DreamBooth multi-token con LoRA para la generación de personajes con estilo consistente mediante pocos ejemplos

Ruben Pascual[1 [0009−0002−5250−8554]], Mikel Sesma-Sara[1 [0000−0002−2949−7909]], Aranzazu Jurio[1 [0000−0002−4087−586X]], Daniel Paternain[1 [0000−0002−5845−887X]], and Mikel Galar[1 [0000−0003−2865−6549]]

Institute of Smart Cities (ISC) and Department of Statistics, Computer Science and Mathematics, Public University of Navarre (UPNA), Campus Arrosadia, 31006 Pamplona, España {ruben.pascual, mikel.sesma, aranzazu.jurio, daniel.paternain, mikel.galar}@unavarra.es

Keywords: Modelos de difusión · DreamBooth · Generación de personajes · Few-shot.

1. RESUMEN

Este trabajo ha sido publicado originalmente bajo el título "Few-shot multi-token DreamBooth with LoRa for style-consistent character generation" (https://doi.org/10.1016/j.eswa.2026.131226) en la revista *Expert Systems with Applications*.

En los últimos años, la Inteligencia Artificial ha cambiado la industria audiovisual, afectando tanto a los flujos de trabajo creativos como a los técnicos. En el ámbito de los videojuegos, herramientas impulsadas por IA como NVIDIA DLSS [5] mejoran el renderizado en tiempo real, mientras que en el cine y la televisión, los modelos de lenguaje de gran escala (LLMs) como Dramatron [6] asisten en la escritura de guiones. A pesar de estos progresos, la creación de personajes únicos que mantengan una consistencia estilística sigue siendo un desafío complejo sin resolver. Si bien se han desarrollado diversas aproximaciones para automatizar este flujo de trabajo [4, 8], actualmente no existe una solución efectiva cuando se pretende generar nuevos personajes partiendo de un conjunto de referencia muy reducido.

Esto se debe principalmente a la dificultad de este problema que radica en varios factores. En primer lugar, la mayoría de estilos artísticos no pueden ser descritos con precisión mediante texto, lo que obliga a los modelos a aprender exclusivamente de ejemplos visuales. Además, el estilo y el contenido se encuentran profundamente entrelazados en las imágenes; el estilo no es un rasgo localizado, sino una distribución global compleja de texturas, iluminación y patrones. Asimismo, en un entorno *few-shot* [10], típicamente limitado a 15-30 imágenes, la escasez de datos impide que el modelo separe las reglas estilísticas de los detalles específicos del personaje. Esto provoca que los métodos de aprendizaje convencionales sufran de sobreajuste, memorizando los sujetos de entrenamiento

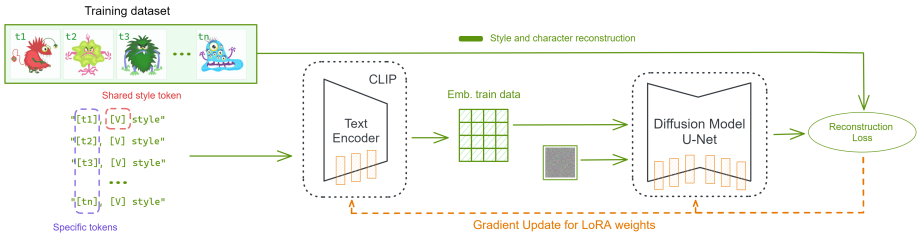


Figura 1. Diagrama del entrenamiento Multi-Token DreamBooth con LoRA. Se asignan tokens individuales ($[t_1], [t_2], \dots$) a cada personaje para capturar su identidad, mientras que el token compartido ($[V]$) se utiliza en todas las imágenes para aprender el estilo global.

en lugar de generalizar la estética. Por último, la evaluación de los resultados es compleja, ya que la subjetividad y la limitación de los datos dificultan que las métricas estándar [3, 12] capturen variaciones estilísticas sutiles.

Como respuesta a este desafío, los modelos de difusión fundacionales, como Flux [1] o Stable Diffusion [7], resultan prometedores al permitir el ajuste fino a partir su gran conocimiento previo. Sin embargo, lograr un control preciso sobre sus resultados es complejo. Para ello se han desarrollado diversas técnicas de personalización [2, 11] entre las que destaca DreamBooth [9], siendo uno de los métodos más utilizados al permitir el aprendizaje de nuevos conceptos de forma sencilla. No obstante, a pesar de su efectividad en ciertos contextos, estos métodos a menudo requieren descripciones textuales explícitas, refinamientos iterativos o condicionamiento externo, lo que los hace menos adecuados para generar personajes ilimitados y consistentes a partir de pocas imágenes.

En este trabajo proponemos una modificación de DreamBooth [9] denominada Multi-Token DreamBooth con LoRA. Nuestra propuesta introduce una estrategia de desentrelazamiento donde se asignan tokens específicos (*[specific_id]*) a cada personaje individual para capturar sus rasgos distintivos, junto a un token compartido (*[shared_id]*) encargado de aprender la estética global del conjunto. Con ello, buscamos separar el aprendizaje del estilo del resto de características de los personajes. Además, eliminamos el conjunto de regularización de clase para priorizar una adaptación profunda al estilo y reducir la interferencia de conocimientos previos no deseados; este proceso de entrenamiento se ilustra en la Figura 1. Con el fin de habilitar la creación automatizada de personajes ilimitados, exploramos tres estrategias de generación que prescinden de prompts manuales: el uso de tokens aleatorios y el muestreo de embeddings basado en distribuciones gaussianas univariables y multivariables. Estas técnicas permiten generar personajes nuevos y visualmente consistentes al integrar identidades latentes inéditas con el estilo aprendido, logrando un sistema escalable.

Para evaluar la efectividad de nuestra propuesta, hemos llevado a cabo un extenso estudio experimental utilizando cinco conjuntos de datos con estilos visuales diferenciados (Virus, Scary, Daily, Hipster y Trans). En primer lugar, el

análisis cuantitativo mediante la métrica de fidelidad adaptada y el estudio de diversidad confirman que el método Multi-Token DreamBooth genera personajes nuevos y visualmente consistentes, destacando el muestreo multivariable como la estrategia más robusta. Esta validez técnica se fundamenta, en última instancia, en un estudio humano a gran escala que comprende 11.140 comparaciones por pares realizadas por 375 participantes generales, junto con 300 valoraciones detalladas de seis artistas profesionales. Los resultados demuestran que los evaluadores integran con éxito los personajes generados en las familias de referencia, superando significativamente a las aproximaciones convencionales.

Acknowledgments. Rubén Pascual recibió apoyo de una beca FPU (ref. FPU22/02961) otorgada por el Ministerio de Ciencia e Innovación. Este trabajo también fue financiado parcialmente por el Ministerio de Ciencia e Innovación bajo el proyecto PID2022-136627NB-I00 (MCIN/AEI/10.13039/501100011033/FEDER, UE) y el proyecto PID2025-174319OB-I00 (MICIU/AEI/10.13039/501100011033) y por el Gobierno de Navarra bajo los proyectos 0011-1365-2022-000130 (INVERSO) y PC24-GENESIS-041-002.

Bibliografía

[1] Black Forest Labs: FLUX (2025)

[2] 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. In: Int. Conf. Learn. Represent. (2022)

[3] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In: Adv. Neural Inf. Process. Syst. vol. 30. Curran Associates, Inc. (2017)

[4] Leigh, M.: The Sims 4 Prima Guide. Prima Games (2014)

[5] Lin, H., Burnes, A.: NVIDIA DLSS 4 Introduces Multi Frame Generation & Enhancements For All DLSS Technologies (2025)

[6] Mirowski, P., Mathewson, K.W., Pittman, J., Evans, R.: Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In: Conf. Hum. Factors Comput. Syst. pp. 1–34. CHI '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3544548.3581225

[7] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models. In: IEEE Conf. Comput. Vis. Pattern Recognit. pp. 10674–10685 (2022). https://doi.org/10.1109/CVPR52688.2022.01042

[8] Ruan, S.: Anime Characters Generation with Generative Adversarial Networks. In: 2022 IEEE Int. Conf. Adv. Electr. Eng. Comput. Appl. AEECA. pp. 1332–1335 (2022). https://doi.org/10.1109/AEECA55500.2022.9918869

[9] 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: Conf. Comput. Vis. Pattern Recognit. CVPR. pp. 22500–22510 (2023). https://doi.org/10.1109/CVPR52729.2023.02155

[10] Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Adv. Neural Inf. Process. Syst. pp. 4080–4090. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)

[11] 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. ArXiv (2023). https://doi.org/10.48550/ARXIV.2306.00983

[12] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In: Conf Comput Vis Pattern Recognit CVPR. pp. 586–595 (2018). https://doi.org/10.1109/CVPR.2018.00068