Revista Científica Multidisciplinar G-ner@ndo ISSN: 2806-5905

# Análisis comparativo de algoritmos basados en Inteligencia Artificial para la generación de imágenes a partir de texto.

## Comparative analysis of Artificial Intelligence-based algorithms for text-to-image generation.

*Edison Gabriel Mero Aragundi & Roberth Abel Alcivar Cevallos*.

**PUNTO CIENCIA.**

**Julio - diciembre, V°6 - N°2; 2025**

**Recibido:** 29-09-2025
**Aceptado:** 02-10-2025
**Publicado:** 30-12-2025

**PAIS**

- Ecuador, Manabí
- Ecuador, Manabí

**INSTITUCION**

- Universidad Técnica de Manabí.
- Universidad Técnica de Manabí.

**CORREO:**

✉ emero8062@utm.edu.ec
✉ roberth.alcivar@utm.edu.ec

**ORCID:**

https://orcid.org/0009-0008-5624-4422
https://orcid.org/0000-0001-6282-8493

**FORMATO DE CITA APA.**

*Mero, E. & Alcivar, R.* (2025). Análisis comparativo de algoritmos basados en Inteligencia Artificial para la generación de imágenes a partir de texto. *Revista G-ner@ndo, V°6 (N°2). Pág. 1970 – 1994.*

**Resumen**

El uso de IA para generar imágenes a partir de texto se ha convertido en un área primordial para la automatización de imágenes visuales. En este estudio se compara el desempeño de tres modelos generativos open-Source: BigGAN, BigGAN+CLIP, Stable diffusion, y uno Comercial: Midjourney, usando una metodología mixta. Se utilizaron instrucciones con niveles de complejidad semántica simple, medio y complejo en inglés con el fin de no intervenir en el proceso de entrenamiento de los modelos. La valoración se efectuó en dos fases: una cuantitativa, basada en la Distancia de Inicio de Fréchet (FID), y otra cualitativa, sustentada en una escala Likert de cinco puntos aplicada por expertos en diseño, e inteligencia artificial. Los resultados evidencian diferencias significativas en función de la complejidad textual y los criterios aplicados, aportando elementos metodológicos y analíticos útiles para la evaluación académica, y creativa de modelos generativos.

**Palabras clave:** inteligencia artificial; generación de imágenes; Stable Diffusion; BigGAN; MidJourney.

**Abstract**

The use of AI to generate images from text has become a primary area for the automation of visual imagery. This study compares the performance of three open-source generative models: BigGAN, BigGAN+CLIP, and Stable Diffusion, and one commercial model: Midjourney, using a mixed methodology. Instructions with simple, medium, and complex levels of semantic complexity in English were used to avoid interfering with the training process of the models. The evaluation was conducted in two phases: a quantitative phase, based on the Fréchet Inception Distance (FID), and a qualitative phase, supported by a five-point Likert scale applied by experts in design and artificial intelligence. The results reveal significant differences based on textual complexity and the applied criteria, providing useful methodological and analytical elements for the academic and creative evaluation of generative models.

**Keywords:** artificial intelligence; image generation; Stable Diffusion; BigGAN; MidJourney.

## Introducción

La creación de imágenes a partir de descripciones textuales se ha consolidado como uno de los avances más relevantes de la inteligencia artificial, al integrar procesamiento de lenguaje natural y generación visual mediante redes neuronales profundas. Este enfoque busca imitar la capacidad humana de imaginar escenas a partir de narraciones verbales, transformando el texto en representaciones visuales con sentido (Frolov et al., 2021; Zhang & Tang, 2023).

El desarrollo de este campo está estrechamente vinculado a la evolución de las Redes Generativas Antagónicas (GANs), propuestas inicialmente por Goodfellow et al. (2014). Estas arquitecturas introdujeron un esquema competitivo entre generador y discriminador, permitiendo producir imágenes con realismo creciente. Posteriores mejoras como DCGAN, WGAN, StackGAN y StyleGAN superaron limitaciones de resolución y estabilidad, consolidando el terreno para modelos de mayor escala como BigGAN (Zhang et al., 2017; Salimans et al., 2016; Brock et al., 2019). A la par, diversas revisiones sistemáticas han analizado el rol de estas arquitecturas, destacando tanto sus aplicaciones como sus limitaciones teóricas (Gui et al., 2021; Huang et al., 2020; Singh et al., 2020).

La débil correspondencia entre texto e imagen en los primeros modelos impulsó enfoques híbridos que integraron representaciones lingüísticas y visuales. Ejemplo de ello es BigGAN+CLIP, donde el texto guía la optimización para alinear contenido semántico y atributos visuales (Radford et al., 2021). Posteriormente, los modelos de difusión inauguraron un nuevo paradigma: la eliminación progresiva de ruido en espacios latentes condicionados por texto. Entre ellos, Stable Diffusion (Rombach et al., 2022) destacó por su equilibrio entre calidad estructural, control creativo y carácter abierto, lo que facilitó su rápida adopción en investigación y aplicaciones creativas (Ho et al., 2020; Du et al., 2023).

En paralelo, han emergido propuestas comerciales como MidJourney, cuyo valor radica menos en la transparencia técnica que en su impacto cultural y accesibilidad para usuarios no especializados, consolidándose como herramienta en diseño, publicidad e ilustración (Müller & Lee, 2023; Domínguez & García, 2023). Estudios recientes señalan que su popularidad se asocia a su capacidad de democratizar la creación visual y expandir horizontes expresivos en contextos artísticos (Cedeño & Ruiz, 2023; Sri Krishna, 2022; Oppenlaender, 2022).

Un aspecto central en este ecosistema es la ingeniería de prompts (prompt engineering), entendida como el diseño estratégico de descripciones lingüísticas para guiar el comportamiento de los modelos. Un prompt efectivo incorpora elementos como sujeto, atributos visuales, contexto, estilo y composición, y su variación puede modificar radicalmente los resultados generados (Reynolds & McDonell, 2021; Ramesh et al., 2021). A ello se suman los denominados metaprompts, estructuras diseñadas para inducir estilos narrativos o enfoques cognitivos específicos, reforzando el papel del lenguaje como motor de creatividad algorítmica (Li & Liang, 2021; Yamada & Nakamura, 2024).

El impacto de estas tecnologías se ha extendido más allá del arte digital hacia el diseño conceptual y la co-creación asistida por computadora. Investigaciones en entornos de ingeniería muestran que los modelos T2I pueden potenciar la exploración de alternativas y enriquecer la práctica profesional del diseño (Alcaide-Marzal & Diego-Mas, 2025; Chiou et al., 2023; Zhang & Ortega, 2023). Sin embargo, la introducción de la IA generativa en estos campos plantea retos éticos y metodológicos asociados a la autoría, la transparencia y la agencia algorítmica (Camacho & Paredes, 2023; Broncano, 2024; Gualdoni, 2024; Hernández, 2024; Martín Prada, 2024; Merino, 2024).

En términos de evaluación, persisten limitaciones. Aunque métricas como Inception Score (IS) y Fréchet Inception Distance (FID) son ampliamente empleadas

para cuantificar la calidad generativa, diversos autores advierten que no capturan de manera integral la coherencia semántica entre texto e imagen, lo que dificulta comparaciones objetivas entre modelos (Borji, 2022; Wang, 2023). Ante ello, la incorporación de evaluaciones humanas, mediante escalas perceptivas y análisis cualitativos, constituye un complemento necesario para validar el desempeño de los algoritmos.

En este contexto, el presente estudio compara el rendimiento de cuatro algoritmos representativos —BigGAN, BigGAN+CLIP, Stable Diffusion y MidJourney— bajo distintos niveles de complejidad semántica. Para ello, se aplica una metodología mixta que combina evaluación cuantitativa mediante FID y evaluación cualitativa con juicio experto a través de escala Likert. El propósito es aportar evidencia metodológica que permita valorar simultáneamente la calidad técnica y la percepción humana, ofreciendo insumos relevantes para aplicaciones académicas, creativas e industriales.

## Métodos y Materiales

El estudio se enmarca en una investigación aplicada con enfoque mixto. Por un lado, se utiliza un componente cuantitativo, sustentado en métricas objetivas como el Fréchet Inception Distance (FID), que permiten medir la fidelidad visual y la similitud estructural de las imágenes generadas frente a un conjunto de referencia. Paralelamente, se incorpora un componente cualitativo, basado en la evaluación perceptual humana mediante escalas tipo Likert, lo cual posibilita captar matices estéticos, semánticos y de coherencia que difícilmente pueden ser reducidos a valores numéricos.

La combinación de ambos enfoques ofrece una visión integral del desempeño de los modelos, articulando precisión técnica con interpretación contextual.

**Modelos seleccionados**

Se seleccionaron cuatro algoritmos representativos por su relevancia investigativa y práctica:

- BigGAN: red generativa adversarial a gran escala, entrenada con clases semánticas complejas y reconocida por la nitidez de sus resultados.
- BigGAN+CLIP: integración que permite guiar la generación con descripciones en lenguaje natural, ampliando la expresividad visual.
- Stable Diffusion: modelo de difusión basado en espacios latentes, con alta flexibilidad gracias a extensiones como ControlNet y LoRA.
- MidJourney: modelo comercial basado en difusión, reconocido por su orientación artística y su popularidad en comunidades creativas.

**Plataformas y herramientas**

Los experimentos se ejecutaron principalmente en Google Colab Pro+, aprovechando infraestructura GPU (NVIDIA Tesla T4/A100). Se emplearon entornos en Python 3.10, con bibliotecas específicas como diffusers, transformers, torch (Stable Diffusion), pytorch_pretrained_biggan, CLIP (BigGAN+CLIP), y herramientas de análisis como Fréchet Inception Distance (FID). Esta infraestructura asegura replicabilidad, disponibilidad inmediata y resultados reproducibles sin necesidad de hardware especializado.

**Diseño experimental**

El estudio se basa en un enfoque controlado y reproducible permitiendo realizar una evaluación comparativa objetiva de los modelos generativos. Cada algoritmo fue sometido a las mismas pruebas estandarizadas utilizando un conjunto común de

estímulos textuales (prompts), diseñado para medir el rendimiento a través de un espectro de complejidad semántica creciente. Para garantizar la equidad en la comparación, todos los modelos generaron imágenes a partir de idénticos prompts en condiciones equivalentes, incluyendo múltiples réplicas por condición que permiten evaluar tanto la estabilidad (consistencia entre repeticiones) como la coherencia conceptual (fidelidad al prompt) de los resultados.

La construcción de la batería de estímulos se basó en principios de ingeniería de prompts, estratificando los estímulos en tres niveles de complejidad lingüística y conceptual creciente. Esta estructura permite observar no solo las diferencias en la interpretación del lenguaje entre algoritmos, sino también cómo el grado de especificidad e información contextual influye en la calidad visual y la fidelidad semántica de los resultados.

En el Nivel 1 – Básico (Conceptos Atómicos) se evalúa la capacidad fundamental del modelo para representar categorías semánticas amplias mediante objetos concretos sin modificadores, con ejemplos como "tiger", "pizza" o "castle". El Nivel 2 – Intermedio (Composición de Atributos) valora la competencia para integrar múltiples atributos visuales, relaciones espaciales básicas y contextos ambientales simples en escenas coherentes, como se aprecia en prompts del tipo "A tiger walking through tall grass" o "Old medieval castle surrounded by mist". Finalmente, el Nivel 3 – Avanzado (Especificación Narrativa y Estilística) prueba la habilidad para sintetizar escenas complejas y adherirse a convenciones estilísticas específicas, mediante descripciones como "A tiger in a dense jungle, photorealistic, golden hour lighting" o "A majestic medieval castle at dawn, in the style of a fantasy RPG concept art".

Todos los prompts se utilizaron exclusivamente en inglés, idioma predominante en los conjuntos de entrenamiento de los modelos evaluados, garantizando así la

máxima compatibilidad semántica y minimizando ambigüedades de traducción que pudieran afectar el rendimiento interpretativo.

La información se recolectó en dos fases complementarias:

Evaluación perceptual humana: realizada por 10 jueces invitados con formación en diseño gráfico, y comunicación visual, y experiencia de al menos dos años. Cada imagen fue valorada en una escala Likert (1–5) según los cinco criterios: fidelidad semántica, calidad visual, coherencia conceptual, diversidad generativa y accesibilidad visual. La evaluación fue ciega respecto al modelo utilizado, reduciendo sesgos de marca.

Evaluación automatizada (FID): se calcularon distancias estadísticas entre las imágenes generadas y conjuntos de 100 por categoría de imágenes reales de referencia tomadas de Pexels . Un menor valor de FID indica mayor similitud estructural.

El Fréchet Inception Distance (FID) mide la similitud estadística entre imágenes reales y generadas, tomando como base las características extraídas por la red Inception-v3. Estas se modelan como distribuciones gaussianas multivariadas, comparadas mediante la distancia de Fréchet:

$$\|\boldsymbol{u_r} - \boldsymbol{u_g}\|^2 + \boldsymbol{Tr}\left(\boldsymbol{\Sigma_r} + \boldsymbol{\Sigma_g} - \boldsymbol{2}\left(\boldsymbol{\Sigma_r}\boldsymbol{\Sigma_g}\right)^{\frac{1}{2}}\right)$$

donde μr Σr corresponden a media y covarianza de imágenes reales, y μg Σg a las generadas. Valores bajos de FID indican mayor fidelidad visual.

Rangos interpretativos:

- 0–10: Excelente (casi indistinguibles de las reales)
- 10–50: Bueno/Aceptable

- 50–100: Regular
- 100–200: Bajo realismo
- 200: Muy baja fidelidad

En este estudio, el FID se aplicó a cuatro modelos generativos para comparar su aproximación estadística frente a 100 imágenes reales por categoría. Para cada uno de los cinco prompts se generaron 40 imágenes por modelo en tres niveles de complejidad (básico, intermedio, avanzado).

## Análisis de resultados

### Análisis de Concordancia entre Evaluadores

Para garantizar la confiabilidad y consistencia de las evaluaciones realizadas por los diez participantes del estudio, se aplicó el coeficiente Alfa de Cronbach, un método empleado para medir la consistencia interna entre evaluadores. Este análisis permite determinar en qué medida los juicios de los participantes son coherentes entre sí al calificar los distintos criterios, asegurando que las puntuaciones reflejen de manera fiable las percepciones sobre las imágenes evaluadas.

El Alfa de Cronbach se calculó mediante la fórmula:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_{y_i}^2}{\sigma_x^2}\right)$$

Donde:

- k representa el número de evaluadores (k=10)
- $\sigma^2 Yi$ corresponde a la varianza de las puntuaciones de cada evaluador
- $\sigma^2 X$ denota la varianza de las puntuaciones totales agregadas.

La implementación se realizó considerando cada imagen evaluada como ítem y cada evaluador como fuente de medición, promediando previamente los cinco criterios de evaluación (Fidelidad Semántica, Calidad Visual, Diversidad Generativa, Coherencia Conceptual y Accesibilidad Visual) para cada imagen

Se establecieron los siguientes criterios de interpretación basados en la literatura especializada:

- 0.90 - 1.00: Confiabilidad Excelente
- 0.80 - 0.89: Confiabilidad Buena
- 0.70 - 0.79: Confiabilidad Aceptable
- < 0.70: Confiabilidad Pobre

Los valores obtenidos reflejan un nivel consistente de confiabilidad entre los evaluadores para los diferentes modelos generativos analizados

**Tabla 1.**

*Resultados del Alfa de Cronbach por Modelo Generativo*

| Modelo | Alfa de Cronbach | Nivel de Confiabilidad |
|---|---|---|
| BigGAN | 0.892 | Buena |
| BigGAN+CLIP | 0.896 | Buena |
| Stable Diffusion | 0.847 | Buena |
| Midjourney | 0.905 | Excelente |

Los valores obtenidos, todos superiores a 0.80, indican una consistencia interna satisfactoria entre los evaluadores, validando la confiabilidad de los datos recopilados para el análisis comparativo de los modelos generativos. El modelo Midjourney presentó el mayor nivel de concordancia (α=0.905), clasificándose en el rango de excelente

confiabilidad, mientras que los restantes modelos mostraron niveles de confiabilidad buena, con coeficientes superiores a 0.84, como se reflejan en la (Tabla 1).

Estos resultados confirman que las evaluaciones realizadas por los diez participantes mantienen patrones consistentes de criterio, permitiendo proceder con el análisis comparativo de los modelos con garantías metodológicas sobre la calidad y confiabilidad de los datos obtenidos. La alta concordancia inter-evaluadores respalda la validez interna del instrumento de evaluación y la capacidad de los participantes para aplicar los criterios establecidos de manera uniforme.

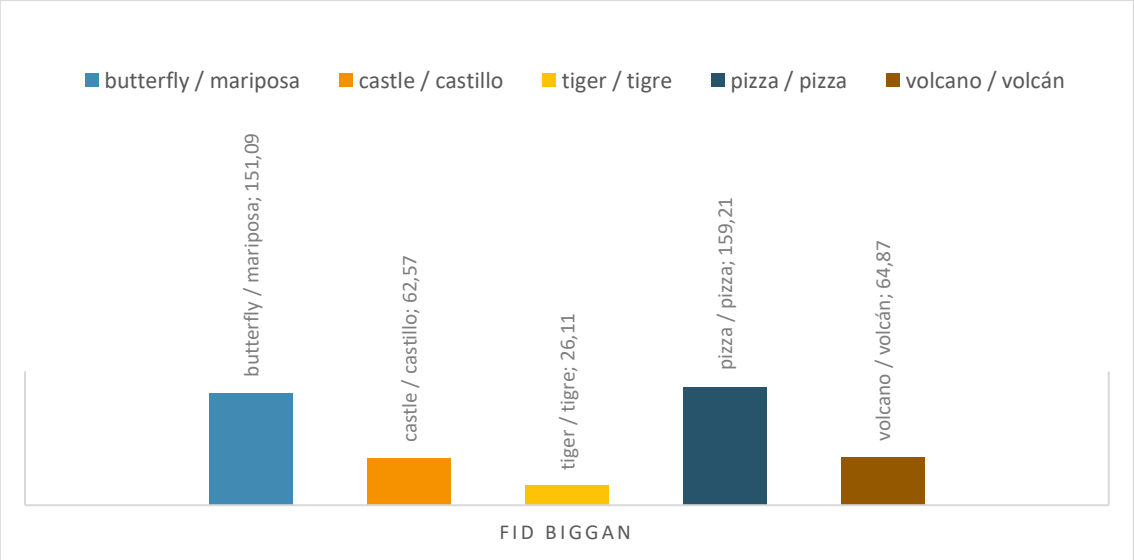
**Resultados cuantitativos mediante Fréchet Inception Distance (FID)**

Resultados de FID – BigGAN

El modelo BigGAN solo permitió la generación de imágenes a partir de etiquetas simples (single-label prompts), lo que restringió el análisis a categorías básicas. Los resultados obtenidos (Gráfico 1) evidencian un desempeño desigual: Tiger alcanzó el mejor valor (FID = 26.11), considerado excelente; Castle (62.57) y Volcano (64.87) mostraron calidad moderada; mientras que Butterfly (151.09) y Pizza (159.21) presentaron deficiencias significativas. En conjunto, estos hallazgos sugieren que BigGAN logra resultados aceptables en clases visualmente definidas, pero muestra limitaciones frente a categorías con mayor complejidad estructural o variabilidad contextual

**Figura 1.**

*Valores de BigGAN obtenidos mediante FID en cinco categorías.*



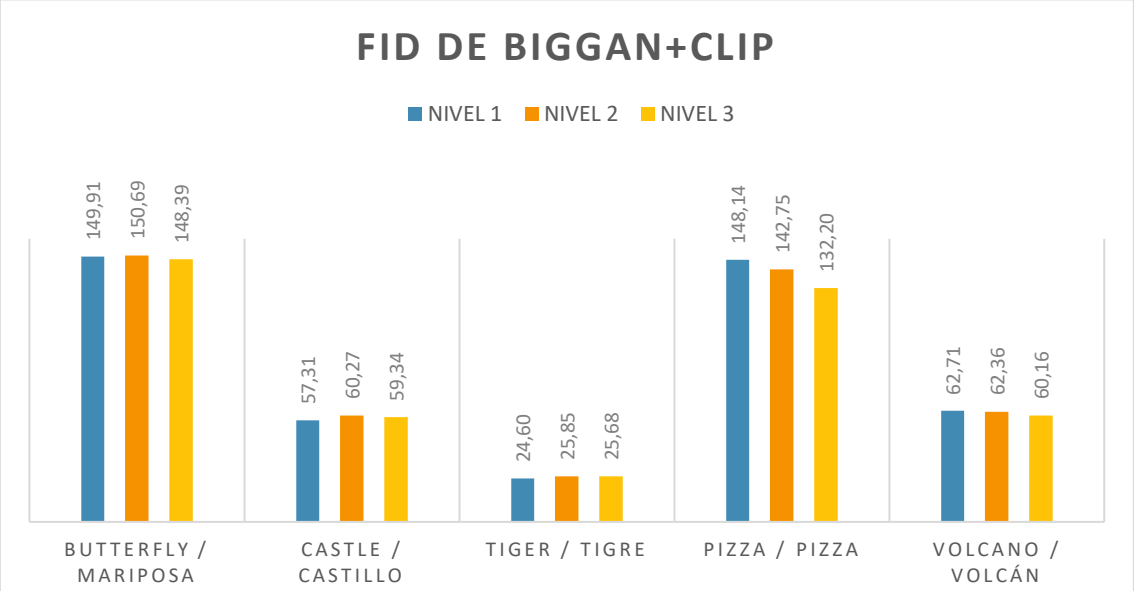

Resultados de FID – BigGAN+CLIP

El modelo BigGAN+CLIP fue evaluado en tres niveles de complejidad semántica (Gráfico 2). Los resultados evidencian un patrón claro: el aumento de detalle en los prompts mejora ligeramente la fidelidad visual, aunque el efecto depende de la categoría. Tiger alcanzó el mejor desempeño (FID ≈ 25), estable en los tres niveles. Castle y Volcano presentaron valores moderados (FID≈58–62) con variaciones mínimas. En Pizza, los valores descendieron de 148.14 a 132.20, reflejando que el enriquecimiento descriptivo favorece categorías con alta variabilidad. En contraste, Butterfly mantuvo valores elevados (FID≈ 150), lo que confirma la dificultad del modelo para reproducir detalles finos y texturas complejas.

Los resultados muestran que BigGAN+CLIP mejora la alineación texto–imagen cuando se incorporan descripciones más ricas, aunque su rendimiento sigue condicionado por la naturaleza del objeto representado.

**Figura 2.**

*Comparación de valores del modelo BigGAN+CLIP obtenidos mediante FID con prompts de tres niveles de complejidad.*



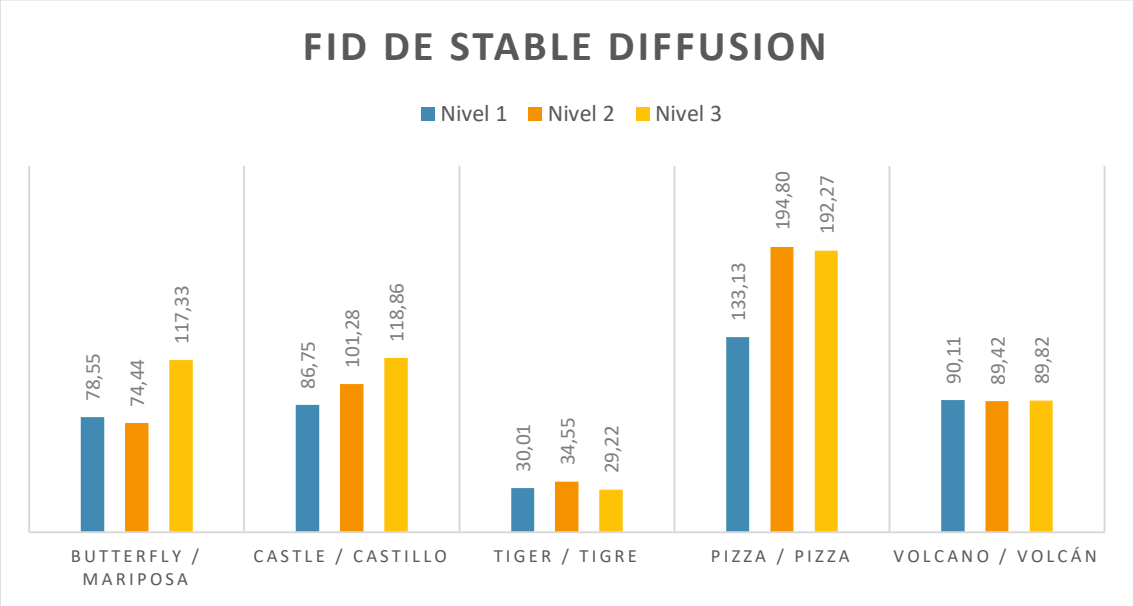

Resultados de FID – Stable Diffusion

El modelo Stable Diffusion mostró un rendimiento competitivo con variaciones según la categoría y el nivel de complejidad semántica (Gráfico 3). Tiger alcanzó valores excelentes (≈ 29–34) estables en los tres niveles, confirmando su solidez en conceptos definidos. Butterfly (≈ 74–117) presentó desempeño regular, con leves mejoras en el nivel intermedio, pero persistentes dificultades en detalles finos. Castle (≈ 86–118) se mantuvo en un rango moderado, con tendencia a empeorar conforme aumenta la complejidad. Pizza (≈ 133–194) evidenció el peor resultado, sin mejoras significativas al incrementar el detalle del prompt. En contraste, Volcano (≈ 89–99) permaneció estable en los tres niveles, con resultados aceptables.

Stable Diffusion genera imágenes de alta calidad en categorías estructurales claras, aunque conserva limitaciones frente a clases variables o de gran complejidad morfológica.

**Figura 3.**

*Comparación de valores del modelo Sable Diffusion obtenidos mediante FID con prompts por niveles de complejidad.*



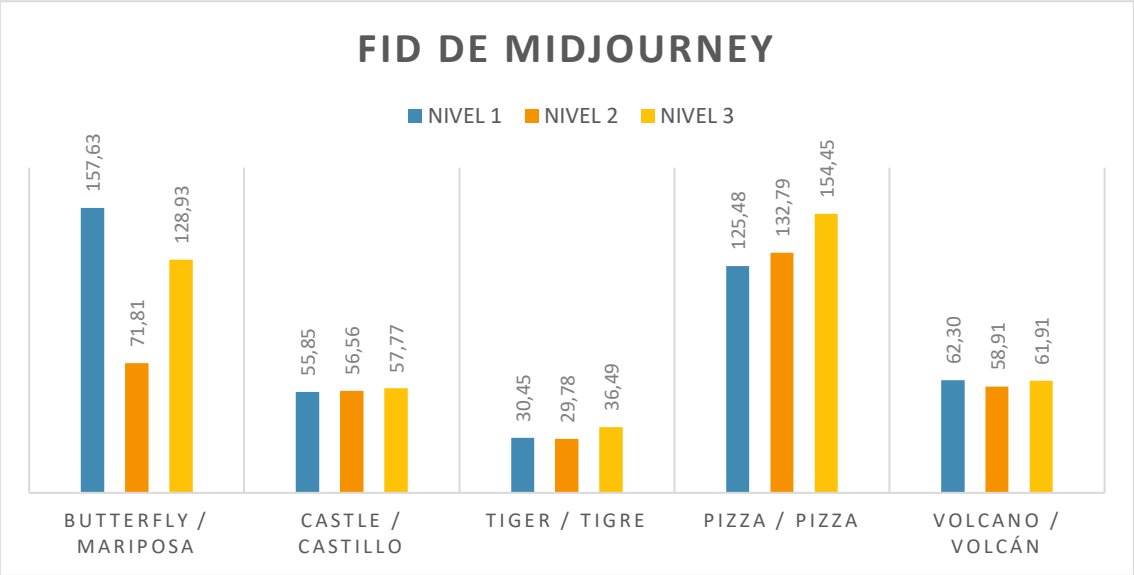

Resultados de FID - MidJourney

El modelo MidJourney mostró una relación directa entre la complejidad semántica de los prompts y la fidelidad visual, aunque con variaciones por categoría (Grafico 4). Butterfly (≈ 71–157) mejoró notablemente en el nivel 2, pero decayó en el nivel 3 por ruido semántico. Castle (≈ 55–57) permaneció estable en los tres niveles, evidenciando solidez en conceptos estructurados. Tiger (≈ 29–36) alcanzó valores excelentes, con ligeras mejoras en prompts intermedios, aunque los más complejos introdujeron inconsistencias (Gráfico 4). Pizza (≈ 125–154) mostró un comportamiento inverso: la complejidad adicional elevó el FID, reflejando que categorías variables no siempre se benefician de mayor detalle. Finalmente, Volcano (≈ 58–62) obtuvo mejores resultados en el nivel intermedio, pero disminuyó en el nivel 3.

En síntesis, MidJourney responde de manera óptima a prompts descriptivos moderados, mientras que narraciones excesivamente complejas pueden degradar la coherencia visual.

**Figura 4.**

*Comparación de valores del modelo MidJourney obtenidos mediante FID con prompts por niveles de complejidad.*




**Resultados cualitativos Basados en Evaluación Humana**

Después de efectuar el análisis cuantitativo usando la medida de Fréchet Inception Distance (FID), se sumó una valoración cualitativa enfocada en registrar la visión de las personas sobre las imágenes producidas por los modelos examinados. Este método intentó comparar los datos numéricos con juicios subjetivos efectuados por expertos en diseño visual y estudio gráfico, facilitando de este modo una interpretación más completa de la calidad, lógica y eficiencia representacional de cada modelo.

La evaluación se estructuró en torno a cinco criterios fundamentales: accesibilidad visual, calidad visual, coherencia conceptual, diversidad generativa y fidelidad semántica. Cada imagen fue valorada en una escala Likert de 1 a 5, donde 1

corresponde a un desempeño muy deficiente y 5 a un desempeño excelente. Estos criterios fueron seleccionados debido a su pertinencia para medir no solo el impacto estético, sino también la correspondencia semántica de las imágenes con los prompts textuales.

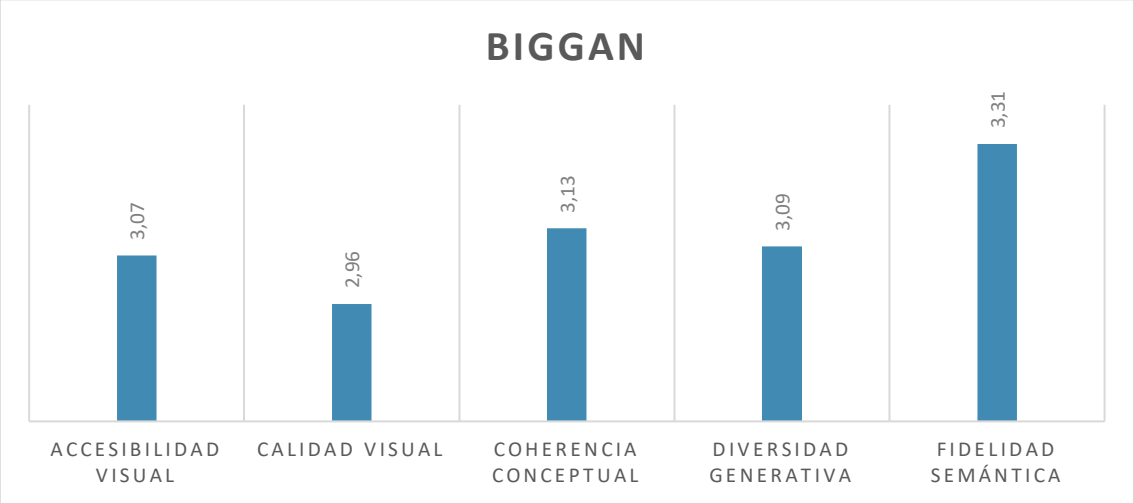
Análisis Cualitativo de Resultados – Modelo BigGAN

El modelo BigGAN, evaluado exclusivamente con prompts básicos debido a sus limitaciones arquitectónicas, demostró competencia en la representación de conceptos simples. Logró su mejor desempeño en fidelidad semántica (3.31), respaldado por coherente conceptual (3.13) y diversidad generativa (3.09), aunque con variaciones creativas limitadas (Gráfico 5).

No obstante, el modelo presentó deficiencias significativas en calidad visual (2.96) y accesibilidad visual (3.07), manifestadas en imágenes poco nítidas y de escaso realismo. Estos resultados posicionan a BigGAN como adecuado para aplicaciones que requieran representación básica, pero limitado para contextos que demanden riqueza estética o narrativa.

**Figura 5.**

*Evaluación cualitativa del modelo BigGAN en cinco criterios perceptivos.*

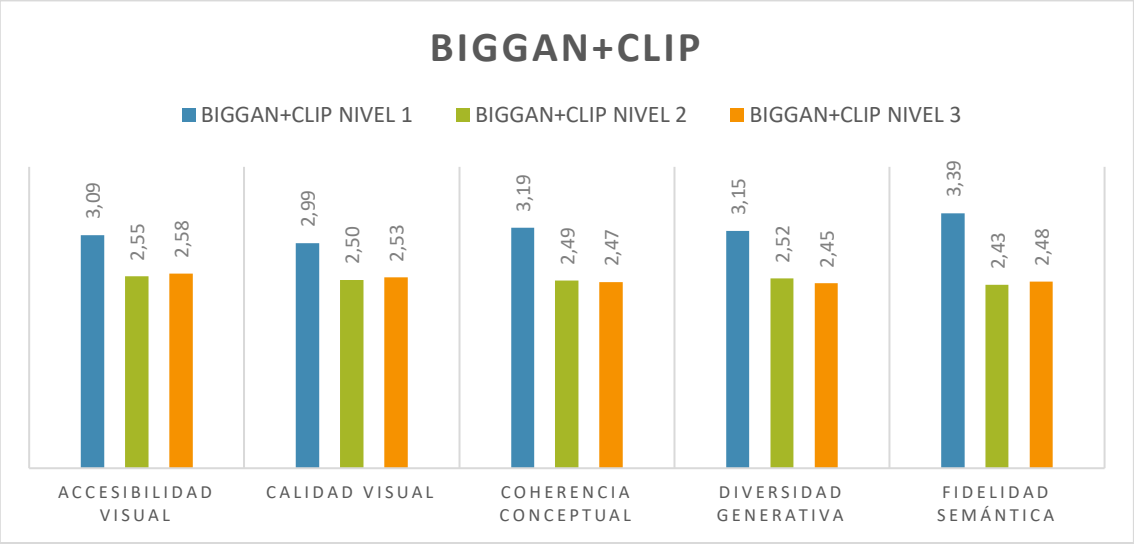


Análisis Cualitativo de Resultados – Modelo BigGAN+CLIP

El modelo BigGAN+CLIP demostró un rendimiento óptimo en el Nivel 1 (Conceptos Básicos), donde alcanzó sus máximas puntuaciones en fidelidad semántica (3.39) y coherencia conceptual (3.19), respaldadas por una diversidad generativa (3.15) y accesibilidad visual (3.09) consistentes. Sin embargo, al avanzar al Nivel 2 (Descripciones Detalladas), se observó un deterioro significativo en todos los criterios, con caídas notables en fidelidad semántica (2.43) y coherencia conceptual (2.49), lo que evidencia sus limitaciones para integrar atributos visuales múltiples. Finalmente, en el Nivel 3 (Narrativa Compleja), el modelo mostró una incapacidad para manejar escenas estilizadas, manteniendo puntuaciones bajas en fidelidad semántica (2.48) y diversidad generativa (2.45), lo que confirma su restricción a aplicaciones de baja complejidad descriptiva.narrativos (Gráfico 6).

**Figura 6.**

*Evaluación cualitativa del modelo BigGAN+CLIP en tres niveles de complejidad semántica.*



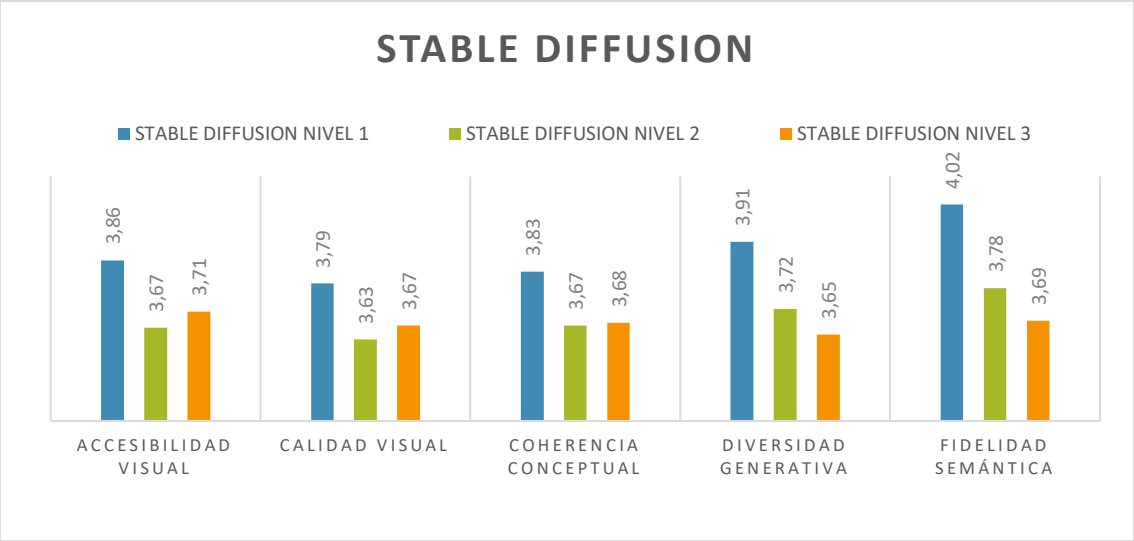
Análisis Cualitativo de Resultados – Modelo Diffusion

Stable Diffusion demostró un rendimiento sólido y consistente a través de todos los niveles de complejidad. En el Nivel 1 alcanzó su máximo desempeño, con destacadas puntuaciones en fidelidad semántica (4.02) y diversidad generativa (3.91), respaldadas por una alta accesibilidad visual (3.86) y coherencia conceptual (3.83), lo que confirma su excelente capacidad para interpretar conceptos básicos.

Al enfrentarse a los Niveles 2 y 3, el modelo mostró una adaptación notable, manteniendo puntuaciones consistentes a pesar del aumento en la complejidad descriptiva. Aunque se observaron ligeros descensos en fidelidad semántica (3.78 en Nivel 2 y 3.69 en Nivel 3) y calidad visual (3.63 en Nivel 2), los valores se mantuvieron por encima del umbral de aceptación, demostrando su robustez para manejar escenas narrativas y especificaciones estilísticas sin comprometer significativamente la comprensibilidad de las imágenes generadas. (Gráfico 7).

**Figura 7.**

*Evaluación cualitativa del modelo Stable Diffusion en tres niveles de complejidad semántica.*



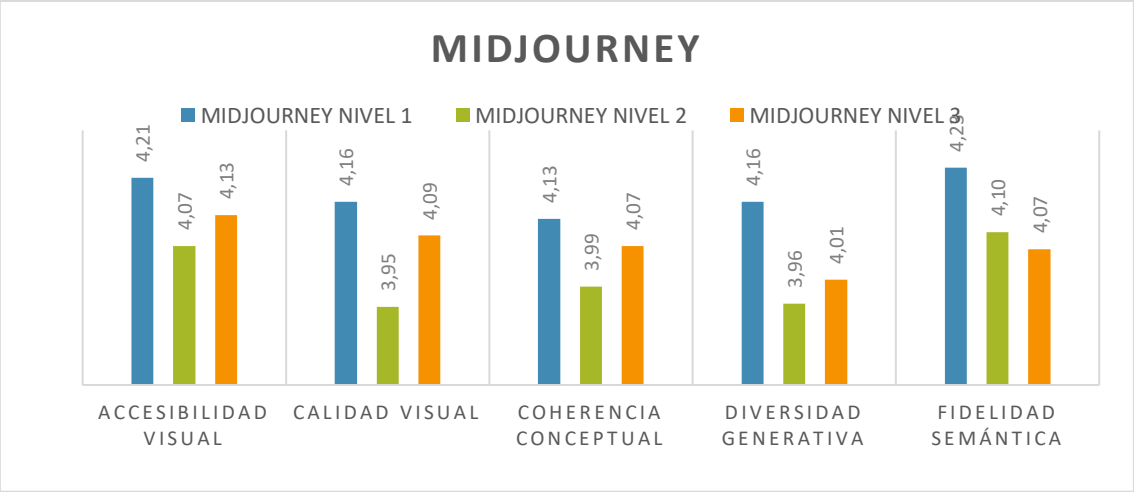
Análisis Cualitativo de Resultados – Modelo MidJourney

MidJourney demostró un desempeño excepcionalmente consistente a través de todos los niveles de complejidad. En el Nivel 1 alcanzó sus puntuaciones máximas, con una destacada fidelidad semántica (4.23) y accesibilidad visual (4.21), respaldadas por una diversidad generativa (4.16) y calidad visual (4.16) sobresalientes, lo que evidencia su capacidad para producir imágenes de alta calidad incluso a partir de instrucciones simples.

Al avanzar a los Niveles 2 y 3, el modelo mantuvo un rendimiento notablemente estable, con apenas variaciones mínimas en sus métricas. En el Nivel 3, todas las puntuaciones se mantuvieron por encima de 4.0, destacando la fidelidad semántica (4.07) y calidad visual (4.09), lo que confirma su superioridad para interpretar prompts narrativos y estilizados sin comprometer la calidad, consolidándose como el modelo más robusto y consistentemente alto en rendimiento del estudio. (Gráfico 8).

**Figura 8.**

*Evaluación cualitativa del modelo MidJourney en tres niveles de complejidad semántica.*



**Comparación de calidad visual (Fréchet Inception Distance (FID) vs Evaluación Humana)**

Tras el análisis cualitativo individual de cada modelo, se realizó una comparación entre la percepción humana de calidad visual y los valores obtenidos mediante la métrica objetiva Fréchet Inception Distance (FID). Para ello fue necesario normalizar los valores de FID, dado que su escala natural es inversa a la lógica de evaluación perceptiva.

El procedimiento se realizó mediante normalización min-max inversa, seguida de un reescalado lineal al rango [1, 5]. La fórmula utilizada fue la siguiente:

$$Valor\ Normalizado = 1 + 4 * (\frac{FID\text{max} \quad - \quad FID\text{modelo}}{FID\text{max} \quad - \quad FID\text{min}})$$

Donde:

FIDmodelo: Es el valor FID del modelo a transformar

FIDmin y FIDmax: Corresponden al menor y mayor valor de FID obtenido entre todos los modelos.

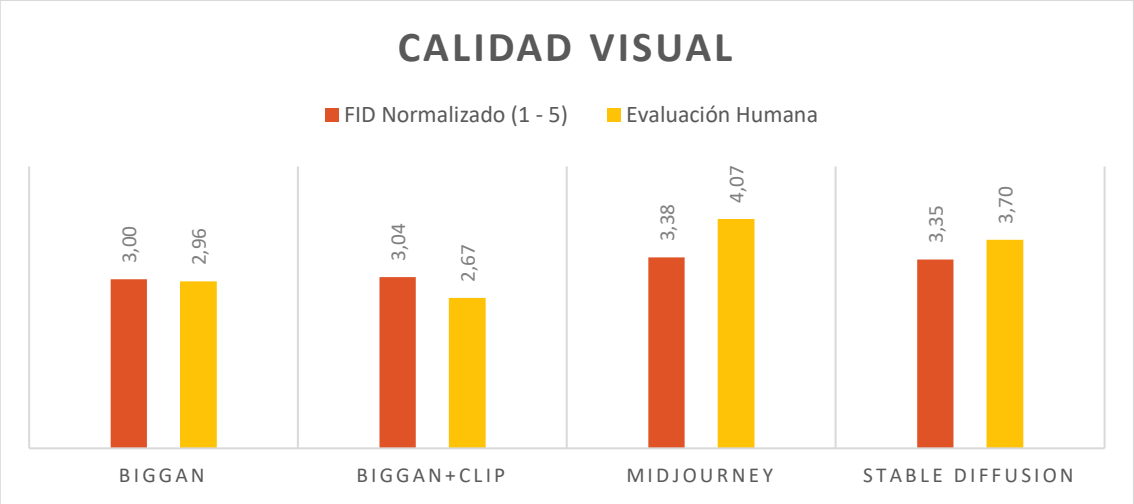
**El factor 1+4×(...)**: Ajusta el valor resultante al rango de 1 a 5.

De esta forma, los valores FID más bajos (mejor desempeño estadístico) se asignaron a puntajes más altos, y los valores FID más altos a puntajes más bajos, asegurando la coherencia interpretativa entre ambas métricas.

La comparación resultante (Gráfico 9) revela que BigGAN y BigGAN+CLIP muestran una correspondencia cercana entre ambas evaluaciones, mientras que en MidJourney se observa una clara divergencia: la evaluación humana lo sitúa como el mejor modelo (4,07), aunque el FID lo ubica solo ligeramente por encima de Stable Diffusion. Este último, en cambio, mantiene un equilibrio más consistente entre percepción humana (3,70) y métrica objetiva (3,35).

**Figura 9.**

*Comparativa de calidad Visual de resultados FID y resultados por evaluadores.*




## Conclusiones

El análisis integral de los resultados, mediante la combinación de la métrica Fréchet Inception Distance (FID) y la evaluación perceptual experta, revela diferencias estructurales y funcionales notables entre los modelos generativos evaluados, lo que permite una comprensión más matizada de la síntesis visual mediada por inteligencia artificial.

BigGAN, basado en una arquitectura clásica de redes generativas antagónicas (GANs), demostró ser efectivo en la representación de conceptos simples, aunque con marcadas limitaciones ante prompts de mayor complejidad semántica. Su dependencia de categorías predefinidas y su propensión al colapso modal confirman lo señalado por Brock et al. (2019) respecto a su escasa capacidad de generalización beyond de su entrenamiento supervisado.

La incorporación de CLIP en BigGAN+CLIP permitió una mejor alineación entre texto e imagen, validando el potencial de los modelos multimodales para mejorar la direccionalidad semántica. Sin embargo, como advierten Reynolds & McDonell (2021),

la integración textual no resuelve por completo las limitaciones intrínsecas del modelo base, especialmente en la representación de conceptos visualmente intricados o con alto nivel de abstracción.

MidJourney destacó por su consistencia perceptual y orientación estética, consolidándose como el modelo más robusto en términos de calidad visual y aceptación por parte de evaluadores humanos. No obstante, su tendencia a la sobre interpretación en escenas narrativas complejas sugiere un equilibrio inestable entre creatividad y fidelidad semántica, un fenómeno también observado por Oppenlaender (2022) en entornos de generación orientados al arte digital.

Stable Diffusion, si bien mostró un desempeño competitivo en términos de accesibilidad y diversidad generativa, presentó irregularidades significativas en escenas estilizadas o narrativas, lo que refleja limitaciones en la arquitectura de difusión para mantener la coherencia estructural ante descripciones complejas, tal como señalan Rombach et al. (2022).

Desde una perspectiva metodológica, este estudio refuerza la necesidad de complementar las métricas cuantitativas —como FID— con evaluaciones humanas basadas en criterios perceptuales y semánticos. Como sostiene Borji (2022), ninguna métrica automatizada captura de forma integral dimensiones como la creatividad, la expresividad o la adecuación estilística, aspectos clave en aplicaciones reales de estos modelos.

En síntesis, los resultados obtenidos evidencian que la eficacia de los modelos generativos no depende exclusivamente de su arquitectura o capacidad técnica, sino también de su aptitud para interpretar y traducir la complejidad del lenguaje natural en representaciones visuales coherentes, estéticamente consistentes y semánticamente alineadas. Estos hallazgos aportan criterios analíticos útiles para la selección de modelos según contextos de uso —académico, creativo o industrial— y abren nuevas

líneas de discusión en torno a la evaluación de sistemas de IA generativa, la ingeniería de prompts y las implicaciones éticas y culturales de la automatización en la creación visual.

**Referencias bibliográficas**


Alcaide-Marzal, J., & Diego-Mas, J. A. (2025). Computers as co-creative assistants: A comparative study on the use of text-to-image AI models for computer aided conceptual design. Computers in Industry, 164, 104168. https://doi.org/10.1016/j.compind.2024.104168

Arboleda Sánchez, C., & Patiño Ávila, J. (2024). Generación creativa con MidJourney: análisis estético y aplicaciones en diseño digital. Revista Latinoamericana de Innovación y Tecnología, 12(1), 45–62.

Borji, A. (2022). Pros and cons of GAN evaluation measures: New developments. Computer Vision and Image Understanding, 208, 103329. https://doi.org/10.1016/j.cviu.2021.103329

Brock, A., Donahue, J., & Simonyan, K. (2019). Large scale GAN training for high fidelity natural image synthesis. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1809.11096

Broncano, F. (2024). De la creación a la reiteración: Estética y repetición en la inteligencia artificial generativa. En G. López de Munain (Ed.), Las fronteras de la historia del arte y los estudios visuales: Reflexiones en torno a su objeto de estudio. Universidad Complutense de Madrid. https://doi.org/10.5209/eiko.90081

Camacho, F., & Paredes, A. (2023). Limitaciones éticas y técnicas en modelos cerrados de IA generativa. Estudios en Inteligencia Artificial Aplicada, 11(1), 1–15. https://doi.org/10.1234/eiaa.v11i1.015

Cedeño, D., & Ruiz, P. (2023). Aplicaciones artísticas de modelos generativos: Una comparación de MidJourney y DALL·E 2. Cuadernos de Arte Computacional, 6(3), 41–58. https://doi.org/10.29076/cac.v6i3.221

Chiou, Y.-C., Kuo, T.-C., & Chen, Y.-J. (2023). Design exploration of generative AI models in conceptual design: A comparison of DALL·E, MidJourney and Stable Diffusion. In Proceedings of the International Conference on Engineering Design (ICED23) (pp. 1–10). Cambridge University Press. https://doi.org/10.1017/pds.2023.161

Domínguez, C., & García, E. (2023). MidJourney: Democratización del arte visual mediante IA. Comunicación y Sociedad Digital, 10(4), 77–92. https://doi.org/10.26441/csd.v10i4.879

Du, Y., Li, S., Tenenbaum, J. B., & Torralba, A. (2023). Understanding diffusion models: A unified perspective. arXiv preprint. https://doi.org/10.48550/arXiv.2208.11970

Frolov, S., Hinz, T., Raue, F., Hees, J., & Dengel, A. (2021). Adversarial text-to-image synthesis: A review. Neural Networks, 144, 187–214. https://doi.org/10.1016/j.neunet.2021.08.021

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems (NeurIPS) (pp. 2672–2680). https://papers.nips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3

-Abstract.html

Gui, J., Sun, Z., Wen, Y., Tao, D., & Ye, J. (2021). A review on generative adversarial networks: Algorithms, theory, and applications. IEEE Transactions on Knowledge and Data Engineering, 35(2), 331–353. https://doi.org/10.1109/TKDE.2021.3075386

Gualdoni, F. (2024). La autoría en crisis: Reflexiones desde la creación visual por IA. En G. López de Munain (Ed.), Las fronteras de la historia del arte y los estudios visuales: Reflexiones en torno a su objeto de estudio. Universidad Complutense de Madrid. https://doi.org/10.5209/eiko.90081

Hernández, J. (2024). La caja negra de la imagen: Problemas de trazabilidad y agencia algorítmica. En G. López de Munain (Ed.), Las fronteras de la historia del arte y los estudios visuales: Reflexiones en torno a su objeto de estudio. Universidad Complutense de Madrid. https://doi.org/10.5209/eiko.90081

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems (NeurIPS), 33, 6840–6851. https://arxiv.org/abs/2006.11239

Huang, H., Yu, P. S., & Wang, C. (2020). An introduction to image synthesis with generative adversarial nets. arXiv preprint arXiv:2001.06937. https://doi.org/10.48550/arXiv.2001.06937

Li, X. L., & Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint. https://doi.org/10.48550/arXiv.2101.00190

Martín Prada, J. (2024). La creación artística visual frente a los retos de la inteligencia artificial: Automatización creativa y cuestionamientos éticos. En G. López de Munain (Ed.), Las fronteras de la historia del arte y los estudios visuales: Reflexiones en torno a su objeto de estudio. Universidad Complutense de Madrid. https://doi.org/10.5209/eiko.90081

Merino, J. (2024). Evaluar lo generado: Crítica de la imagen artificial y legitimación cultural. En G. López de Munain (Ed.), Las fronteras de la historia del arte y los estudios visuales: Reflexiones en torno a su objeto de estudio. Universidad Complutense de Madrid. https://doi.org/10.5209/eiko.90081

Müller, J., & Lee, S. (2023). Human-centered evaluation of generative design workflows with AI models. Journal of Design Research, 21(2), 115–132. https://doi.org/10.1504/JDR.2023.129876

Oppenlaender, J. (2022). Text-to-image synthesis for abstract and artistic prompts: Analyzing the creative potential of generative AI. In Proceedings of the Creativity & Cognition Conference. ACM. https://doi.org/10.1145/3527927.3532790

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. arXiv preprint. https://doi.org/10.48550/arXiv.2103.00020

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation. arXiv preprint.

https://doi.org/10.48550/arXiv.2102.12092

Reynolds, L., & McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. arXiv preprint. https://doi.org/10.48550/arXiv.2102.07350

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 10684–10695). IEEE. https://doi.org/10.1109/CVPR52688.2022.01042

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training GANs. Advances in Neural Information Processing Systems (NeurIPS), 29. https://arxiv.org/abs/1606.03498

Singh, M., Sharma, S., Kumar, N., & Deb, D. (2020). A survey and taxonomy of adversarial neural networks for text-to-image synthesis. In International Conference on Computational Intelligence in Pattern Recognition (pp. 45–57). Springer. https://doi.org/10.1007/978-981-15-4288-6_5

Sri Krishna, R. (2022). MidJourney vs DALL·E 2: A comparative analysis. AI Practitioners Digest, 5(1), 22–30. https://doi.org/10.5281/zenodo.7007517

Tao, M., Tang, H., Wu, F., & Chen, Y. (2023). Text-Guided Image Generation with CLIP-Conditioned GANs. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2023.3241234

Wang, Z. (2023). AI-based text-to-image synthesis: A review. IEEE Access, 11, 100234–100249. https://doi.org/10.1109/ACCESS.2023.3306422

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. (2017). StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 5907–5915). IEEE. https://doi.org/10.1109/ICCV.2017.629

Zhang, N., & Tang, H. (2023). Text-to-image synthesis: A decade survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2023.3286457

Zhang, M., & Ortega, L. (2023). Evaluating creativity and control in text-to-image generation tools. International Journal of Creative Interfaces and Computer Graphics, 13(1), 65–78. https://doi.org/10.4018/IJCICG.2023010105