Evaluación de diferentes alternativas para la generación de personajes consistentes en estilo con pocos ejemplos

Roberto González[1 [0009−0000−2726−9569]], Rubén Pascual[1 [0009−0002−5250−8554]], Mikel Sesma-Sara[1 [0000−0002−2949−7909]], and Mikel Galar[1 [0000−0003−2865−6549]]

Institute of Smart Cities (ISC) y Departamento de Estadística, Informática y Matemáticas, Universidad Pública de Navarra (UPNA), Campus Arrosadia s/n, 31006, Pamplona, Navarra, España
{roberto.gonzalezmiranda, ruben.pascual, mikel.sesma, mikel.galar}@unavarra.es

Resumen Los modelos de generación de imágenes han acelerado la producción creativa en toda la industria audiovisual; sin embargo, producir un número potencialmente ilimitado de nuevos personajes que preserven el estilo artístico y los rasgos visuales compartidos de un pequeño conjunto de diseños de referencia creados por humanos sigue siendo un desafío persistente. Este trabajo compara los dos enfoques dominantes para esta tarea: el fine-tuning y la generación zero-shot. Partiendo de nuestro método anterior de fine-tuning, DreamBooth multi-token con LoRA, evaluamos su adaptación a una selección de arquitecturas base actualizada y lo comparamos con los principales métodos zero-shot, incluidos IPAdapter y la generación basada en prompts con Grandes Modelos Multimodales (LMMs). A través de una combinación de análisis cuantitativo y una evaluación humana exhaustiva, evaluamos la capacidad de cada método para equilibrar la fidelidad estilística y la diversidad de personajes. Nuestros experimentos revelan un compromiso claro: mientras que los métodos zero-shot ofrecen resultados notables e instantáneos, los enfoques de fine-tuning mantienen una ventaja clara en la obtención de una fidelidad estilística superior, junto con una mayor diversidad creativa.

Keywords: Generación de personajes · Consistencia de estilo · Pocos ejemplos (few-shot) · Fine-tuning · Grandes modelos multimodales (LMMs).

1. Introducción

Los modelos generativos han evolucionado de proyectos de investigación a componentes centrales del software creativo convencional. Herramientas como Generative Fill de Adobe permiten a los artistas realizar tareas complejas como la eliminación de objetos o la extensión de imágenes (outpainting) con prompts sencillos [3], mientras que otras apoyan la generación rápida de arte conceptual y recursos [2]. Aunque estas herramientas pueden producir imágenes de alta calidad, no están diseñadas para aprender un estilo coherente a partir de unos pocos

ejemplos ni para generar nuevos personajes ilimitados que sigan consistentemente dicho estilo. Este trabajo aborda ambos desafíos mediante el aprendizaje del estilo artístico y los rasgos visuales compartidos de un pequeño conjunto de referencias diseñadas por humanos para crear personajes novedosos ilimitados, como se ilustra en la Fig. 1.

Esta tarea es intrínsecamente desafiante. El estilo objetivo suele ser sutil y difícil de describir con un prompt de texto, lo que requiere que el modelo lo infiera a partir de un conjunto limitado de ejemplos, un problema few-shot. Además, más allá de mantener el estilo, los nuevos personajes deben incorporar elementos novedosos para evitar ser meras réplicas de las referencias. Finalmente, la evaluación objetiva resulta difícil debido a la limitación de los datos de referencia y a la naturaleza subjetiva de la evaluación estilística.



Figura 1. La tarea de generación de personajes con consistencia de estilo a partir de pocos ejemplos. Izquierda: dieciséis personajes de referencia creados por un artista que definen el estilo objetivo. Derecha: ejemplos de personajes recién generados que conservan los rasgos estilísticos e introducen diseños novedosos.

Entre las diferentes alternativas para abordar estos desafíos, destacan dos aproximaciones comunes en la literatura actual.

Una de ellas es el fine-tuning [30], donde los pesos del modelo preentrenado se adaptan utilizando las imágenes de referencia para codificar explícitamente el estilo objetivo. Los métodos eficientes en parámetros, como DreamBooth [26] y la adaptación de bajo rango (LoRA) [14], adaptan eficientemente modelos grandes y se han convertido en herramientas estándar de personalización. Nuestro trabajo previo [23] introdujo un método novedoso: un enfoque DreamBooth multi-token combinado con LoRA, y demostró que separar los tokens para la identidad del personaje y el estilo artístico puede producir resultados estilísticamente cohesivos pero diversos.

Otra aproximación es la generación zero-shot, que elimina por completo el entrenamiento. Enfoques como IPAdapter [32] integran características visuales

de las imágenes de referencia directamente en el proceso de difusión, mientras que los Grandes Modelos Multimodales (LMMs) [12,21] utilizan el aprendizaje en contexto para inferir y reproducir el estilo a partir de las imágenes de referencia. Estos métodos prometen una adaptabilidad inmediata, pero su capacidad para preservar la estructura estilística o sostener la diversidad creativa sigue siendo incierta.

A falta de una comparación directa entre ambos paradigmas, este trabajo evalúa el compromiso entre la personalización profunda del fine-tuning y la rapidez y accesibilidad de los métodos zero-shot para generar personajes consistentes en estilo. De este modo, se proporciona información valiosa para artistas, desarrolladores e investigadores, guiando la elección entre priorizar la fidelidad estilística y la diversidad frente a la velocidad y la usabilidad.

2. Trabajo relacionado

Esta sección revisa la evolución de los modelos generativos y los principales enfoques para la personalización del estilo, enmarcando el contexto para nuestro estudio comparativo.

Evolución de los métodos de generación de imágenes Tras el éxito de las GANs [11,16], el campo ha convergido en los modelos de difusión [13,7] por su calidad y flexibilidad. Arquitecturas de pesos abiertos como Stable Diffusion [25,8] y FLUX [4], junto con modelos recientes como GPT Image 1 [21] y Nano Banana [12], definen el estado del arte. Sin embargo, reproducir un estilo artístico lleno de matices a partir de pocos ejemplos sigue siendo un desafío abierto.

Fine-tuning para la personalización del estilo El fine-tuning adapta modelos preentrenados para aprender un estilo de forma reutilizable. Trabajos pioneros proponen desde aprender nuevos embeddings (Textual Inversion [10]) o ajustar capas de atención (CustomDiffusion [18]), hasta actualizar el modelo completo (DreamBooth [26]). Métodos recientes como StyleDrop [27] o Ada-Adapter [20] se enfocan en la transferencia de estilo pura. Ante el coste de actualizar todo el modelo, las técnicas eficientes en parámetros (PEFT) [30] como LoRA [14] son el estándar actual. Nuestro trabajo previo [23] combina DreamBooth y LoRA en un esquema multi-token para desacoplar la identidad del personaje de su estilo artístico.

Condicionamiento de estilo zero-shot Los métodos zero-shot prescinden del entrenamiento adaptándose instantáneamente a nuevas entradas visuales. Mientras que técnicas como ControlNet [33] aportan guía estructural y STAM [9] realiza transferencia de estilo uno a uno, nuestro foco es generar contenido libre y novedoso a partir de referencias. En esta línea, IPAdapter [32] destaca al inyectar las imágenes mediante atención cruzada desacoplada como prompts visuales. Recientemente, los LMMs [12,21,31] se han posicionado como potentes generadores zero-shot; mediante aprendizaje en contexto, estos modelos infieren el estilo objetivo y guían la generación de personajes sin modificar sus pesos.

Evaluación del arte generativo Evaluar cualidades subjetivas como la coherencia de estilo es complejo. Métricas de distribución como FID [15] son poco fiables con pocos ejemplos, mientras que CLIP-I [26] no logra desacoplar estilo y contenido. Para solucionarlo, propuestas recientes introducen métricas especializadas como la fidelidad [23], que usa un codificador específico de estilo [28], y la diversidad [23], adaptada del análisis de textos [19] para medir la variación en el espacio latente. Con todo, las métricas automáticas siguen siendo limitadas, por lo que la evaluación humana [22] resulta indispensable.

3. Métodos

Comparamos dos aproximaciones para la generación de personajes consistentes en estilo a partir de un pequeño conjunto de referencias: fine-tuning y métodos zero-shot.

3.1. Enfoque de fine-tuning

Nuestro enfoque DB^{L}_{MT} [23] utiliza una estrategia multi-token en el marco de DreamBooth [26] con LoRA [14] para desacoplar el estilo de la identidad del personaje.

Durante el entrenamiento, a cada uno de los N personajes de referencia se le asigna un token único ($\texttt{[c}_1\texttt{]}$, $\texttt{[c}_2\texttt{]}$, ..., $\texttt{[c}_N\texttt{]}$) para aprender características específicas del personaje, mientras que se utiliza un token compartido (`[S]`) en todas las imágenes para capturar el estilo artístico. Los prompts de entrenamiento tienen el formato "$\texttt{[c}_i\texttt{]}$`, [S] style`". Esta separación permite al modelo aprender una representación generalizada del estilo, que luego puede aplicarse para generar personajes completamente nuevos.

Consideramos dos estrategias para seleccionar los tokens específicos:

- DB^{L}_{MT-C}: Tokens seleccionados mediante k-means sobre embeddings de CLIP para maximizar la separación semántica, como en [23].
- DB^{L}_{MT-R}: Tokens elegidos según su rareza en el vocabulario del modelo.

Para la generación de nuevos personajes, el token de estilo compartido `[S]` se combina con un token aleatorio no utilizado extraído del vocabulario.

3.2. Enfoque zero-shot

Consideramos dos enfoques representativos zero-shot: IPAdapter y LMMs basados en prompts.

IPAdapter IPAdapter [32] inyecta características visuales de las imágenes de referencia directamente en las capas de atención cruzada del modelo de difusión congelado, permitiendo que actúen como un "prompt visual" guiando el estilo del resultado.

Generación basada en prompts con LMMs Los LMMs [12,21,31] analizan las imágenes de referencia proporcionadas utilizando sus codificadores de visión, infieren el estilo objetivo mediante el aprendizaje multimodal en contexto y luego guían un módulo interno de generación de imágenes para producir nuevos personajes consistentes con ese estilo.

4. Configuración experimental

Esta sección detalla el marco de trabajo para nuestro estudio comparativo (conjunto de datos, configuración de modelos, pipeline de generación, métricas de evaluación y metodologías de evaluación humana).

4.1. Conjunto de datos

Todos los experimentos emplean el conjunto de datos cartoon *scary* [1] (16 personajes, Fig. 1). Este dataset ofrece un escenario few-shot realista al combinar una fuerte coherencia de estilo con rasgos complejos imposibles de describir puramente con texto.

4.2. Configuración de los modelos

Esta sección detalla las elecciones específicas de implementación para cada uno de los métodos evaluados en nuestro estudio.

Modelos de fine-tuning Evaluamos dos variantes de nuestro enfoque de fine-tuning en diferentes arquitecturas:

- DB^{L}_{MT-C}: La variante basada en clustering de [23], aplicada a Stable Diffusion 1.5 [25]. Esta sirve como nuestra línea base establecida, representando la metodología original en una arquitectura de codificador único donde el clustering es una estrategia efectiva de selección de tokens.
- DB^{L}_{MT-R}: La variante basada en rareza de [23], aplicada a Stable Diffusion 3.5 Medium [8] y FLUX.1-dev [4]. Estos modelos emplean múltiples codificadores de texto, donde las distancias de los embeddings no son directamente comparables, de ahí la elección de tokens basados en rareza.

Modelos zero-shot Para los métodos zero-shot, se proporcionan cuatro imágenes de referencia seleccionadas empíricamente. En IPAdapter (utilizando el modelo `ip-adapter-plus_sd15` [32] en ComfyUI [6]), las características visuales extraídas con CLIP ViT-H [24] se inyectan con peso máximo (1.0) mediante atención cruzada desacoplada sin prompt de texto. Para preservar un fondo neutro, la latente inicial se deriva de una imagen real. Para los LMMs (GPT Image 1 [21] y Nano Banana [12]), se utiliza el prompt en inglés:

"Generate a character that preserves the artistic style, color palette, brushstrokes, facial expression style, and shared visual traits of the reference characters. The image should be completely new and original, not a copy, but fully consistent with the style, color palette, expressions, and shared visual traits of the reference characters."

4.3. Generación de imágenes y filtrado

Para cada configuración, generamos 250 muestras iniciales y aplicamos un filtrado de cuatro etapas [23] para excluir resultados inválidos o sesgados:

1. **Copias de entrenamiento:** Identificadas mediante similitud CLIP-I [26] respecto a las imágenes reales y descartadas si superan un umbral para evitar la memorización.
2. **Imágenes defectuosas:** Descartadas si su puntuación de fidelidad (ver sección 4.4) está por debajo de un umbral aceptable contra el conjunto de referencia.
3. **Sujetos múltiples:** Descartados usando el Segment Anything Model (SAM) [17] y posterior verificación manual.
4. **Salidas duplicadas:** Detectadas mediante el Índice de Similitud Estructural (SSIM) [29] para evitar el colapso de modo.

Solo las imágenes que superan todas las comprobaciones se utilizan para las evaluaciones cuantitativa y humana.

4.4. Métricas de evaluación cuantitativa

Utilizamos dos métricas de [23] diseñadas para la evaluación del estilo y la variación de personajes:

Fidelidad Mide la similitud de estilo entre las imágenes generadas (V) y las de referencia (D). Se calcula mediante la similitud coseno promedio por pares entre sus respectivos embeddings, v_i y d_j, de acuerdo con (1):

$$\text{Fidelity}(V, D) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \text{máx}(\cos(v_i, d_j), 0), \tag{1}$$

donde un mayor valor indica una mejor correspondencia estilística.

Diversidad Mide la variedad de los personajes generados. Se calcula mediante la media geométrica de las desviaciones estándar de las componentes de sus embeddings sobre su dimensionalidad (H), según se define en (2):

$$\text{Diversity}(V) = \sqrt[H]{\prod_{i=1}^{H} \sigma_i}, \tag{2}$$

donde σ_d representa la desviación estándar de la componente d-ésima de los embeddings en el conjunto V.

4.5. Evaluación humana

Para complementar las métricas cuantitativas [22], realizamos dos estudios de evaluación humana con 98 participantes (3399 respuestas en total). Ambos fueron diseñados para medir la preferencia general del usuario y diagnosticar la fidelidad estilística. Las interfaces de usuario para los estudios se muestran en la Fig. 2.

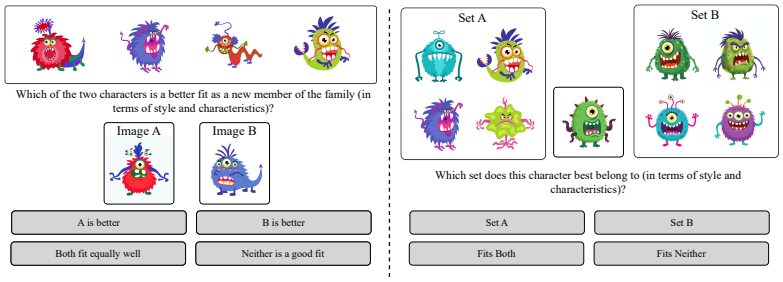


Figura 2. Diagrama de las UI para nuestros estudios de evaluación humana. Izquierda: Prueba de preferencia basada en comparaciones, en la que el usuario selecciona qué personaje (Image A o Image B) coincide mejor con el estilo de la galería de referencia. Derecha: Prueba de fidelidad estilística frente a consistencia del modelo, donde un participante debe decidir si un personaje generado es estilísticamente más cercano al dataset de referencia (Set A) o a otras imágenes de su propio modelo (Set B).

Clasificación de preferencias basada en comparaciones Este estudio establece una clasificación de preferencias. En cada prueba, los usuarios seleccionaron cuál de dos imágenes generadas coincidía mejor con una galería de referencia dinámica. La galería se construyó a partir de 12 de las 16 imágenes de entrenamiento, seleccionando los personajes estilísticamente más similares a cada candidato. Esto evita el sesgo de galerías estáticas, asegurando que se juzgue la coherencia estilística en lugar de los rasgos superficiales. Las preferencias se modelaron mediante Bradley-Terry [5], incluyendo las cuatro imágenes del dataset de referencia excluidas como control.

Fidelidad de estilo frente a consistencia del modelo Este estudio evalúa la fidelidad de un modelo al estilo objetivo frente a su propia consistencia interna. En cada prueba, se clasificó una imagen generada como más cercana al conjunto de referencia (los cuatro personajes reales más similares) o al conjunto del propio modelo (cuatro imágenes generadas al azar). Las respuestas se agregaron como una puntuación de adherencia a la referencia: porcentaje de pruebas donde el usuario asoció la imagen al estilo real (“Referencia” o “Ambas”), penalizando modelos que consolidan un estilo propio erróneo.

5. Resultados y discusión

Esta sección presenta un análisis exhaustivo de nuestros hallazgos experimentales. El rendimiento cuantitativo de cada método se resume en la Tabla 1. Los resultados de nuestros dos estudios de evaluación humana se detallan en la Tabla 2 (clasificación de preferencias) y la Tabla 3 (consistencia al estilo de referencia). Finalmente, se proporciona una comparación cualitativa de las imágenes generadas en la Tabla 4.

Cuadro 1. Comparación cuantitativa de los métodos de fine-tuning y zero-shot para la generación de personajes consistentes en estilo. Los mejores valores se destacan en **negrita**.

Fine-tuning	Método	Modelo	Fidelidad ↑	Diversidad ↑
Sí	DB^{L}_{MT-C}	SD 1.5	0.7600	0.2349
Sí	DB^{L}_{MT-R}	SD 3.5M	0.7869	**0.2461**
		FLUX.1-dev	**0.8133**	0.2039
No	IPAdapter	SD 1.5	0.7468	0.2394
No	Prompt	GPT Image 1	0.7754	0.2092
		Nano Banana	0.7991	0.2067

5.1. Métricas cuantitativas

El fine-tuning resulta la aproximación más sólida. La variante DB^{L}_{MT-R} con FLUX.1-dev alcanza la mayor fidelidad (0.8133), validando que un entrenamiento específico garantiza una reproducción estilística óptima. Asimismo, su aplicación en SD 3.5M logra la mayor diversidad (0.2461) manteniendo una fidelidad alta (0.7869), ofreciendo el mejor equilibrio global.

Los LMMs zero-shot ofrecen resultados competitivos. Nano Banana logra una fidelidad notablemente alta (0.7991), superando a varios modelos. Sin embargo, este rendimiento se acopla con una baja diversidad (0.2067). La inspección cualitativa sugiere que los LMMs tienden a reinterpretar el estilo objetivo a través de fuertes prioris estilísticos, produciendo imágenes de alta calidad e internamente coherentes que, no obstante, convergen hacia un estilo erróneo.

Se observa un patrón complementario con IPAdapter. Alcanza una diversidad relativamente alta (0.2394) pero muestra la fidelidad más baja entre todos los métodos, lo que indica que su mayor variabilidad se produce a expensas de la preservación del estilo.

5.2. Resultados de la evaluación humana

Clasificación de preferencias basada en comparaciones Los resultados (Tabla 2) muestran que ambas variantes de DB^{L}_{MT-R} superan al conjunto de

referencia. Esto indica que los modelos capturan un "estilo central" generalizado, generando personajes que se perciben como más representativos que algunas muestras reales que exhiben mayor variabilidad.

Cuadro 2. Clasificación de preferencias basada en comparaciones de la evaluación humana.

Fine-tuning	Método	Modelo	Clasificación
Sí	DB^{L}_{MT-R}	SD 3.5M	1 (1093)
		FLUX.1-dev	1 (1093)
No	Referencia	*(Dataset scary)*	2 (1060)
No	Prompt	Nano Banana	3 (1022)
		GPT Image 1	4 (978)
Sí	DB^{L}_{MT-C}	SD 1.5	5 (920)
No	IPAdapter	SD 1.5	6 (830)

Los LMMs (Nano Banana y GPT Image 1) también logran valoraciones altas debido a su consistencia y alta calidad estética libre de artefactos, aunque dicha calidad no se traduzca necesariamente en fidelidad estilística, como se analiza en la siguiente sección.

Fidelidad de estilo frente a consistencia del modelo Los resultados del segundo estudio (Tabla 3) confirman esta discrepancia. Los modelos fine-tuneados (SD 3.5M y FLUX.1-dev) obtienen la mayor adherencia a la referencia (58.46 % y 45.02 %), validando que se perciben alineados con el estilo objetivo.

Cuadro 3. Resultados de la evaluación humana para la fidelidad de estilo, mostrando la adherencia a la referencia obtenida por cada modelo.

Entrenamiento	Método	Modelo	Adherencia a la ref.	N.º de eval.
Sí	DB^{L}_{MT-C}	SD 1.5	32.68 %	257
Sí	DB^{L}_{MT-R}	SD 3.5M	58.46 %	260
		FLUX.1-dev	45.02 %	291
No	IPAdapter	SD 1.5	27.24 %	279
No	Prompt	GPT Image 1	18.57 %	280
		Nano Banana	19.78 %	268

En cambio, los métodos zero-shot muestran una fuerte consistencia interna pero una alineación débil con el estilo real. GPT registra una baja adherencia a

la referencia (16.28 %), evidenciando su tendencia a consolidar un estilo artístico divergente (Tabla 4).

Cuadro 4. Ejemplos de imágenes generadas con la mayor y menor fidelidad para cada método de generación.

Métodos de fine-tuning									
	DB^{L}_{MT-C} (SD 1.5)			DB^{L}_{MT-R} (SD 3.5M)			DB^{L}_{MT-R} (FLUX.1-dev)		
Mejor fidelidad									
	0.8236	0.8212	0.8171	0.8314	0.8311	0.8259	0.8472	0.8440	0.8439
Peor fidelidad									
	0.5679	0.6537	0.6597	0.6775	0.6824	0.6877	0.7497	0.7538	0.7585
Métodos zero-shot									
	IPAdapter (SD 1.5)			GPT Image 1			Nano Banana		
Mejor fidelidad									
	0.8190	0.8185	0.8167	0.8281	0.8276	0.8265	0.8358	0.8318	0.8287
Peor fidelidad									
	0.6194	0.6370	0.6416	0.6811	0.6910	0.6929	0.7438	0.7500	0.7550

6. Conclusiones

En este trabajo, comparamos el fine-tuning y métodos zero-shot para la generación de personajes con consistencia de estilo few-shot. Contrastamos el enfoque DreamBooth multi-token con LoRA frente a alternativas como IPAdapter y LMMs (GPT Image 1 y Nano Banana) mediante análisis cuantitativo y una detallada evaluación humana.

Los hallazgos confirman que, aunque las aproximaciones zero-shot han avanzado, el fine-tuning sigue siendo la opción más sólida para internalizar una representación de estilo fiel y diversa. Los LMMs, por su parte, destacan en accesibilidad e ideación rápida sin necesidad de entrenamiento. Como trabajo futuro, se planea explorar esquemas híbridos integrando nuestra estrategia multi-token con técnicas de adaptación rápida como Ada-Adapter [20], así como realizar evaluaciones humanas con mayor variedad de estilos artísticos.

Agradecimientos. Rubén Pascual recibió apoyo de una ayuda FPU (ref. FPU22/02961) otorgada por el Ministerio de Ciencia, Innovación y Universidades de España (MICIU). Este trabajo también fue financiado parcialmente por el MICIU bajo los

proyectos PID2022-136627NB-I00 (MCIN/AEI/10.13039/501100011033/FEDER, UE) y PID2025-174319OB-I00 (MICIU/AEI/10.13039/501100011033). Y por el Gobierno de Navarra bajo los proyectos INVERSO (0011-1365-2022-000130) y Genesis (PC24-GENESIS-041-002). Fuente del conjunto de datos: Freepik.

Referencias

1. Scary dataset. https://www.freepik.com/free-vector/scary-monster-set-colored-monsters-with-teeth-eyes-illustration-funny-monsters_13031939.htm
2. Adobe: Adobe Firefly. https://www.adobe.com/products/firefly.html
3. Adobe: The Next Generation of Generative AI is now in Photoshop. https://blog.adobe.com/en/publish/2024/04/23/the-next-generation-generative-ai-now-in-photoshop
4. black-forest-labs: Flux. black-forest-labs (2025)
5. Bradley, R.A., Terry, M.E.: Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. Biometrika **39**(3/4), 324–345 (1952). https://doi.org/10.2307/2334029
6. comfyanonymous: ComfyUI, https://github.com/comfyanonymous/ComfyUI
7. Dhariwal, P., Nichol, A.: Diffusion Models Beat GANs on Image Synthesis. In: Adv. Neural Inf. Process. Syst. vol. 34, pp. 8780–8794. Curran Associates, Inc. (2021)
8. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Rombach, R.: Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In: International Conference on Machine Learning (2024)
9. Fahim, M.A.N.I., Saqib, N., Boutellier, J.: STAM: Zero-Shot Style Transfer using Diffusion Model via Attention Modulation. In: Conference on Computer Vision and Pattern Recognition. pp. 6387–6397 (2025)
10. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion (2022). https://doi.org/10.48550/arXiv.2208.01618
11. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. In: Adv. Neural Inf. Process. Syst. vol. 27. Curran Associates, Inc. (2014)
12. Google: Introducing Gemini 2.5 Flash Image, our state-of-the-art image model. https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/
13. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: Adv. Neural Inf. Process. Syst. vol. 33, pp. 6840–6851. Curran Associates, Inc. (2020)
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models. CoRR **abs/2106.09685** (2021)
15. Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., Kumar, S.: Rethinking FID: Towards a Better Evaluation Metric for Image Generation. In: Conference on Computer Vision and Pattern Recognition. pp. 9307–9315 (2024)
16. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training Generative Adversarial Networks with Limited Data. In: Adv. Neural Inf. Process. Syst. vol. 33, pp. 12104–12114. Curran Associates, Inc. (2020)
17. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment Anything. In: International Conference on Computer Vision. pp. 4015–4026 (2023)

18. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-Concept Customization of Text-to-Image Diffusion. In: Conference on Computer Vision and Pattern Recognition. pp. 1931–1941 (2023)
19. Lai, Y.A., Zhu, X., Zhang, Y., Diab, M.T.: Diversity, Density, and Homogeneity: Quantitative Characteristic Metrics for Text Collections. CoRR **abs/2003.08529** (2020)
20. Liu, J., Li, C., Sun, Q., Ming, J., Fang, C., Wang, J., Zeng, B., Liu, S.: Ada-adapter:Fast Few-shot Style Personlization of Diffusion Model with Pre-trained Image Encoder (2024). https://doi.org/10.48550/arXiv.2407.05552
21. OpenAI: Introducing GPT-5. https://openai.com/index/introducing-gpt-5/ (2025)
22. Otani, M., Togashi, R., Sawai, Y., Ishigami, R., Nakashima, Y., Rahtu, E., Heikkilä, J., Satoh, S.: Toward Verifiable and Reproducible Human Evaluation for Text-to-Image Generation. In: Conference on Computer Vision and Pattern Recognition. pp. 14277–14286 (2023)
23. Pascual, R., Sesma-Sara, M., Jurio, A., Paternain, D., Galar, M.: Few-shot multi-token DreamBooth with LoRa for style-consistent character generation. Expert Syst. Appl. **309**, 131226 (2026). https://doi.org/10.1016/j.eswa.2026.131226
24. 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.: Learning Transferable Visual Models From Natural Language Supervision. In: Proc. 38th Int. Conf. Mach. Learn. pp. 8748–8763. PMLR (2021)
25. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis With Latent Diffusion Models. In: Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
26. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In: Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)
27. Sohn, K., Ruiz, N., Lee, K., Chin, D.C., Blok, I., Chang, H., Barber, J., Jiang, L., Entis, G., Li, Y., Hao, Y., Essa, I., Rubinstein, M., Krishnan, D.: StyleDrop: Text-to-Image Generation in Any Style (2023). https://doi.org/10.48550/arXiv.2306.00983
28. Somepalli, G., Gupta, A., Gupta, K., Palta, S., Goldblum, M., Geiping, J., Shrivastava, A., Goldstein, T.: Measuring Style Similarity in Diffusion Models (2024). https://doi.org/10.48550/arXiv.2404.01292
29. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
30. Xu, L., Xie, H., Qin, S.Z.J., Tao, X., Wang, F.L.: Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment (2023). https://doi.org/10.48550/arXiv.2312.12148
31. Xu, Y., Wang, W., Zhang, Y., Tang, B., Yan, P., Feng, F., He, X.: Personalized Image Generation with Large Multimodal Models. In: Proc. ACM Web Conf. 2025. pp. 264–274. WWW '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3696410.3714843
32. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models (2023). https://doi.org/10.48550/arXiv.2308.06721
33. Zhang, L., Rao, A., Agrawala, M.: Adding Conditional Control to Text-to-Image Diffusion Models. In: International Conference on Computer Vision. pp. 3836–3847 (2023)