicamente estas matrices sin modificar los parámetros principales.

El mantener los pesos originales congelados ademas de tener ventajas con respecto a los recursos y energia necesarios tiene otros beneficios ya que facilita la reversión al estado preentrenado. Ademas se elimina el impacto del *catastrophic forgetting* donde el modelo pierde todas sus capacidades útiles. Por otro lado los *adapters* entrenados con LoRA se pueden almacenar y cargarlos como modulos al modelo, permitiendo combinar varios *adapters* juntos para darle al modelo mas de una cualidad.

Esta técnica presenta limitaciones, ya que depende en gran medida de las capas seleccionadas del modelo a adaptar, donde se introducen las matrices de bajo rango. Una elección inadecuada puede limitar la capacidad de adaptación del modelo. Además, aunque LoRA es eficaz para ajustes específicos, puede no ser suficiente para tareas que requieren modificaciones más profundas en el modelo.

LoRA fine-tuning ha revolucionado el ajuste de modelos grandes al ofrecer un enfoque eficiente y accesible para personalizar modelos de difusión. Su capacidad para adaptar modelos a nuevos estilos o conceptos sin reentrenar completamente los pesos lo convierte en una herramienta esencial para aplicaciones prácticas y de investigación. Este método equilibra la necesidad de personalización con los recursos computacionales disponibles, haciendo que los modelos de difusión sean más versátiles y permitiendo que mucha gente con recursos computacionales reducidos pueda crear sus propias personalizaciones.

4.5. Descripciones de humanos en ingles como input de Stable Diffusion finetuneado con LORA

Con la motivación de explorar otra técnica de *alignment* para conseguir generar imágenes más geométricas y del estilo deseado se realizaron varios experimentos utilizando *LoRA fine-tuning* en Stable Diffusion v1.5. Este enfoque se aplicó a imágenes de composiciones geométricas generadas de manera sintética. Durante estos experimentos, se probaron diferentes valores de rango con el objetivo de identificar el más adecuado para nuestro caso. Este parámetro es particularmente significativo, ya que no afecta la estabilidad del entrenamiento, aunque puede influir en el tiempo y los recursos necesarios.

Para el fine-tuning se utilizó Diffusers, un paquete de HuggingFace. Así se puedo realizar el entrenamiento de manera sencilla con distintos scripts en Python. Además, para el proceso de *fine-tuning*, fue necesario generar descripciones de cada una de las imágenes de los datasets. Para esto, utilizamos BLIP [7], un modelo que sirve para hacer captioning de imágenes.

Se crearon varios datasets sintéticos de composiciones geométricas. Algunos contenían figuras individuales, como triángulos, mientras que otros incluían composiciones más complejas con varias figuras. También se consideraron datasets con alta variabilidad entre las imágenes y otros más uniformes.

Aunque se realizaron múltiples experimentos, a continuación se presentan los detalles de los más relevantes, que serán incluidos en la evaluación de la fase 3 del experimento.

4.5.1. Experimento 4: Stable Difussion con LoRA fine-tuning con imágenes de triángulos

Se realizo *fine-tuning* con un dataset sintético de composiciones que solo incluían triángulos con el objetivo de dar una ventaja a la familia 4 de las imágenes originales. Se utilizaron 45 imágenes y se definió rango 4. Como se puede ver en Figura 4.4 las imágenes generadas fueron mas alineadas con el estilo deseado en contraste a los experimentos anteriores. Luego en la Capítulo 6 se muestran imágenes de este experimento que resultan muy similares a las originales. En este experimento se destacó de manera más notable cómo todas las imágenes generadas respetaban fielmente el estilo deseado.

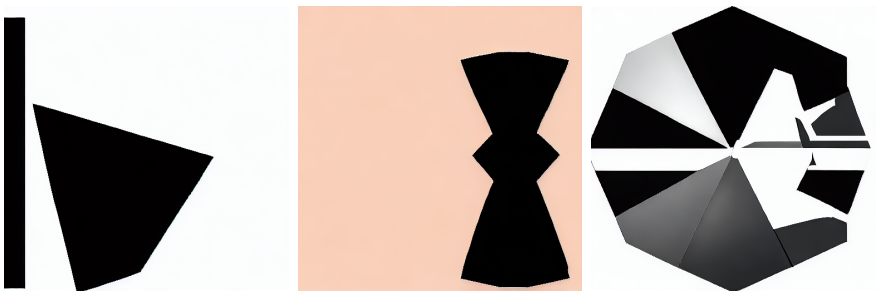


Fig. 4.4: Imágenes generadas utilizando las descripciones humanas en inglés con Stable Diffusion usando LoRA *fine-tuning* generado con triángulos y `rank=4`

4.5.2. Experimento 5: Stable Diffusion con LoRA fine-tuning con imágenes geométricas y rango 4

Se realizo *fine-tuning* con un dataset sintético de composiciones de figuras geométricas variadas. Se utilizaron 97 imágenes y se definió rango 4. Nuevamente, se puede observar que las imágenes generadas en este experimento respetan mucho más el estilo 2D y el formato de dibujo en blanco y negro deseado, como se muestra en Figura 4.5.

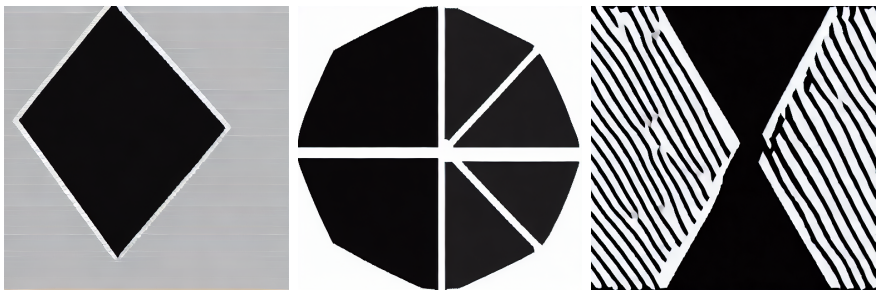


Fig. 4.5: Imágenes generadas utilizando las descripciones humanas en inglés con Stable Diffusion usando LoRA fine-tuning generado con diversas figuras geométricas y `rank=4`

4.5.3. Experimento 6: Experimento 5: Stable Diffusion con LoRA fine-tuning con imágenes geométricas y rango 32

Se realizo *fine-tuning* con un dataset sintético de composiciones de figuras geométricas variadas. Se utilizaron 228 imágenes y se definió rango 32. Además, las descripciones de las imágenes necesarias para el *fine-tuning* se generaron automáticamente utilizando un *prompt* al crear las imágenes, en lugar de emplear BLIP como se hizo en los experimentos anteriores.

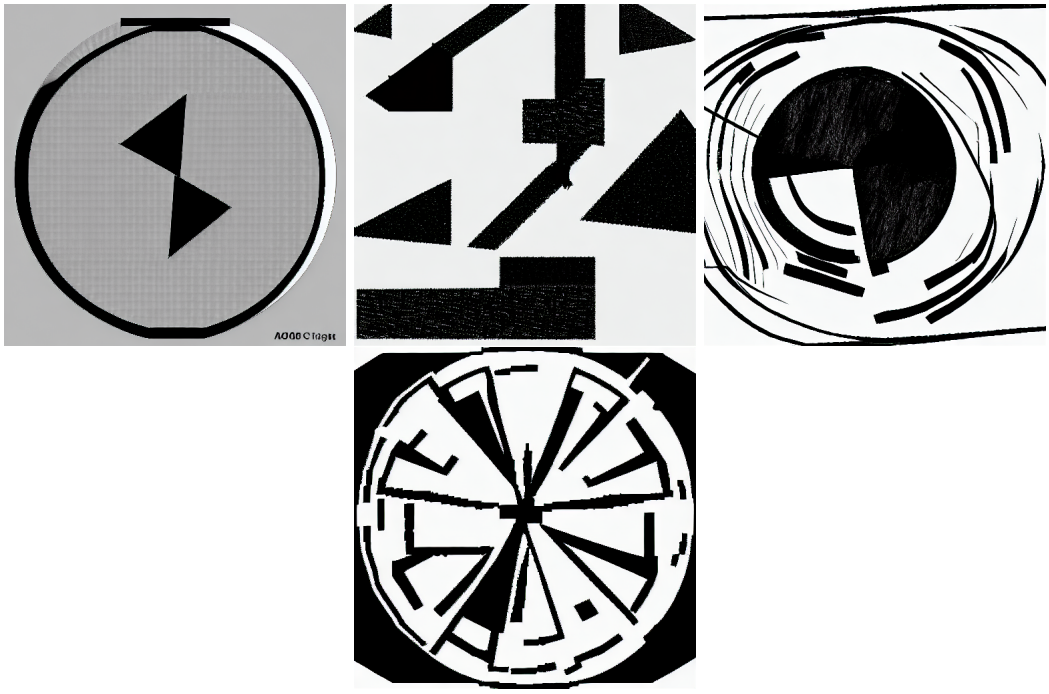


Fig. 4.6: Imágenes generadas utilizando las descripciones humanas en inglés con Stable Diffusion usando LoRA fine-tuning generado con diversas figuras geométricas y `rank=32`

4.6. Prompt tuning

El *prompt tuning* es una técnica utilizada en el ámbito de la generación de imágenes mediante modelos de inteligencia artificial para ajustar y optimizar las instrucciones textuales que se proporcionan al modelo. Su objetivo principal es lograr que las imágenes generadas mantengan una coherencia estilística o respondan a criterios específicos de diseño y temática. Este proceso implica la redacción cuidadosa y refinada de los *prompts*, considerando aspectos como el vocabulario empleado, la estructura de las descripciones y los detalles que se incluyen en ellas.

En primer lugar, el vocabulario es un elemento esencial en el *prompt tuning*, ya que las palabras elegidas influyen en el condicionamiento que recibe el modelo. Por ejemplo, términos como “surrealista” o “realista” ayudan a guiar al sistema hacia un estilo particular. Asimismo, el uso consistente de un léxico asociado a colores, texturas, formas y estilos visuales refuerza la homogeneidad en los resultados.

La estructura de los *prompts* también desempeña un papel crucial. Al utilizar una estructura fija en las descripciones, se establece un patrón que facilita que las imágenes

generadas compartan características comunes. Un enfoque típico puede consistir en describir primero el contexto general, seguido de los elementos principales y, finalmente, los aspectos estilísticos o técnicos, como la iluminación o la composición.

Por último, el nivel de detalle en los *prompts* influye directamente en la especificidad de las imágenes generadas. Incluir descripciones claras y completas permite que el modelo produzca resultados que se alineen con las expectativas del usuario, mientras que la omisión de detalles puede llevar a interpretaciones más abstractas o aleatorias. Este equilibrio entre precisión y flexibilidad es clave para adaptar los prompts a diferentes necesidades creativas.

El *prompt tuning*, al combinar estos elementos, se posiciona como una herramienta fundamental para el control creativo de los modelos generativos. Su implementación no solo permite obtener resultados más consistentes y estilísticamente coherentes, sino que también optimiza el proceso creativo al reducir la cantidad de iteraciones necesarias para alcanzar el objetivo deseado.

4.7. Experimento 7: Prompt tuning sobre DALL·E 3 con las descripciones en inglés

Dado que el entendimiento del lenguaje de Stable Diffusion v1.5 no era suficientemente robusto para interpretar con precisión las descripciones humanas y generar imágenes alineadas con las expectativas, se optó por realizar el experimento de generación de imágenes utilizando otro modelo, DALL·E 3. Este modelo, más reciente y desarrollado con un enfoque más avanzado para la interpretación del lenguaje natural, mostró una mayor capacidad para comprender las instrucciones y producir resultados más consistentes con lo solicitado.

Como parte de la estrategia para mejorar la alineación entre los *prompts* y las imágenes generadas, se implementó una técnica de *aligment* simple pero efectiva: se añadió un prefijo estándar a todos los *prompts*, especificando explícitamente “Genera un dibujo digital de”. Este prefijo proporcionó una guía clara que estableció el contexto creativo deseado, reduciendo la ambigüedad y facilitando que el modelo interpretara las instrucciones de manera más consistente. Este enfoque no solo permitió obtener imágenes más acordes con las expectativas, sino que también garantizó una mayor uniformidad en el estilo de las imágenes generadas.

Las imágenes generadas lograron respetar notablemente el estilo propuesto de dibujo digital. Además, muchas de ellas incorporaron elementos geométricos, alineándose con los objetivos planteados. Esto queda evidenciado en las imágenes presentadas en la Figura 4.7.

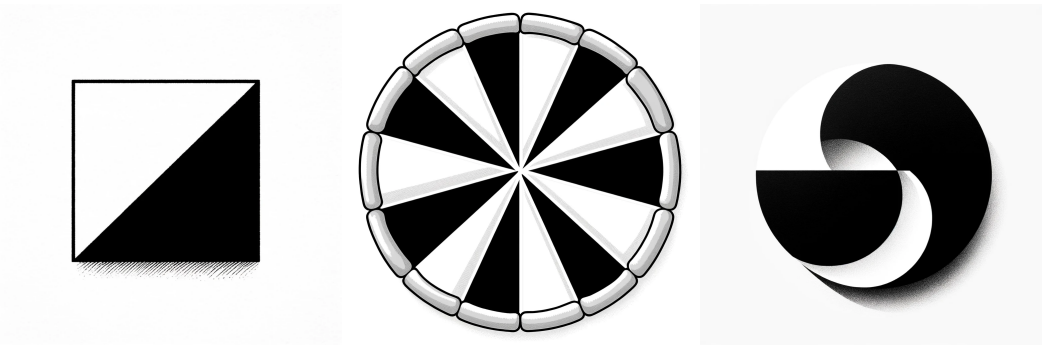


Fig. 4.7: Imágenes generadas por DALLE3 en Sección 4.7 utilizando la técnica de *prompt tuning*.

5. APLICACIÓN WEB PARA LA EVALUACIÓN CROWDSOURCED DE LOS EXPERIMENTOS

Para realizar la evaluación *crowdsourced* correspondiente a la fase 3 del experimento, se desarrolló una aplicación web. El propósito de esta plataforma es evaluar la calidad de las imágenes generadas en los distintos experimentos de generación de imágenes con IA en comparación con los dibujos producidos durante el experimento descrito en [8]. Esta herramienta permite recopilar y analizar opiniones de múltiples usuarios para establecer una evaluación objetiva y comparativa de los resultados obtenidos.

La app web sigue una arquitectura de *serverless app* donde el backend es quien se ocupa de gestionar la base de datos de la aplicación. Esta no depende de servidores tradicionales que se mantengan de forma constante, sino de servicios gestionados que ejecutan funciones en respuesta a eventos. En lugar de mantener un servidor activo en todo momento, las operaciones del backend se ejecutan bajo demanda. La aplicación consta de dos componentes: un cliente, que gestiona la interfaz gráfica y envía los datos ingresados por los participantes, y un backend, que utiliza una API para almacenar dichos datos en la base de datos. En la API cada ruta es manejada de manera independiente de forma que cada vez que se realiza una solicitud se establece la conexión a la base de datos y se ejecuta una operación. Esto nos permitió que tanto el cliente como la API están hosteados en Vercel. Los datos se guardan en una base de datos no estructurada hosteada en Atlas MongoDB. El codigo se encuentra disponible en Experiment Web Page Repo.

El flujo de un participante en el experimento comienza en una página inicial, donde debe ingresar su número de teléfono (ver Figura 5.1a). Este dato se utiliza para generar un identificador único mediante un *hash*, garantizando la anonimidad del participante a lo largo del estudio.

Una vez registrado, el participante accede a un tutorial que explica en detalle el funcionamiento del experimento. En esta etapa, se presentan las instrucciones generales junto con un par de imágenes de prueba para evaluar. Esta evaluación preliminar permite que el participante se familiarice con el formato y las expectativas del experimento. La interfaz del tutorial se puede observar en Figura 5.1b.

Finalizado el tutorial, el participante avanza hacia el experimento real, donde se le presentan pares de imágenes para evaluar según la consigna de *"Suponiendo que la imagen de la derecha fue generada a partir de una descripción basada en la imagen de la izquierda, ¿qué tan bien consideras que salio el proceso?"*. Estas evaluaciones se realizan de manera secuencial, y cada decisión se registra junto con datos relevantes, como los timestamps y las imágenes evaluadas. La vista de la página durante esta fase del experimento se ilustra en Figura 5.1c. El flujo finaliza cuando el participante completa todas las evaluaciones asignadas.

Dado que la página está diseñada para un experimento con humanos, es esencial registrar todos los datos de la actividad de los participantes para garantizar la trazabilidad y permitir la reconstrucción detallada de lo que hizo cada uno durante el experimento. Este enfoque nos permite analizar el comportamiento individual y colectivo, identificar patrones, y asegurar la integridad de los datos para futuros análisis.

Para lograr este objetivo, se almacena toda la información relevante sobre las interacciones que los participantes tienen con la página. En particular, se registran los siguientes

datos: el *hash* del celular del participante como identificador, el *timestamp* del *login*, el *timestamp* correspondiente a la salida del tutorial y el inicio del experimento, el número de serie asignado y un array con los datos de las calificaciones. De cada calificación se almacenan los datos de la imagen original, la imagen generada y el *timestamp* en el que se envía.

Además, esta metodología asegura la reproducibilidad del experimento, ya que contar con un registro completo de las actividades de los participantes permite validar los resultados y detectar errores.

(a) Pantalla de inicio de sesión, donde los participantes ingresan para acceder al experimento.

(b) Vista del tutorial interactivo que explica a los participantes las instrucciones y pasos necesarios para completar el experimento.

(c) Interfaz principal del experimento, donde los participantes evalúan las imágenes generadas.

Fig. 5.1: Vistas de la aplicación web para la evaluación *crowdsourced*.

En el diseño del experimento, se identificó la necesidad de establecer el orden en que se presentan las imágenes a cada participante. Dado que se preveía que algunos participantes podrían no completar el experimento, se implementó un sistema para determinar el orden de presentación de manera equitativa y controlada.

Se generó aleatoriamente un conjunto de series, cada una con un orden diferente de las imágenes a evaluar. A cada participante se le asigna una de estas series. Para gestionar esta asignación, se mantiene en la base de datos un índice que indica la última serie asignada. Cuando un nuevo participante accede al experimento, se le otorga la siguiente serie en el orden y el índice se incrementa en la base de datos. Al alcanzar la última serie, el sistema reinicia el índice y vuelve a asignar desde la primera serie.

Este método garantiza una distribución uniforme de las series entre los participantes y permite que el experimento se realice de manera consistente, incluso si algunos participantes no completan la evaluación de todos los pares de imágenes.

6. EVALUACIÓN CROWDSOURCED: PRUEBA PILOTO

6.1. Experimento piloto

Con el objetivo de probar la aplicación web y obtener resultados preliminares, se decidió llevar a cabo una prueba piloto de la evaluación *crowdsourced*. Para este propósito, se restringió el conjunto de imágenes originales, buscando asegurar que cada imagen generada obtuviera una cantidad considerable de calificaciones sin necesitar una gran cantidad de personas. Durante este piloto solo se evalúan imágenes pertenecientes a la familia 4 (ver Figura 2.2).

En este piloto, se incluyeron los siguientes métodos de generación:

- Experimento 1: Descripciones de humanos en español como entrada para Stable Diffusion
- Experimento 2: Descripciones de humanos en inglés como entrada para Stable Diffusion
- Experimento 3: Descripciones de humanos en en inglés como entrada para Stable Diffusion con Textual Inversion generado con figuras geométricas
- Experimento 4: Stable Difussion con LoRA fine-tuning con imágenes de triángulos
- Experimento 5: Stable Diffusion con LoRA fine-tuning con imágenes geométricas y rango 4
- Experimento 7: Prompt tuning sobre DALL·E 3 con las descripciones en inglés
- Experimento Original con Humanos

En total, se seleccionaron 8 imágenes generadas por cada uno de los 7 experimentos para cada una de las 4 imágenes originales, resultando en un total de 224 imágenes que cada participante debía calificar. El criterio de selección consistió, para cada imagen original y experimento, en incluir cuatro imágenes consideradas de buena calidad y cuatro de mala calidad, según un criterio propio. Las imágenes se presentaron a los participantes en un orden aleatorio. Para garantizar un balance en las calificaciones, incluso si algunos participantes no completaban el experimento, se diseñaron 10 series con diferentes órdenes de presentación de las imágenes. Siguiendo un enfoque similar al método *round robin*, a cada participante se le asignó uno de estos órdenes de manera cíclica, asegurando una distribución uniforme de las series entre todos los participantes.

El experimento se llevó a cabo durante el mes de febrero de 2024 y se compartió entre estudiantes de la facultad, logrando la participación de 95 personas. En total, se recopilaron 15,133 calificaciones, ya que solo el 50 % de los participantes completó el experimento, como se muestra en la Figura 6.2. Sin embargo, la distribución de las calificaciones entre los distintos métodos de generación fue bastante uniforme, lo que asegura que todas las imágenes tienen cantidades similares de calificaciones. Tal como se preveía, el tiempo promedio que tomó a cada participante evaluar un par de imágenes fue de aproximadamente 5 segundos, y el tiempo total de permanencia en el experimento se concentro entre 10 y

15 minutos, como se puede ver en Figura 6.1. No obstante, varios participantes señalaron que la cantidad total de imágenes a evaluar era excesiva, lo que podría haber afectado la completitud de sus tareas o su nivel de atención durante el experimento.

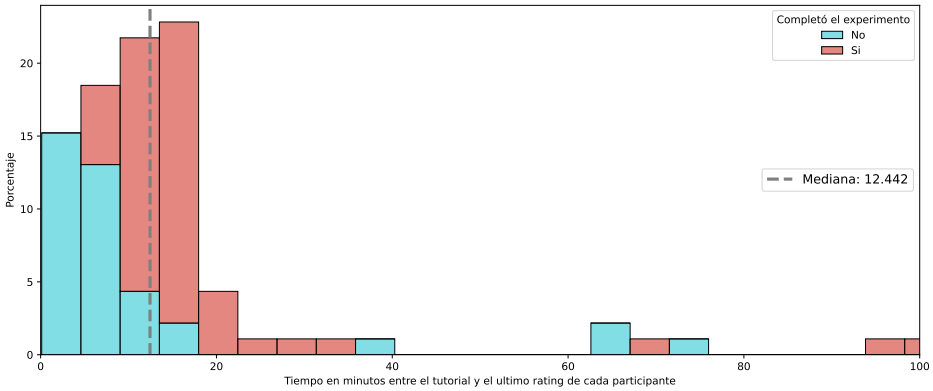


Fig. 6.1: Histograma que muestra la cantidad de tiempo que los participantes permanecieron en el experimento, desglosado según si completaron o no el experimento.

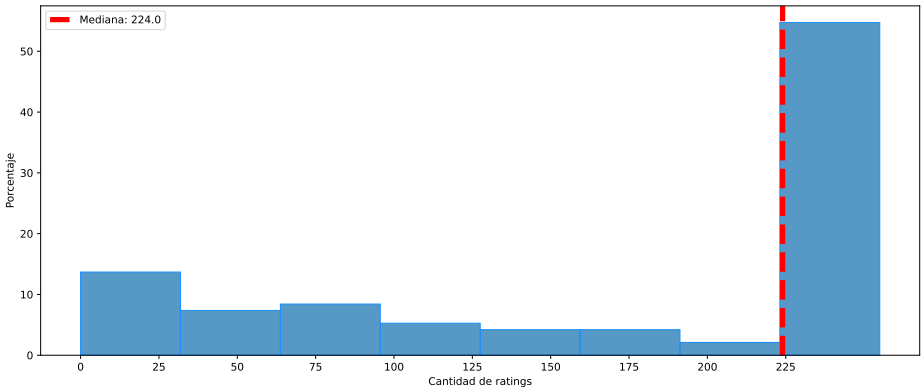


Fig. 6.2: Histograma de la cantidad de pares evaluados por cada participante.

6.2. Resultados piloto

Frente a los resultados del piloto, se confirmó la efectividad de la estrategia de balanceo para garantizar una distribución uniforme de las calificaciones, incluso en los casos en los que los participantes abandonaron el experimento prematuramente. Además, estos resultados permiten ajustar el diseño experimental para iteraciones futuras. El enfoque se centra en mejorar la experiencia de los participantes y optimizar el número de imágenes evaluadas. El objetivo es lograr un equilibrio entre la calidad de los datos obtenidos y la carga cognitiva de los evaluadores. En general, las estrategias implementadas lograron preservar el balance y la representatividad de las evaluaciones, proporcionando datos confiables para el análisis de los métodos de generación utilizados.

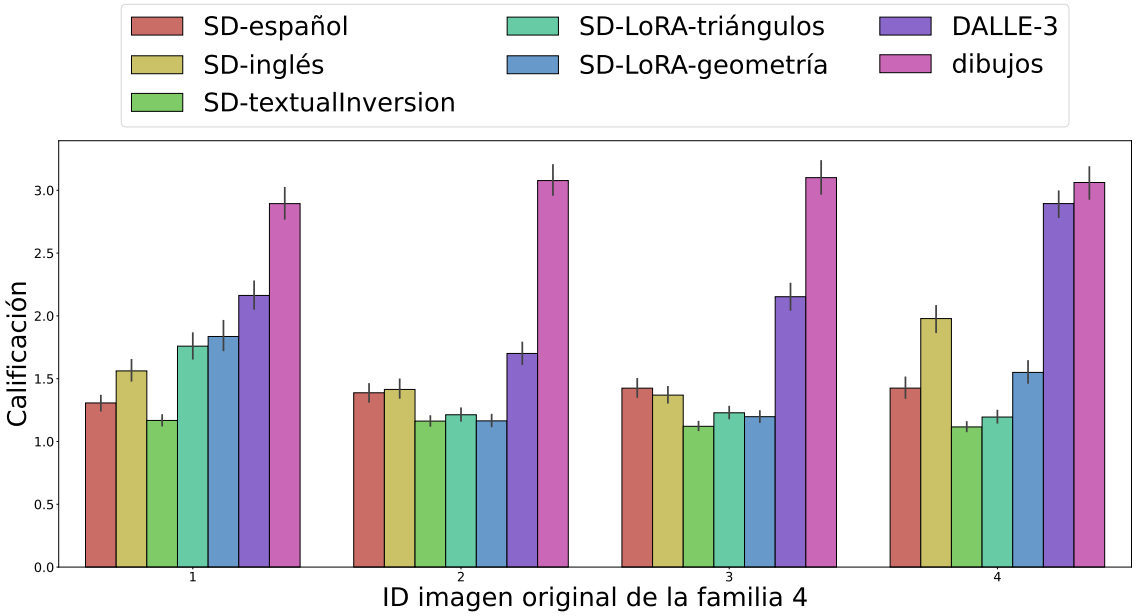
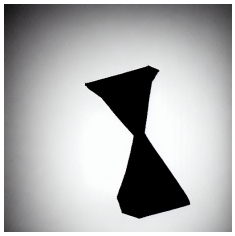
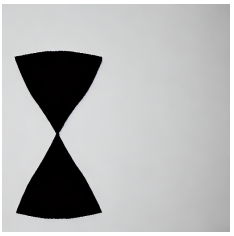


Fig. 6.3: Calificación promedio de las imágenes generadas en cada experimento dividido por el ID de la imagen original dentro de la familia 4.

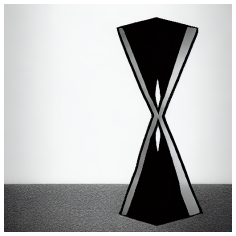
Tras la exploración de los datos generados durante el piloto, se identificaron varias conclusiones relevantes. Tal como se preveía, los humanos obtuvieron las mejores calificaciones en comparación con todos los otros métodos de generación basados en modelos generativos (ver Figura 6.3). Además, las imágenes generadas por DALLE-3 en Sección 4.7 fueron las mejores en comparación de todos los demás métodos de generación evaluados. Aunque esto podría parecer sorprendente dado que el método de *alignment* utilizado con DALLE-3 fue muy simple, este resultado tiene sentido ya que este modelo es significativamente más reciente que Stable Diffusion v1.5. Resulta interesante ademas que las imágenes de los experimentos donde se utilizo *fine-tuning* con LoRA no fueron superiores a los experimentos con los modelos base. Esto puede atribuirse a que los experimentos con LoRA que incluimos en este piloto fueron con un rango muy pequeño, de todas formas la imagen con mejor puntuación fue generada por este método y es efectivamente muy parecida a la original (ver Figura 6.4a). Inicialmente, se esperaba que el uso de descripciones directamente en español produjera resultados significativamente inferiores en comparación con las descripciones en inglés. Sin embargo, esto no fue así, ya que las imágenes generadas a partir de Sección 3.3 y Sección 3.2 obtuvieron calificaciones similares en promedio. No obstante, esto no implica que el modelo Stable Diffusion tenga la misma capacidad de comprensión del texto en español que en inglés. Más bien, se relaciona con el hecho de que las imágenes generadas por ambos métodos están, en términos de estilo, notablemente alejadas de las originales.



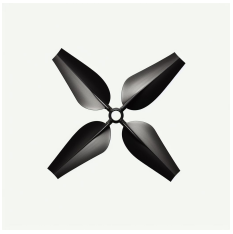
(a) Generada por el experimento Subsección 4.5.1.



(b) Generada por el experimento Subsección 4.5.2.



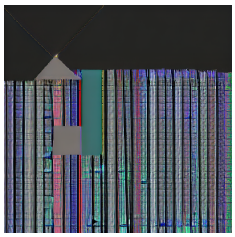
(c) Generada por el experimento Subsección 4.5.2.



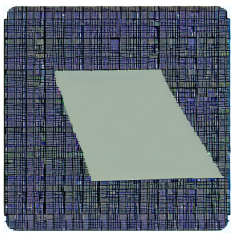
(d) Generada por el experimento Sección 4.7.

(e) Generada por el experimento Sección 4.7.

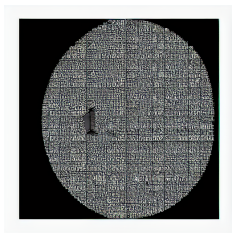
Fig. 6.4: 5 imágenes con mayor calificación promedio de las evaluadas durante el piloto.



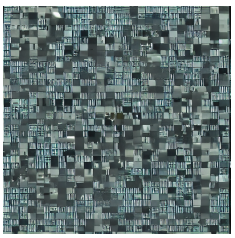
(a) Generada por el experimento Sección 4.3.



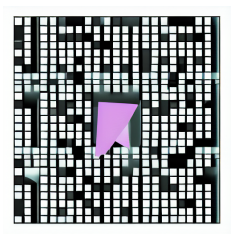
(b) Generada por el experimento Sección 4.3.



(c) Generada por el experimento Sección 4.3.



(d) Generada por el experimento Sección 4.3.



(e) Generada por el experimento Sección 4.3.

Fig. 6.5: 5 imágenes con menor calificación promedio de las evaluadas durante el piloto.

El análisis del experimento piloto permitió recopilar diversas observaciones que profundizan la comprensión de los resultados obtenidos en los distintos experimentos de generación, estableciendo una base sólida para orientar la fase final de evaluación de las

imágenes.

7. EVALUACIÓN CROWDSOURCED FINAL

Para la evaluación final, se diseñó cuidadosamente el procedimiento con el objetivo de determinar la cantidad necesaria de participantes y el número de imágenes a evaluar. Este diseño inicial permitió establecer los parámetros necesarios para garantizar que los resultados fueran estadísticamente significativos y que cada imagen generada recibiera un número adecuado de calificaciones para su análisis comparativo. La planificación incluyó la estimación de los recursos necesarios y la implementación de estrategias para maximizar la calidad y balance de los datos recopilados.

Frente a los resultados del piloto se reemplazo uno de los experimentos con LoRA por otro usando un rango mayor. Finalmente, se incluyeron los siguientes métodos de generación:

- Experimento 1: Descripciones de humanos en español como entrada para Stable Diffusion
- Experimento 2: Descripciones de humanos en inglés como entrada para Stable Diffusion
- Experimento 3: Descripciones de humanos en en inglés como entrada para Stable Diffusion con Textual Inversion generado con figuras geométricas
- Experimento 4: Stable Difussion con LoRA fine-tuning con imágenes de triángulos
- Experimento 6: Experimento 5: Stable Diffusion con LoRA fine-tuning con imágenes geométricas y rango 32
- Experimento 7: Prompt tuning sobre DALL·E 3 con las descripciones en inglés
- Experimento Original con Humanos

Se seleccionaron 896 imágenes para la evaluación. Por cada uno de los 7 métodos de generación y por cada una de las 16 figuras originales se tomaron 8 imágenes. La selección de imágenes se realizó de manera manual bajo el criterio de que, para un experimento y una figura original específicos, la mitad de las imágenes fueran consideradas de buena calidad y la otra mitad de mala calidad. Es decir, por ejemplo, para DALLE-3 y la figura 3 de la familia 2 tomamos 4 imágenes buenas y 4 malas. Debido a que la cantidad de imágenes asignadas a cada participante durante el piloto fue excesiva, en esta evaluación se decidió reducir la cantidad de imágenes evaluadas por participante a 152. Dado que se quieren obtener 90 calificaciones por imagen se necesitan 538 personas que completen el experimento.

A diferencia de la evaluación realizada en el piloto, donde cada participante evaluaba todas las imágenes, en esta ocasión fue necesario decidir qué imágenes serían asignadas a cada participante. Se estableció como criterio que cada participante debía evaluar imágenes generadas por todos los métodos y correspondientes a las 16 figuras originales. Para ello, se diseñaron 6 series, cada una de las cuales incluye entre 5 y 6 imágenes asociadas a cada una de las 16 figuras originales y a los 7 métodos de generación. Finalmente, se generaron 10 permutaciones de cada serie para garantizar la variedad en el orden en que las imágenes son presentadas a los participantes.

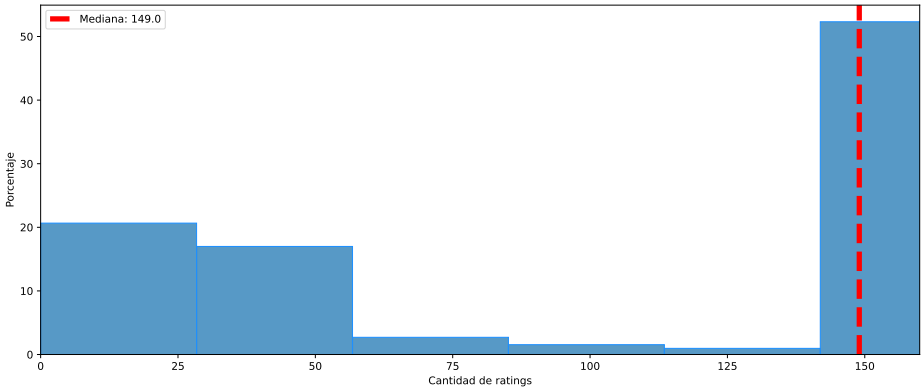


Fig. 7.1: Histograma de la cantidad de pares evaluados por cada participante.

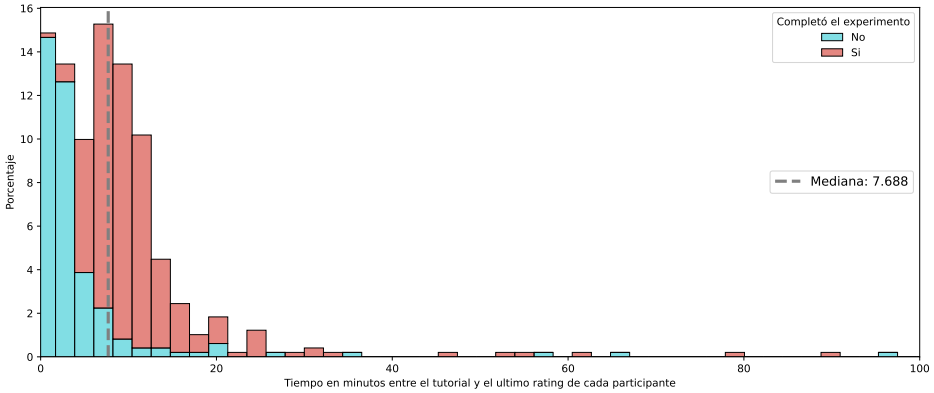


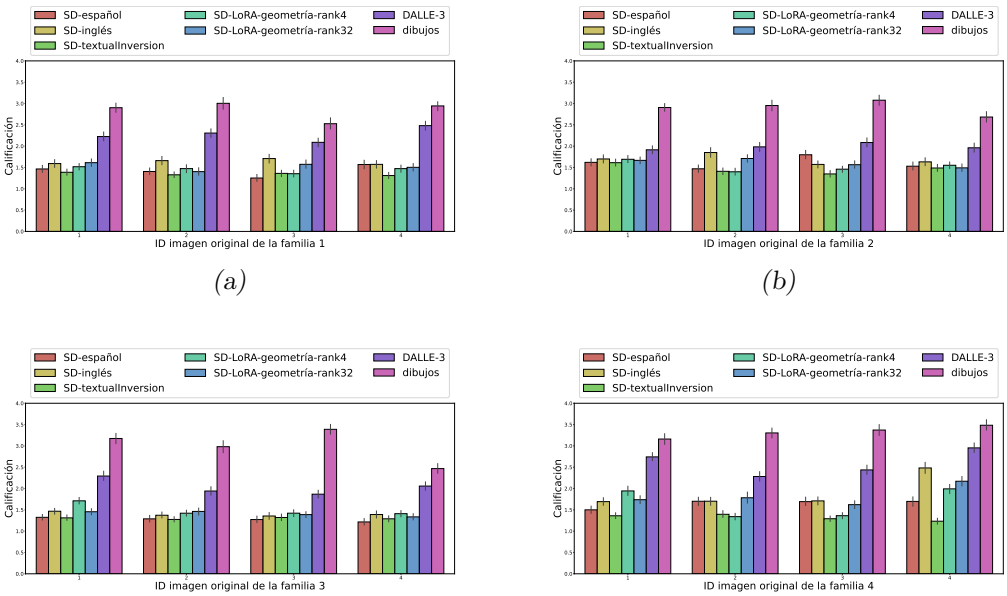
Fig. 7.2: Histograma que muestra la cantidad de tiempo que los participantes permanecieron en el experimento, desglosado según si completaron o no el experimento

7.1. Resultados experimento

El experimento se compartió entre estudiantes de la facultad, redes sociales y personas conocidas, logrando la participación de 524 personas. Si bien esta no es la cantidad deseada esta muy cerca y requirió de mucho esfuerzo poder lograr este nivel de participación. En total, se obtuvieron 53,114 calificaciones, dado que de igual forma que ocurrió durante el piloto muchos participantes no completaron el experimento. De todas maneras mas de la mitad de los participantes completaron el experimento como se puede ver en Figura 7.1. Al disminuir la cantidad de imágenes el tiempo total de permanencia en el experimento se concentro entre 7 y 10 minutos, como se puede ver en Figura 7.2.

Al igual que en el experimento piloto, los humanos continuaron obteniendo los mejores resultados en la generación de imágenes, demostrando una clara superioridad en comparación con los modelos generativos para todas las familias de imágenes como se puede ver en Figura 7.3a, Figura 7.3b, Figura 7.3c y Figura 7.3d. Sin embargo, se destacó también el rendimiento sobresaliente de DALLE·3, que logró generar imágenes de alta calidad, superando las expectativas en cuanto a coherencia y alineación con las descripcio-

nes textuales. Este resultado resalta no solo la capacidad de los humanos para interpretar contextos complejos, sino también el avance significativo de modelos generativos como DALLE·3, que continúan acercándose a los niveles de precisión y creatividad alcanzados por los participantes humanos. Además, resulta notable que las imágenes pertenecientes a la familia 4 obtuvieron las calificaciones más altas, lo que sugiere que este tipo de figuras podrían ser más fáciles de generar o que las descripciones asociadas a ellas son de mayor calidad.



(a) *(b)* *(c)* *(d)*

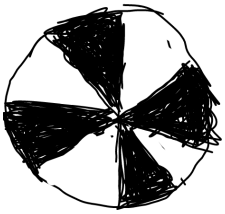
Fig. 7.3: Calificaciones promedio obtenidas por cada experimento, desglosadas según la familia de las imágenes originales.

A diferencia de lo que observamos en el piloto donde las mejores imágenes correspondían a DALLE-3 y a LoRA, en esta nueva evaluación las mejores imágenes fueron todos dibujos como se puede ver en Figura 7.4. Ademas también en contraste con lo que paso en el piloto, las peores imágenes no corresponden todas al experimento con Textual Inversion sino que corresponden al experimento de Stable Diffusion en español Sección 3.2, esto lo podemos ver en Figura 7.5. La performance de LoRA resultó decepcionante, ya que, aunque se esperaba que un aumento en el rango mejorara las calificaciones promedio de las imágenes, esto no ocurrió.

(a) Generada por el experimento Sección 2.1.

(b) Generada por el experimento Sección 2.1.

(c) Generada por el experimento Sección 2.1.

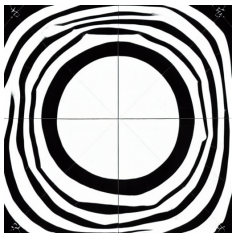


(d) Generada por el experimento Sección 2.1.

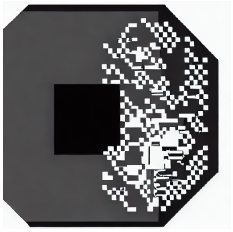
(e) Generada por el experimento Sección 2.1.

Fig. 7.4: Imágenes con mayor calificación promedio durante la evaluación final.

(a) Generada por el experimento Sección 3.2.

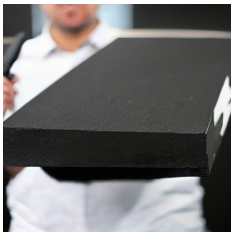


(b) Generada por el experimento Sección 3.3.



(c) Generada por el experimento Subsección 4.5.2.

(d) Generada por el experimento Sección 3.2.



(e) Generada por el experimento Sección 3.2.

Fig. 7.5: Imágenes con peor calificación promedio durante la evaluación final.

8. CONCLUSIONES

El análisis de los distintos experimentos de generación de imágenes realizados durante esta tesis permitió evaluar las capacidades de estos modelos en contraste con las de los humanos en esta tarea particular. Gracias a esto se llegan a las siguientes conclusiones:

- Los resultados mostraron que los dibujos generados por humanos obtuvieron las calificaciones más altas durante la fase 3 del experimento, confirmando su superioridad en esta tarea específica. Esto destaca las limitaciones actuales de los modelos generativos para replicar composiciones visuales con la precisión y el contexto que los humanos son capaces de interpretar y reproducir. Por otro lado es cierto que los humanos contaban con la ventaja que el estilo blanco y negro era algo que estaba marcado por la aplicación en que se recolectaron los dibujos. Es decir que aunque en las descripciones no se pida explícitamente respetar los colores o el estilo igual lo hacían. De todas formas también tiene sentido que los humanos sean muy buenos entendiendo texto generado por otro humanos mientras que los modelos generativos sean peores en este aspecto.

- Aunque se esperaba que las descripciones en español produjeran resultados significativamente peores que aquellas en inglés, las calificaciones promedio entre ambos idiomas fueron similares. Sin embargo, este resultado no implica que el modelo interprete ambos idiomas con la misma eficacia, sino que refleja el hecho de que las imágenes generadas, independientemente del idioma, estaban estilísticamente muy alejadas de las originales.

- DALLE-3 se destaco como el modelo generativo con mejores calificaciones promedio, a pesar de utilizar un método de *alignment* relativamente simple como el *prompt tuning*. Esto resalta las capacidades avanzadas del modelo para interpretar texto y generar imágenes estilísticamente coherentes. Igualmente el *prompt tuning* que se utiliza hace que se defina bien el estilo de las imágenes a generar por lo que eso acerca a las condiciones que tenían los humanos al dibujar. Por otro lado, Stable Diffusion, aunque limitado en algunos aspectos, demostró ser una opción que genero imágenes muy parecidas a las originales aunque tambien muchas muy distintas por lo que en promedio no obtuvo las mejores calificaciones. Las imágenes generadas utilizando *fine-tuning* con Stable Diffusion, aunque lograron mejores resultados en comparación con otras técnicas, no alcanzaron la performance esperada. Esto podría estar relacionado tanto con las limitaciones conocidas de Stable Diffusion para interpretar composicionalidad, como con la posible falta de optimización en los parámetros utilizados para entrenar los LoRA. Este método resulta particularmente difícil de optimizar debido al alto costo asociado a la realización de las evaluaciones necesarias.

- Tanto el diseño de la plataforma para la evaluación *crowdsourced* como las técnicas utilizadas para garantizar el balance de las calificaciones fueron exitosas. Esto permitió obtener resultados representativos.

En conclusión, los resultados de esta tesis confirman que, aunque los modelos generativos actuales son herramientas poderosas y efectivas para la creación de imágenes, todavía presentan limitaciones significativas, particularmente en tareas que exigen un alto nivel de comprensión semántica y contexto. Además, aunque la tarea diseñada permite evaluar la performance de los modelos, la fase 3 de evaluación con humanos representa un proceso costoso en términos de recursos y tiempo.

9. FUTURAS LINEAS DE INVESTIGACIÓN

Los resultados obtenidos en esta tesis abren varias oportunidades para explorar nuevas direcciones de investigación, con el objetivo de seguir avanzando en la comprensión y optimización de los modelos generativos de texto a imagen para esta tarea particular. Algunas de las proyecciones que quedaron pendientes o resultan de interes son:

- **Reescritura de las descripciones de humanos:** Se observó que las descripciones utilizadas no siempre producían imágenes alineadas con las expectativas tanto por falta de detalles importantes sobre el estilo de las imágenes a generar o porque las descripciones era demasiado metafóricas. Una de las lineas a explorar seria utilizar un LLM para que convierta estas descripciones en *prompts* similares a los que utiliza la comunidad. Dado que el *prompting* en general es muy distinto a simplemente describir lo que se quiere generar tiene sentido que las imágenes generadas mejoren al utilizar las descripciones modificadas para que se alineen con el formato de *prompt*.

- **Explorar otras técnicas de aligment:** Dado que tanto LoRA como Textual Inversion no fueron técnicas muy efectivas para generar imágenes para este experimento, se pueden explorar otras técnicas como DreamSync [13] que mejoren esto.

- **Explorar otros modelos:** Los modelos de texto a imagen están mejorando todos los días. En esta tesis nos centramos en usar Stable Diffusion v1.5 dado que es Open-Source y es un modelo rápido para la generación de imágenes con el hardware disponible. Pero actualmente hay otros modelos mas grandes o que no son de código abierto como Midjourney y Stable Diffusion 3 que pueden tener una mejor performance. Aunque no es claro si es suficiente para sobrepasar la performance de los humanos en esta tarea particular.

- **Humanos dibujando prompts:** Dado que las descripciones de los humanos fueron una gran limitación para los modelos generativos, se puede analizar cual es la performance de los humanos si en vez de leer descripciones de humanos leen *prompts* generados por modelos de captioning y LLMs y se basan en eso para dibujar. Esta es una linea de investigación interesante ya que primero hay que encontrar un formato de *prompt* que sirva para generar estas imágenes geométricas, lo cual es un desafío dado que los modelos todavía tienen problemas con la composición.

Bibliografía

[1] Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. A maximum entropy approach to natural language processing. Comput. Linguist., 22(1):39–71, March 1996. 1

[2] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. 19, 21

[3] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 9

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 9, 11

[5] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 22

[6] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. 9

[7] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 6, 23

[8] Matías Lopez-Rosenfeld, Facundo Carrillo, Gerry Garbulsky, Diego Fernandez Slezak, and Mariano Sigman. Quantitative pedagogy: A digital two player game to examine communicative competence. PLOS ONE, 10(11):1–10, 11 2015. 1, 3, 4, 13, 27

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1, 7, 9, 20

[10] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 9

[11] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. 9

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 6, 9, 11

[13] Jiao Sun, Deqing Fu, Yushi Hu, Su Wang, Royi Rassin, Da-Cheng Juan, Dana Alon, Charles Herrmann, Sjoerd van Steenkiste, Ranjay Krishna, and Cyrus Rashtchian. Dreamsync: Aligning text-to-image generation with image understanding feedback, 2023. 43

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. 9

[15] Luis von Ahn and Laura Dabbish. Designing games with a purpose. Communications of the ACM, 51(8):58–67, August 2008. 3

[16] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, 2017. 9