Revista Innovación y Software
Vol. 7, No. 1, Mes Marzo - Agosto, 2026
ISSN: 2708-0935
Pág. 95-108
https://revistas.ulasalle.edu.pe/innosoft

Tipo de artículo: Artículos de revisión
Temática: Inteligencia artificial
Recibido: 10/10/2025 | Aceptado: 17/11/2025 | Publicado: 30/3/2026

Identificadores persistentes:
DOI: 10.48168/innosoft.s29.a356
ARK: ark:/42411/s29.a356

Generación de imágenes a partir de texto mediante inteligencia artificial: una revisión sistemática

Text-to-Image Generation Using Artificial Intelligence: A Systematic Review

Zaleth Rivas Calderón[1][000-0002-9797-151]*, **Estefany Villanueva Rosales**[2][], **Marcelino Torres Villanueva**[3][0000-0002-9797-1510]

[1]Universidad Nacional de Trujillo. Trujillo, Perú.. zrivasca@unitru.edu.pe
[2]Universidad Nacional de Trujillo. Trujillo, Perú.. elvillanuevaro@unitru.edu.pe
[3]Universidad Nacional de Trujillo. Trujillo, Perú.. mtorres@unitru.edu.pe

*Autor para correspondencia: zrivasca@unitru.edu.pe

Resumen
Este estudio aborda distintos enfoques empleados en la generación de imágenes a partir de texto mediante inteligencia artificial, con especial atención a la relación semántica que se establece entre la descripción textual y la imagen generada en los modelos texto–imagen. Asimismo, se revisa la fiabilidad de las métricas empleadas para evaluar su desempeño. Esto con la finalidad de conocer sus capacidades y limitaciones actuales. La investigación se llevó a cabo siguiendo la metodología PRISMA, para lo cual se seleccionaron 18 artículos de acuerdo con los criterios establecidos, que abordaban temas relacionados con arquitecturas de difusión, mecanismos de control semántico, atención a nivel de frase y prompt engineering. Los resultados señalan que los modelos basados en difusión son los más utilizados, mientras que los modelos GAN y VAE se emplean mayormente en aplicaciones de nicho. A partir del análisis realizado, se identificaron tres niveles de control: atributos visuales, composición y estilo. Sin embargo, actualmente se observan diversas limitaciones en las métricas usadas para evaluar el alineamiento semántico y la persistencia de ciertos sesgos asociados a modelos preentrenados. Las conclusiones señalan que los modelos de difusión son los más utilizados en la literatura reciente y que el uso de técnicas como LoRA ayuda a mejorar la coherencia entre texto e imagen. Estos resultados sugieren que todavía es necesario profundizar en el estudio de la atención relacional, en particular en el desarrollo de métricas estandarizadas en futuras investigaciones.

Palabras claves: Generación de imágenes a partir de texto, Inteligencia artificial generativa, Modelos multimodales, Modelos de difusión, Alineamiento semántico

Abstract
This study examines different approaches used in text-to-image generation through artificial intelligence, with particular emphasis on the semantic relationship established between textual descriptions and the images generated by text–image models. In addition, the reliability of the metrics used to evaluate their performance is reviewed, with the aim of identifying their current capabilities and limitations. The research was conducted following the PRISMA methodology, through which 18 articles were selected according to predefined criteria. These studies addressed topics related to diffusion architectures, semantic control mechanisms, phrase-level attention, and prompt engineering. The results indicate that diffusion-based models are the most widely used,

while GAN and VAE models are primarily applied in niche applications. Based on the analysis, three levels of control were identified: visual attributes, composition, and style. However, several limitations are currently observed in the metrics used to assess semantic alignment, as well as the persistence of certain biases associated with pretrained models. The conclusions indicate that diffusion models dominate the recent literature and that the use of techniques such as LoRA contributes to improving text–image coherence. These findings suggest that further research is still required on relational attention, particularly regarding the development of standardized metrics in future studies.

Keywords: *Text-to-Image Generation, Generative Artificial Intelligence, Multimodal Models, Diffusion Models, Semantic Alignment*

Introducción

La investigación en inteligencia artificial generativa, particularmente en los modelos que generan imágenes a partir de texto, ha crecido de forma exponencial en los últimos años gracias a los avances de los modelos de difusión condicionados y las arquitecturas Transformer. Los nuevos modelos de generación visual multimodal han cambiado la manera en que se crean imágenes a partir de descripciones textuales, permitiendo obtener resultados de alta calidad. Este avance representa una innovación en la automatización de contenidos, la visualización educativa y diversas aplicaciones creativas en distintos sectores [1]. En este contexto, Text-to-Image Diffusion Models han emergido como la metodología dominante dentro de la generación de imágenes condicionadas por texto, gracias a su robustez, versatilidad y capacidad para producir resultados visuales comparables a fotografías reales [2].

Sobre esta base, durante los últimos cinco años, diversos estudios han mostrado interés en la capacidad de los modelos de generación de imágenes por texto para representar correctamente las descripciones textuales del usuario. Por ejemplo, estudios recientes han incorporado el uso de funciones de recompensa durante el entrenamiento de modelos de difusión [3]. Los resultados indican que este tipo de estrategias mejora el alineamiento semántico, en

especial cuando se emplea retroalimentación para reforzar la relación entre texto e imagen en aspectos como la cantidad y el tipo de objetos representados.

De forma complementaria, se ha analizado el control del estilo mediante estrategias de semantic guidance, las cuales permiten ajustar determinadas características visuales sin comprometer la coherencia con la descripción textual [4]. Combinadas con técnicas de atención cruzada refinada y mecanismos adaptativos, estas estrategias han demostrado reducir errores relacionados con el conteo de objetos y la representación de relaciones espaciales complejas, lo que supone un avance respecto a trabajos previos.

A pesar de los avances recientes, todavía existen vacíos importantes en la literatura. En particular, los mecanismos y las arquitecturas diseñadas para mejorar el control y la correspondencia semántica en modelos texto–imagen no se han consolidado de manera clara ni bajo un marco metodológico uniforme. Muchos estudios se centran en aplicaciones muy específicas o en ajustes puntuales relacionados con el estilo y el control de atributos, lo que dificulta ver sus fortalezas o limitaciones [5]. Además, la comparación objetiva entre métodos suele ser complicada debido a la falta de métricas estandarizadas que midan de forma consistente la alineación entre texto e imagen.

En este contexto, en este trabajo se propone examinar de manera sistemática los diferentes métodos, arquitecturas y mecanismos que se utilizan en los modelos texto–imagen, con el propósito de mejorar el control y la correspondencia entre las imágenes generadas y las descripciones que las acompañan. A partir de esta revisión, se busca descubrir cuáles son las tendencias más relevantes en la literatura reciente y también identificar tanto las fortalezas como las limitaciones de los enfoques que se han propuesto. Es así que, se realiza una revisión sistemática de trabajos indexados en bases de datos científicas, con el fin de responder las siguientes preguntas clave: ¿qué mecanismos de control semántico son más comunes?, ¿qué tipos de arquitecturas se han propuesto?, y ¿qué retos todavía no se han superado?

Materiales y métodos o Metodología computacional

En el presente trabajo se ha llevado a cabo una revisión sistemática de la literatura científica indexada en diferentes bases de datos, siguiendo los lineamientos de la declaración PRISMA (Preferred Reporting Ítems for Systematic Reviews and Meta-Analyses). Este enfoque se seleccionó con el propósito de detectar, analizar y sintetizar de forma adecuada la evidencia disponible sobre los mecanismos de control y alineamiento semántico en modelos texto–imagen.

La búsqueda bibliográfica se realizó en las bases de datos Scopus, SpringerLink y Redalyc, las mismas fueron elegidas por su relevancia y calidad académica, así como por la extensa cantidad de literatura científica disponible en los temas relacionados con la inteligencia artificial, la generación de imágenes y los modelos multimodales.

Estrategia de búsqueda

La búsqueda se realizó durante los meses de noviembre y diciembre del 2025, tomando en cuenta artículos publicados entre 2021 y 2025.

Se usaron combinaciones de términos clave asociados con modelos texto–imagen, generación de imágenes,

modelos de multimodales y mecanismos de alineamiento semántico. Las ecuaciones de búsqueda se adaptaron ligeramente según las diferentes características de cada base de datos con el fin de incrementar la recuperación de estudios relevantes. A continuación, se detallan las fórmulas de búsqueda empleadas y la cantidad de artículos obtenidos en la búsqueda en cada repositorio bibliográfico.

Tabla 1. Fórmulas de búsqueda empleadas en las bases de datos y el número de artículos encontrados respectivamente.

Base de datos	**Fórmula de búsqueda**	**Cantidad de artículos encontrados**
Scopus	("text-to-image.ºR "text image.ºR ïmage generation from text") AND (ïmage generation.ºR ïmage synthesis") AND ("multimodal models.ºR "vision- language models") AND ("semantic alignment.ºR "text-image alignment.ºR "semantic consistency") OR (çontrollable image generation.ºR "generation control")	45
SpringerLink	("text-to-image") AND (.ªrchitectures.ºR .ªpproaches") AND ("semantic alignment") AND (ïmage generation")	27
Redalyc	(text-to-image) AND (image generation) AND (architectures OR models) AND (control)	19

Antes de seleccionar los artículos que se van a incluir en la revisión sistemática, se establecieron los criterios de inclusión y exclusión, con el fin de garantizar la selección de material bibliográfico conforme a las características y objetivos planteados en el estudio.

Criterios de inclusión

- Incluir únicamente artículos de investigación y no de revisión, estudios de caso único, libros o manuales.
- Material bibliográfico presentado en idioma español o inglés.
- Artículos de acceso abierto y con estado finalizado.
- Abordarán modelos texto–imagen basados en arquitecturas de difusión o similares.
- Analizarán mecanismos de control o alineamiento semántico entre texto e imágenes generadas.
- Artículos que se hayan publicado entre 2025 y 2021, ambos inclusive.

Criterios de exclusión

- Se excluyen los estudios que se refieran a generación de otros formatos como video o audio, que no sean imágenes.
- No abordarán explícitamente el problema del alineamiento semántico en modelos texto–imagen.
- Los centrados únicamente en generación de texto.
- Los que aborden modelos de generación que no sean específicamente texto a imagen.
- Presentarán enfoques puramente teóricos sin validación experimental.
- Las publicaciones que no se encuentran completas o disponibles en su totalidad en los repositorios seleccionados.

Proceso de selección de estudios

El proceso de selección de estudios se realizó siguiendo las cuatro fases de la metodología PRISMA, cuyo flujo se resume en el diagrama correspondiente.

En la fase de identificación, se recuperaron un total de 91 registros a partir de las búsquedas realizadas en repositorios bibliográficos especializados, distribuidos de la siguiente manera: Scopus (n = 45), Redalyc (n = 19) y SpringerLink (n = 27).

Posteriormente, en la fase de cribado, se realizó una revisión inicial de los títulos, a partir de la cual se excluyeron 40 registros por no estar relacionados con el objetivo de estudio. No se encontraron registros duplicados en esta etapa (n

= 0), por lo que el número de estudios cribados se redujo a 51.

En la fase de evaluación de idoneidad, se llevó a cabo la lectura de los resúmenes y posteriormente se excluyeron 33 estudios, por no cumplir los criterios de inclusión establecidos. Las principales razones de exclusión fueron la falta de relación directa con la generación texto–imagen (n = 30) y el enfoque en tareas de generación audio–imagen en lugar de texto–imagen (n = 3). Luego de la eliminación, se seleccionaron preliminarmente 18 artículos para evaluar su elegibilidad.

Finalmente, en la fase de inclusión, los 18 estudios seleccionados en la fase anterior cumplieron todos los criterios de inclusión y fueron seleccionadas para realizar la revisión sistemática. Estos trabajos constituyen el conjunto final de artículos analizados en el presente trabajo.

Revista Innovación y Software
Vol. 7, No. 1, Mes Marzo - Agosto, 2026
ISSN: 2708-0935
Pág. 95-108
https://revistas.ulasalle.edu.pe/innosoft

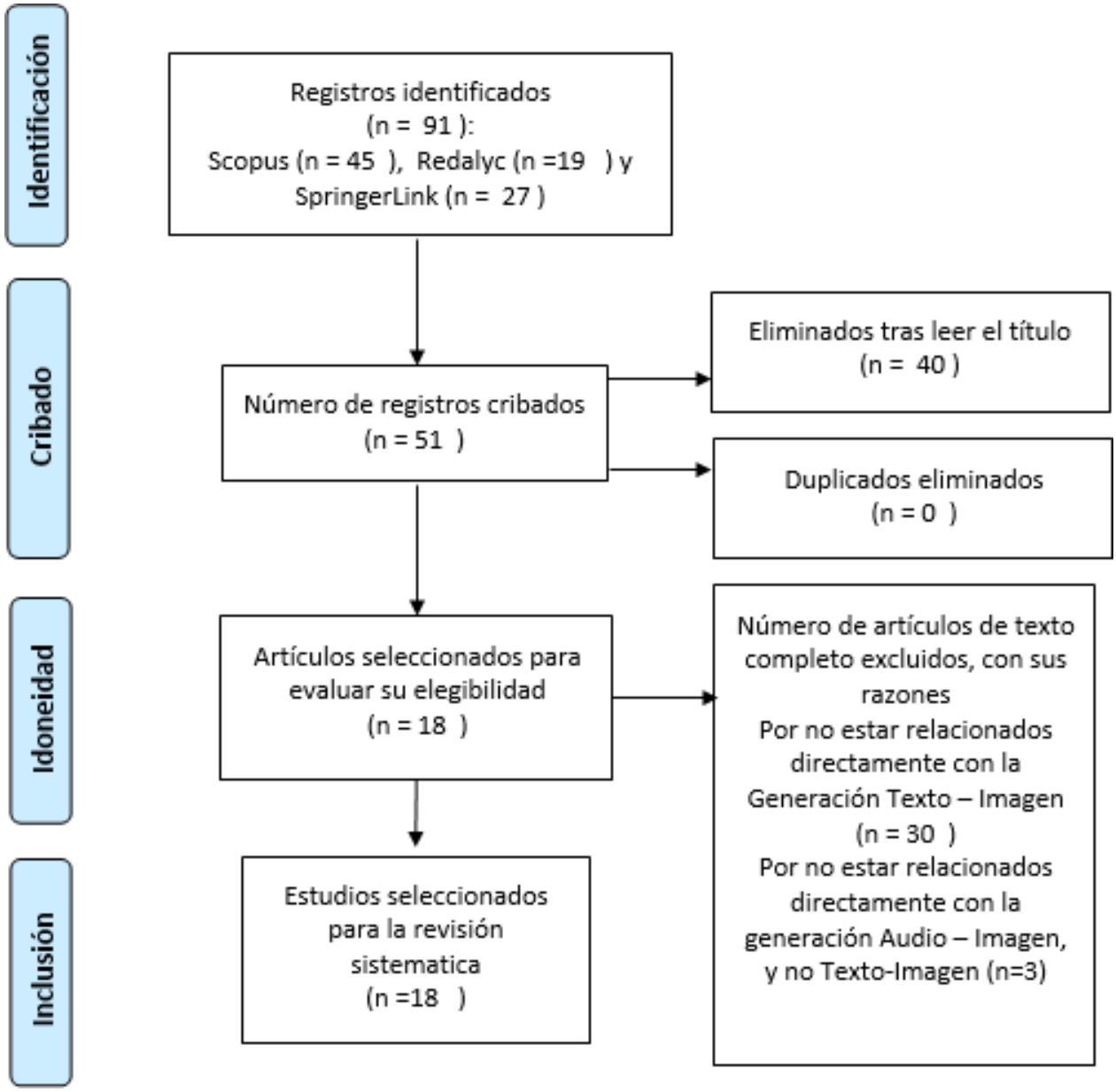

Figura 1. Diagrama de flujo PRISMA en cuatro niveles.

Análisis de los estudios incluidos

Los artículos seleccionados fueron analizados de forma cualitativa y comparativa, identificando las investigaciones más destacadas con relación al objetivo planteado previamente. Este análisis permitió identificar tendencias

actuales y desafíos abiertos en el desarrollo de modelos texto–imagen.

Resultados y discusión

Para este análisis se consideran 18 estudios recientes publicados entre 2021 y 2025, los cuales se centran en modelos texto–imagen con mecanismos de control y alineamiento semántico. No todos los artículos incluyen modelos con arquitecturas explícitas para cada tipo de análisis; por lo tanto, cada gráfico refleja únicamente los estudios relevantes para cada dimensión (arquitecturas, mecanismos, estrategias de control, métricas y tendencias).

En los estudios analizados, las arquitecturas basadas en diffusion models predominan sobre GANs y VAEs, especialmente para tareas de control semántico y generación de contenido multimodal [6–8]. Por ejemplo, modelos como Blended Latent Diffusion, que se emplea en la edición de imágenes y Stable Diffusion, usado en la conservación del patrimonio arquitectónico, muestran mayor capacidad para incorporar información textual compleja [6,9]. Los GANs se utilizan principalmente en áreas específicas como arte y patrimonio, mientras que los VAE se emplean en aplicaciones científicas y médicas [10].

Tabla 2. Número de artículos por Arquitectura

Arquitectura principal	**Artículos**
Diffusion-based (incluye Latent / Stable / adaptaciones con LoRA/ControlNet/CLIP)	12
GAN (hierarchical / domain-specific)	1

Nota: Datos usados: los 13 artículos que proponen modelos. Clasificación por arquitectura principal observada en cada artículo.

La Tabla 2 indica la prevalencia de los modelos de difusión en cantidad de publicaciones, evidenciando una tendencia orientada a arquitecturas capaces de integrarse con mecanismos de control y alineamiento semántico más exactos.

Por otro lado, asimismo se muestra una diversidad de estrategias para alinear texto e imagen, que incluyen el uso de prompts estructurados, embeddings semánticos de grano fino, técnicas de adaptación de bajo rango (LoRA) y mecanismos de ajuste fino guiados por texto [8,11,12]. Además, se resaltan métodos de corrección de sesgos empleando GradBias y ajustes de pesos de palabras con el fin de mejorar la fidelidad semántica [13].

Los métodos actuales utilizan técnicas de control en tres grados: características (color, forma, estilo), composición (posición y relación entre objetos) y estilo artístico o cultural, como se muestra en la figura 2.

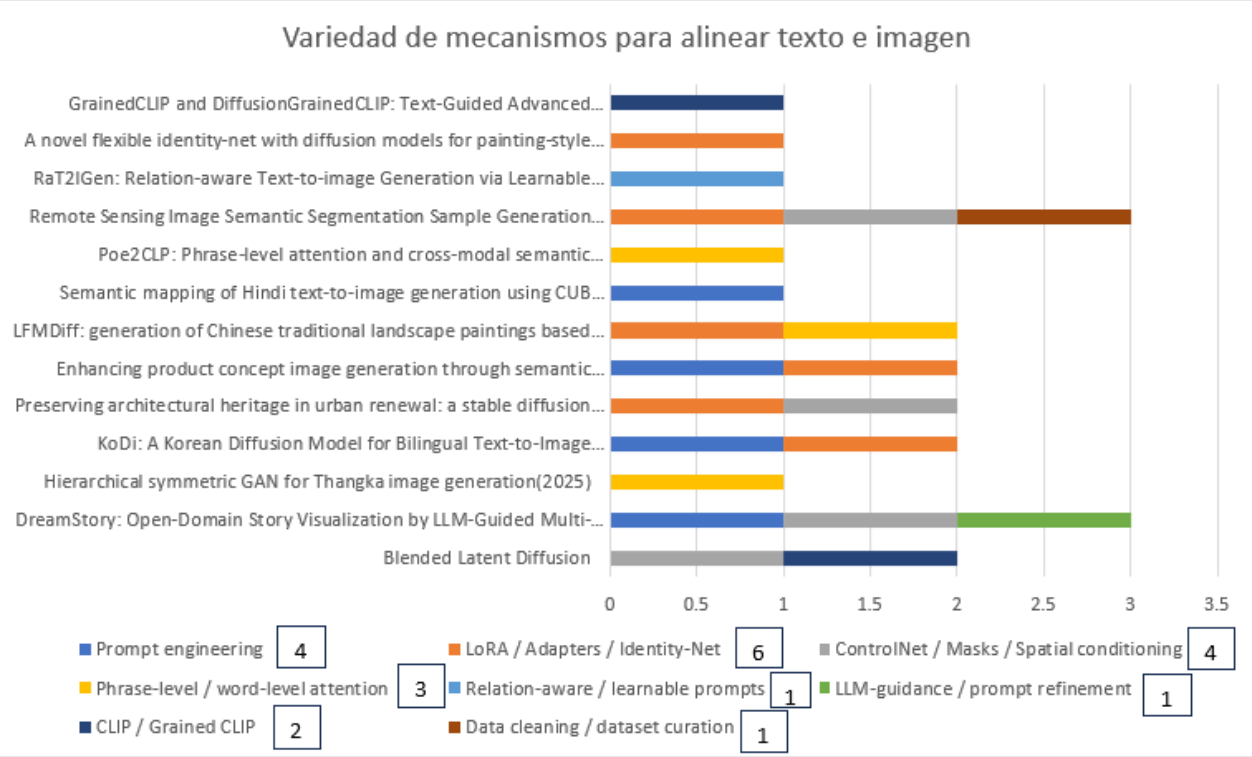

Figura 2. Datos usados: los 13 artículos que proponen modelos. Gráfico de barras apiladas sobre la Variedad de mecanismos para alinear texto e imagen en los artículos seleccionados

Esta indica que los ajustes de dominio específico mediante LoRA son los más frecuentes, seguidos de la combinación de attention mechanisms y prompt engineering o entrenamiento jerárquico [11, 14]. Estos mecanismos han mostrado un impacto positivo en la coherencia semántica de las imágenes generadas, aunque su efectividad depende de la complejidad del prompt y de la diversidad de datos de entrenamiento [15, 16].

También se observa que los enfoques actuales aplican principalmente estrategias de control en tres niveles: atributos, composición y estilo artístico o cultural.

Revista Innovación y Software
Vol. 7, No. 1, Mes Marzo - Agosto, 2026
ISSN: 2708-0935
Pág. 95-108
https://revistas.ulasalle.edu.pe/innosoft

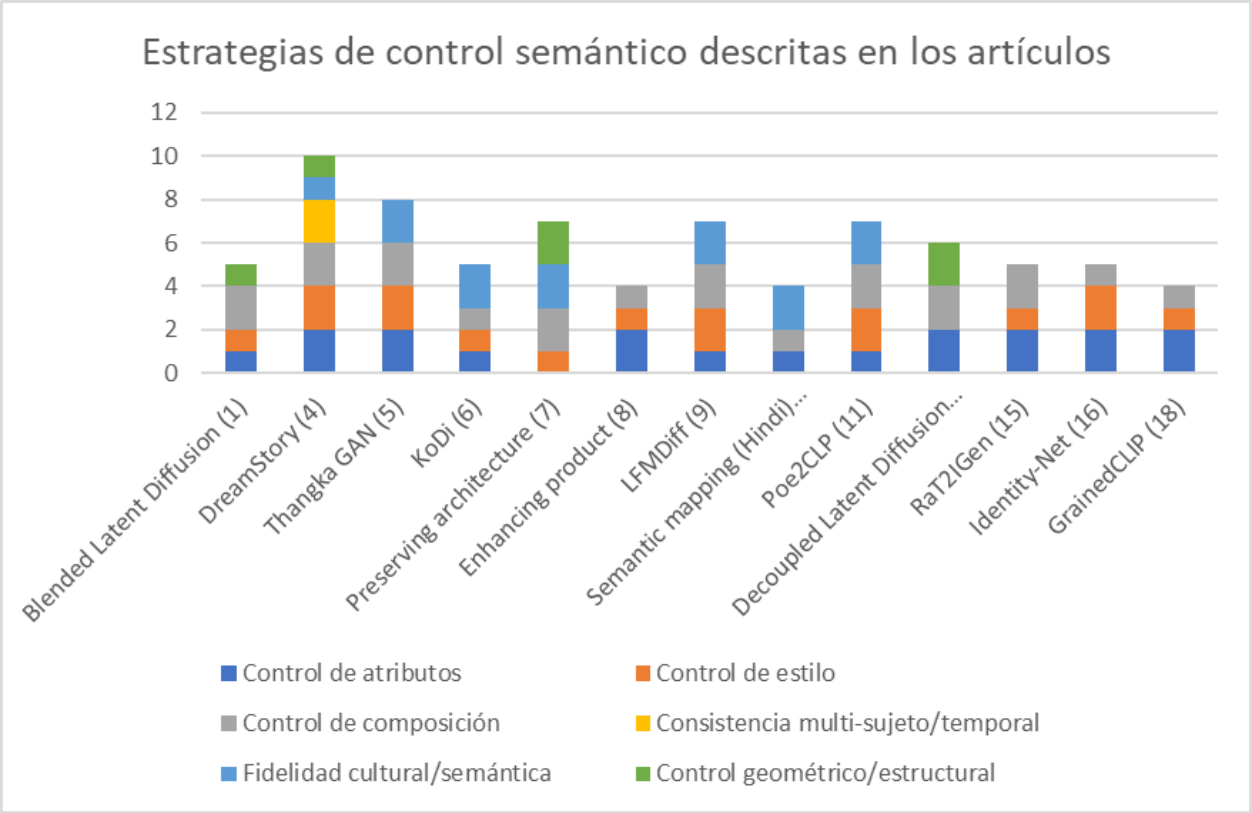

Figura 3. Datos usados: los 13 artículos que proponen modelos. Gráfico de barras apiladas de estrategias de control utilizadas en los artículos seleccionados. Para la graficación se asignó un puntaje según la presencia de la estrategia: No se usó = 0, Sí se usó = 2, Parcialmente usado = 1.

El gráfico descriptivo de "Estrategias de control" organiza por artículos y tipos de control, mostrando que la mayoría de los estudios combinan control de atributos y composición, con un menor número incorporando control de estilo cultural o artístico. Esto indica un interés creciente en generar imágenes coherentes no solo con la descripción textual, sino también con convenciones estéticas específicas [8, 17].

Asimismo, se identifican diversas métricas para evaluar la correspondencia texto–imagen como: CLIPScore, FID, IS, métricas de coherencia semántica y evaluación humana [11, 12, 14]. Cada métrica presenta limitaciones y oportunidades como se describe a continuación:

Tabla 3. Métricas usadas para evaluar la correspondencia texto-imagen

Métrica	**Sensible al alineamiento semántico**	**Reproducibilidad**
FID	Baja: captura diferencias visuales, no intención textual	Alta: automática y reproducible con dataset fijo
IS (Inception Score)	Baja: enfocado en calidad/diversidad, no match textual	Alta: cálculo automático
CLIPScore / CLIP similarity	Media-Alta: correlación con semántica general, falla en compositionality	Alta: evaluación automática y consistente
LPIPS / SSIM	Baja-Media: mide similitud perceptual, no intención textual	Alta: reproducible automáticamente
VQA-based metrics	Alta: evalúa correspondencia vía preguntas dirigidas	Media: depende del modelo VQA usado
Attribute accuracy	Alta: evalúa atributos específicos	Media-Alta: requiere clasificadores entrenados
Evaluación humana	Muy alta: referencia para intención semántica	Baja: costosa y variable
Distributional diagnostics	Media: detecta tendencias y hallucinations	Media: interpretación parcial necesaria

Los resultados de esta revisión sistemática revelan que el avance en la generación de imágenes a partir de texto ha transitado desde estructuras puramente generativas hacia arquitecturas híbridas que priorizan el control semántico y la fidelidad cultural. En los estudios revisados, se observa que los modelos de difusión han adquirido mayor relevancia en comparación con los VAE y las GAN, debido a su capacidad para capturar detalles finos y lograr una alineación más precisa entre texto e imagen [17].

Algunos estudios, por otro lado, destacan la puesta en práctica de técnicas de atención a nivel de frase y el uso de modelos de lenguaje a gran escala (LLM) para ayudar en la visualización de narrativas complejas, lo que ayuda a que haya más coherencia entre diferentes objetos generados [11,18]. Adicionalmente, se han empleado métodos como la ingeniería de prompts jerárquicos y LoRA para mejorar la exactitud del resultado y la calidad visual, sobre todo en aplicaciones concretas como el patrimonio arquitectónico y el diseño de productos [7,10].

No obstante, a pesar de los avances alcanzados, la literatura especializada sigue identificando diversas limita-

ciones. Entre las más relevantes se encuentran la persistencia de sesgos implícitos heredados de los modelos preentrenados y las dificultades para mantener un control fino durante procesos de edición local sin comprometer la coherencia global de la imagen generada [6, 9]. Además, la generación de contenido altamente especializado, como la teledetección o el arte histórico, se ve restringida por la falta de datos y las particularidades del dominio. En este contexto, algunos estudios

han propuesto estrategias automáticas de depuración de datos, entre ellas los esquemas de doble bucle empleados en teledetección, con el fin de mejorar la validez técnica de los modelos [8,12]. Por esa razón, se recomienda que futuras investigaciones se enfoquen en desarrollar métodos de .ªtención consciente de la relación"(relation-aware) para optimizar la interacción entre varios objetos, así como en elaborar modelos de difusión avanzada que procesen atributos faciales de grano fino de manera más eficiente [19, 20].

A partir del análisis se identifican tendencias claras en la investigación reciente, entre ellas un interés creciente en modelos híbridos que incorporan difusión, aprendizaje multimodal y mecanismos de control explícito mediante señales de identidad o estructurales. Del mismo modo, los sistemas tienden hacia una mayor interpretabilidad y adaptabilidad, al tiempo que ganan la capacidad de equilibrar el control y la creatividad según el contexto de uso [15, 18, 21]. Finalmente, se proyecta un incremento en el desarrollo de modelos de texto e imagen especializados por dominio, junto con la adopción de arquitecturas más ligeras y eficientes, lo que generará nuevas oportunidades para su aplicación práctica en ámbitos como el diseño, el arte digital y la generación de contenido asistida por inteligencia artificial [14, 16, 19].

Conclusiones

La revisión sistemática sugiere que los modelos de difusión son los más utilizados para crear imágenes a partir de texto, ya que tienen la capacidad de combinar mecanismos de control semántico y producir contenido multimodal con gran fidelidad. Estos modelos ofrecen ventajas significativas en comparación con las arquitecturas basadas en VAE y GAN, especialmente en trabajos que requieren precisión cuando se trata de atributos visuales complejos y descripciones textuales. El análisis indica que la implementación de técnicas como adaptación de bajo rango (LoRA), prompt engineering jerárquico y atención a nivel de frase ayuda significativamente en el aumento de la fidelidad visual y la coherencia semántica en distintas áreas de aplicación, entre ellas el arte digital, el diseño de productos y el patrimonio arquitectónico.

Además, se observa que los mecanismos de control utilizados en los modelos de texto e imagen funcionan, sobre todo, a tres niveles: la configuración de la escena, el estilo cultural o artístico y las características visuales. La unión de estos posibilita la producción de imágenes más exactas y personalizadas, sim embargo todavía existen restricciones para el manejo a detalle de objetos individuales sin que esto impacte la coherencia total

de la imagen. También es un desafío

continuar evaluando el alineamiento semántico debido a que las métricas existentes son limitadas y requieren ser complementadas con valoraciones humanas y cambios específicos según el contexto de uso.

También, se lograron identificar limitaciones, fortalezas y tendencias en las investigaciones actuales en cuanto a las arquitecturas y mecanismos existentes. Se percibe existe una tendencia creciente hacia la aplicación de modelos híbridos, adaptables e interpretables que sean capaces de equilibrar el control y la creatividad dependiendo el contexto en el que apliquen. En este sentido, este trabajo resulta útil como una guía para el desarrollo de soluciones nuevas para la generación de imágenes a partir de texto.

Igualmente, en base a los resultados, resulta necesario emplear estrategias más sofisticadas para el control relacional y la atención consciente hacia varios objetos, además de incorporar modelos eficaces y especializados por dominio para mejorar su rendimiento.

Finalmente, se sugieren futuras vías de investigación, tales como el desarrollo de arquitecturas de difusión avanzadas para mejorar la gestión de atributos específicos; la elaboración de métricas e índices estandarizados que permitan calcular y medir la correlación entre texto e imagen; y el análisis de modelos adaptativos que hagan las aplicaciones prácticas más exactas en campos como el diseño, la educación, el arte digital y la producción de contenido con asistencia de la inteligencia artificial. Estas acciones mejorarán la fiabilidad, la creatividad y la habilidad de interpretación en los sistemas texto-imagen, lo que ayudará a progresar en esta nueva área de la inteligencia artificial generativa de imágenes.

Contribución de Autoría

Zaleth Valentina Rivas Calderón: Conceptualización, Investigación, Metodología, Software, Validación, Redacción - borrador original. Estefany Lucia Villanueva Rosales: Conceptualización, Investigación, Metodología, Software, Validación, Redacción - borrador original. Marcelino Torres Villanueva: Análisis formal, Visualización, Supervisión, Administración de proyectos, Curación de datos, Escritura, revisión y edición.

Referencias

[1] J. Xu, J. Du, and J. Wang, "A survey of generative models used in text-to-image," *Applied and Computational Engineering*, vol. 79, pp. 38–48, 2024.

[2] C. Zhang, C. Zhang, M. Zhang, I. S. Kweon, and J. Kim, "Text-to-image Diffusion Models in Generative

AI: A Survey," 2023. [Online]. Available: https://arxiv.org/abs/2303.07909

[3] K. Wang, X. Liu, Y. Chang, D. Zhao, T. Xian, and X. Geng, "Semantic guidance for precise style control in diffusion image generation," *Scientific Reports*, 2025.

[4] R. Li, W. Li, Y. Yang, H. Wei, J. Jiang, and Q. Bai, "Swinv2-Imagen: hierarchical vision transformer diffusion models for text-to-image generation," *Neural Computing and Applications*, vol. 36, pp. 17 245–17 260, 2024.

[5] H. Ma and H. Zheng, "Text Semantics to Image Generation: A Method of Building Facades Design Base on Stable Diffusion Model," in *Phygital Intelligence (CDRF 2023), Computational Design and Robotic Fabrication*, 2024, pp. 24–34.

[6] O. Avrahami, O. Fried, and D. Lischinski, "Blended Latent Diffusion," *ACM Transactions on Graphics*, vol. 42, no. 4, p. art. no. 3592450, 2023.

[7] H. He, H. Yang, Z. Tuo, Y. Zhou, Q. Wang, Y. Zhang, Z. Liu, W. Huang, H. Chao, and J. Yin, "DreamStory: Open-Domain Story Visualization by LLM-Guided Multi-Subject Consistent Diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 12, pp. 11 874–11 891, 2025.

[8] Z. Ye, X. He, and Y. Peng, "RaT2IGen: Relation-aware Text-to-image Generation via Learnable Prompt," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 21, no. 5, p. art. no. 151, 2025.

[9] Z. Kuang, J. Zhang, Y. Li *et al.*, "Preserving architectural heritage in urban renewal: a stable diffusion model framework for automated historical facade generation," *npj Heritage Science*, vol. 13, p. art. no. 256, 2025.

[10] Z. Sordo, E. Chagnon, Z. Hu *et al.*, "Synthetic Scientific Image Generation with VAE, GAN, and Diffusion Model Architectures," *Journal of Imaging*, vol. 11, no. 8, p. art. no. 252, 2025.

[11] M. Gao, Q. Zhang, C. Song, X. Zhang, and Y. Li, "Hierarchical Prompt Engineering and Task-Differentiated Low-Rank Adaptation for Artificial Intelligence-Generated Content Image Quality Assessment," *Information (Switzerland)*, vol. 16, no. 11, p. art. no. 1006, 2025.

[12] J. Zhu and L. Mu, "GrainedCLIP and DiffusionGrainedCLIP: Text-Guided Advanced Models for Fine-Grained Attribute Face Image Processing," *IEEE Access*, vol. 11, pp. 99 030–99 045, 2023.

[13] M. DÍncà, E. Peruzzo, M. Mancini, X. Xu, H. Shi, and N. Sebe, "GradBias: Unveiling Word Influence on Bias in Text-to-Image Generative Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 11, pp. 9863–9875, 2025.

[14] J. Li, S. Zhang, L. Sun *et al.*, "Enhancing product concept image generation through semantic feature prompts and LoRA training," *Scientific Reports*, vol. 15, p. art. no. 40795, 2025.

[15] W. Hu, Y. Zhao, L. Yin *et al.*, "Hierarchical symmetric GAN for Thangka image generation," *npj Heritage Science*, vol. 13, p. art. no. 568, 2025.

[16] N. S. Mudiraj and S. Singh, "Semantic mapping of Hindi text-to-image generation using CUB dataset," *Scientific Reports*, vol. 15, p. art. no. 36632, 2025.

[17] Y. Zhao, Z. Liang, Y. Qiu *et al.*, "A novel flexible identity-net with diffusion models for painting-style generation," *Scientific Reports*, vol. 15, p. art. no. 27896, 2025.

[18] X. Peng, T. Sun, Q. Hu *et al.*, "Poe2CLP: Phrase-level attention and cross-modal semantic alignment for poem generate Chinese landscape paintings," *npj Heritage Science*, vol. 13, p. art. no. 656, 2025.

[19] K. Jung, N. Lee, and S. Choi, "KoDi: A Korean Diffusion Model for Bilingual Text-to-Image Generation and Cultural Fidelity," *IEEE Access*, vol. 13, pp. 200 290–200 307, 2025.

[20] Y. Zhao, M. Li, and M. Berger, "CUPID: Contextual Understanding of Prompt-conditioned Image Distributions," *Computer Graphics Forum*, vol. 43, no. 3, p. art. no. e15086, 2024.

[21] Y. Xu, H. Liu, R. Yang, and Z. Chen, "Remote Sensing Image Semantic Segmentation Sample Generation Using a Decoupled Latent Diffusion Framework," *Remote Sensing*, vol. 17, no. 13, p. art. no. 2143, 2025.

[22] Z. Li, Y. Wang, C. Li *et al.*, "LFMDiff: generation of Chinese traditional landscape paintings based on diffusion model," *npj Heritage Science*, vol. 13, p. art. no. 564, 2025.

[23] T. Xing, H. Yan, X. Wang, K. Sun, H. Yu, P. Li, and Q. Zhao, "DLDC: A Dual Loop Data Cleaning Method for Fine-Tuning Remote Sensing Image Generative Models," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 28 709–28 725, 2025.