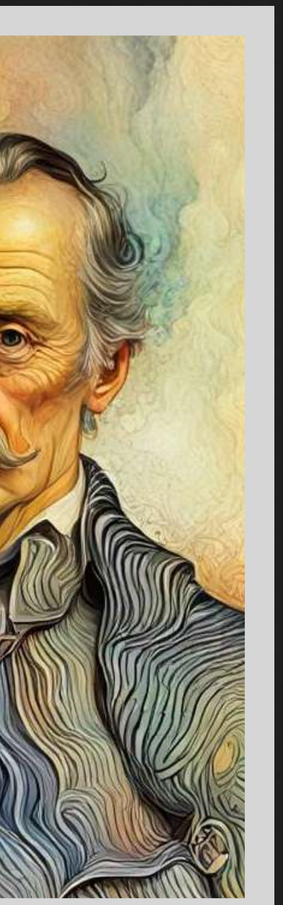
CÁLCULO DE SIMILITUD ENTRE IMÁGENES ARTÍSTICAS Y GENERADAS POR STABLE DIFFUSION UTILIZANDO REDES CONVOLUCIONALES SIAMESAS

Navil Pineda Rugerio, Diego Castro Elvira
Ingeniería en Inteligencia Artificial

Dr. Ricardo Ramos Aguilar

Dr. Jesús García Ramírez



CREACIÓN DEL CONJUNTO DE DATOS

1

Conjunto de datos WikiArt de obras artisticas originales, con un preprocesamiento de eliminación de datos nulos y redimensionamiento.

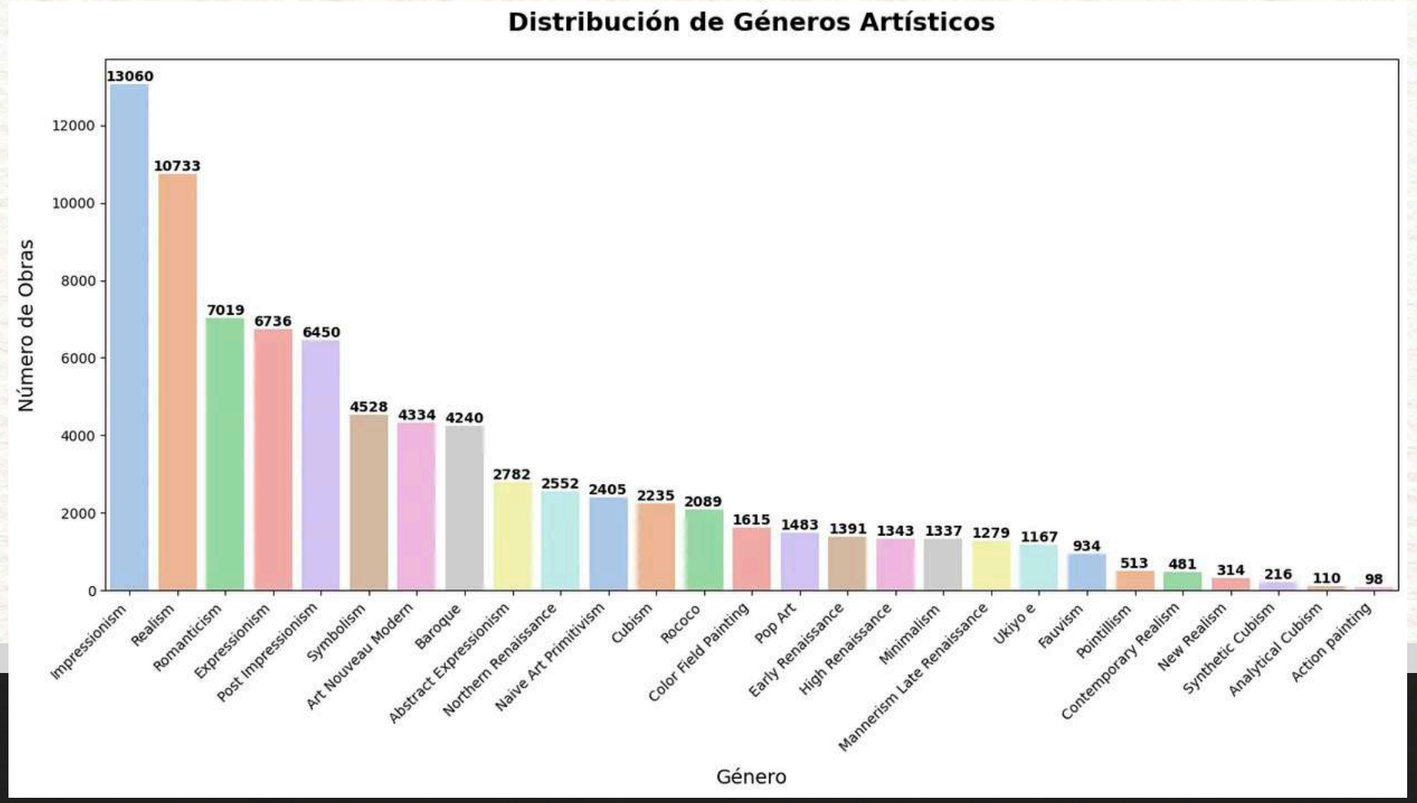
2

Conjunto de datos de imágenes generadas por Stable Diffusion XL Refiner 1.0.

Análisis del conjunto de datos original WikiArt

Total de datos del conjunto original: 81,445

De los cuales, se tenían las siguientes cantidades de imágenes de diferntes géneros artísticos.



Preprocesamiento de WikiArt: Eliminación de datos nulos.

Se eliminaron las filas con datos vacíos, es decir, si no tenía autor, género artístico o el *subset* de datos que pertenece (train o test).

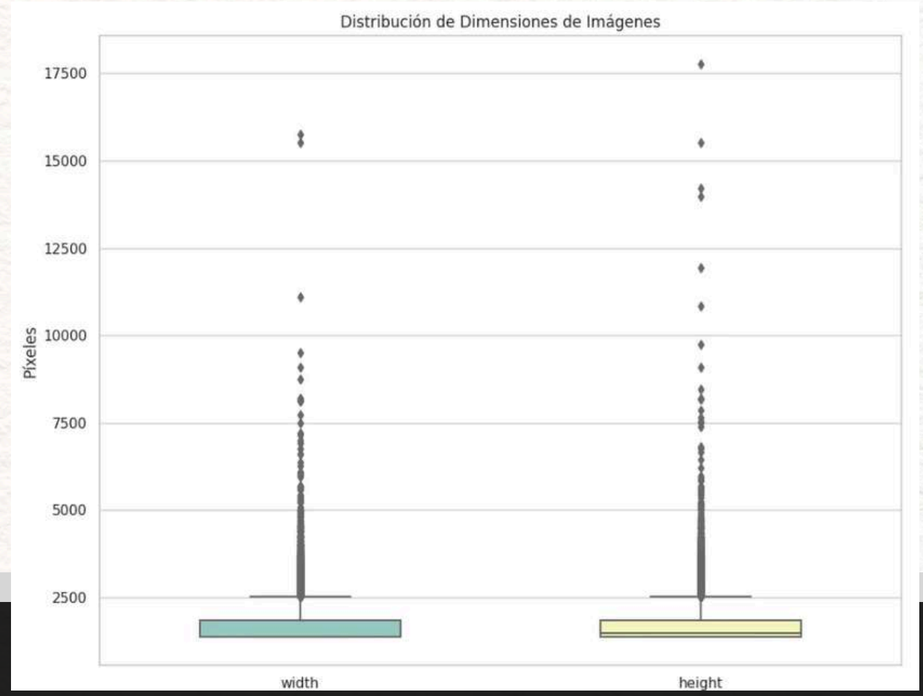
Datos eliminados por cada columna:

Columna	No. de datos eliminados
subset	2097
genre	2097
artist	2097

Preprocesamiento de WikiArt: Redimensionamiento.

De acuerdo con la documentación oficial de Stable Diffusion, el modelo trabaja mejor con imágenes de 1024x1024, pero también tiene soporte para imágenes de 768x768

Las imágenes originales tenían la siguiente distribución de dimensiones, por lo que se decidió que era posible transformarlas a 1024x1024 sin afectar su calidad significativamente.



Creación de conjunto de imágenes generadas por Stable Diffusion

Se necesita una imagen de referencia (imagen original de obra artística) y un prompt. Para ello se realizaron los siguientes procesos:

1. Generación de descripciones textuales de imágenes originales.
2. Utilizar las descripciones textuales para generar prompts personalizados para cada imagen.
3. Pasar la imagen original y su prompt personalizado para generar la imagen con Stable Diffusion

Resultados

Generación de descripciones textuales de imágenes originales.

Se utilizó el modelo BLIP-2, que combina visión y lenguaje para interpretar imágenes y generar un texto descriptivo, tales como los siguientes.

“a painting of three boats in a harbor “
(una pintura de tres barcos en un puerto)

“a painting of a house and trees in a field“
(una pintura de de una casa y árboles en un campo)

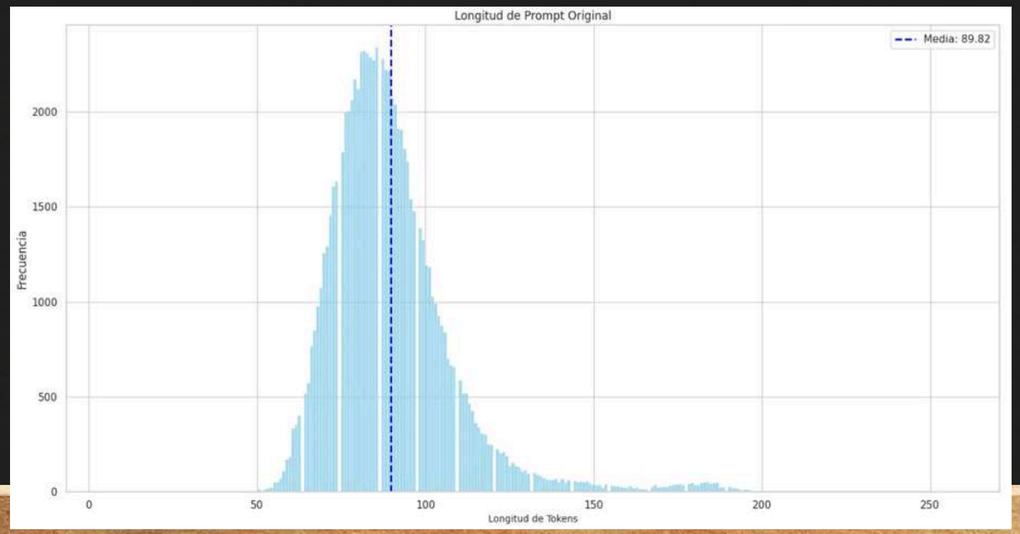
“a painting of two women in a kitchen with a barrel“
(una pintura de dos mujeres en una cocina con un barril)

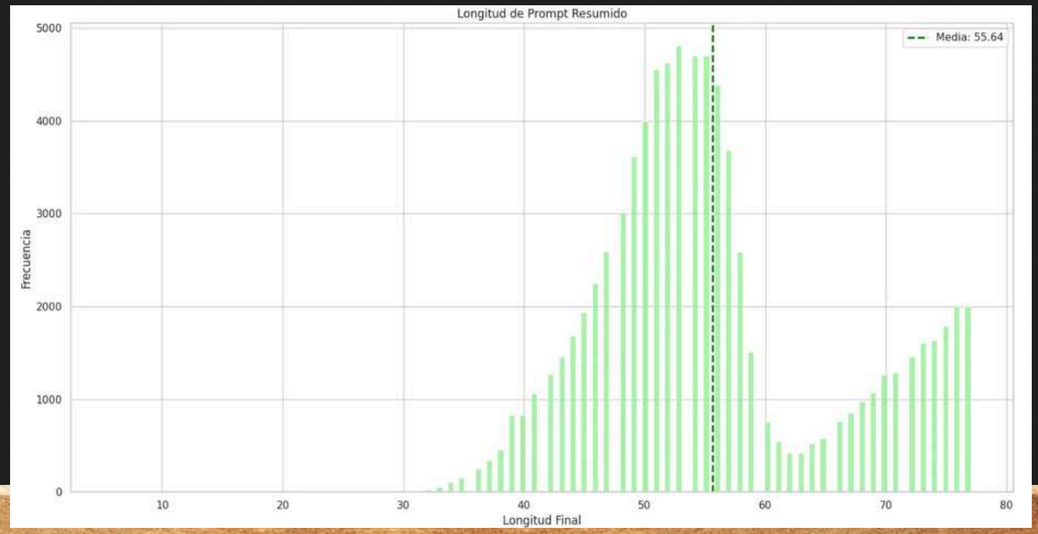
Resultados

Prompts personalizados.

Se crearon prompts con dos niveles de transformación: moderado (cambios sutiles en color, iluminación, estilo) y radical (transformaciones más profundas).

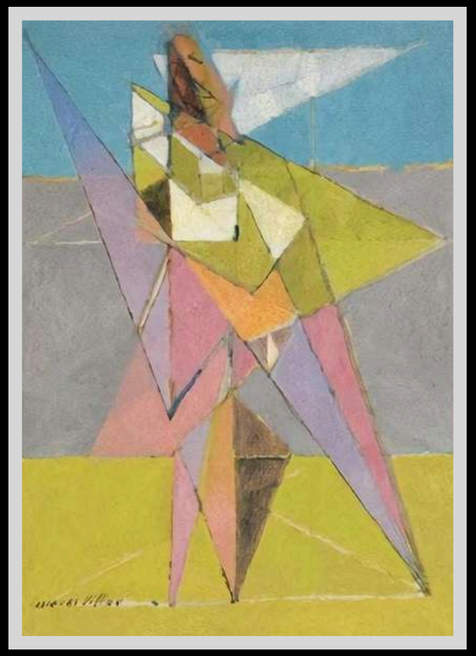
Se resumieron los prompts a un promedio de 55 tokens ya que Stable Diffusion acepta máximo 75, y en un inicio la media de tokens era de 89, tal como se muestra en las gráficas.





Resultados

Generación de imágenes con Stable Diffusion pasándole la imágen original y el prompt personalizado
Se muestra un ejemplo de transformación radical.



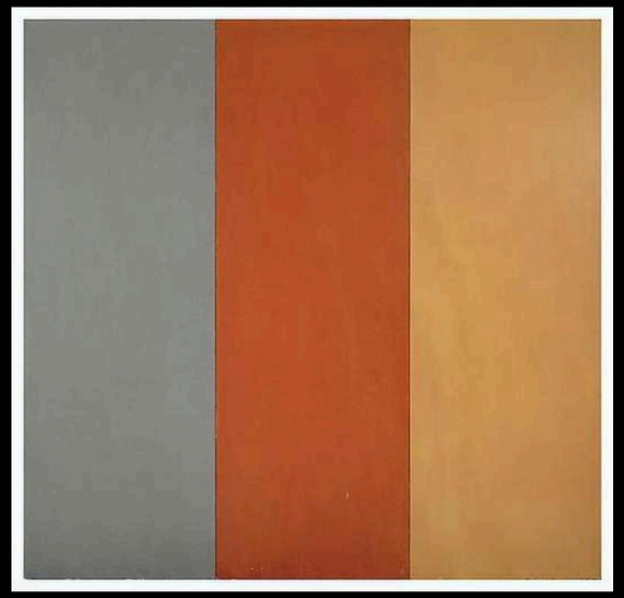
+

“19th-century Baroque-inspired, ornate, gilded, and intricately patterned tapestry. reimagining subjects as majestic, regal, and mythological creatures, with iridescent, shimmering scales and delicate, lacy wings.”

=

Resultados

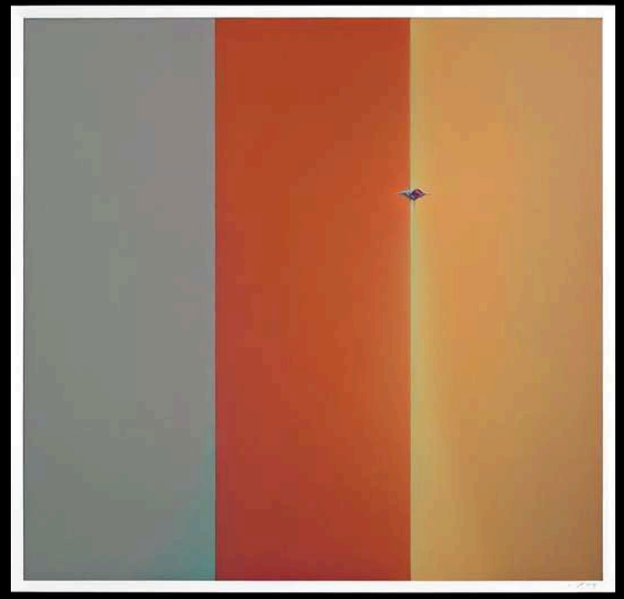
Ejemplo de transformación moderada.



+

A kaleidoscopic, geometric, and biomorphic form, suspended in mid-air, surrounded by a halo of soft, pulsing light, as if it's alive. colors shifting between turquoise, amethyst, and rose, like a living, breathing entity.

=



Entrenamiento de Red Siamesa

1

Se reciben tres imágenes de entrada:

- Ancla: Imagen artística original (de WikiArt).
- Positiva: Imagen generada por Stable Diffusion.
- Negativa: Imagen aleatoria de otro estilo o categoría.

2

Extracción de características, para lo que se utiliza CLIP (clip-vit-base-patch32) preentrenado. Cada imagen pasa por CLIP, generando un embedding de 512 dimensiones.

3

Tras obtener los embeddings de CLIP, se agregan capas adicionales:

- Capas lineales densas (FC Layers).
- Batch Normalization y Dropout para regularización.
- Función de activación ReLU o similar.

Se obtiene un embedding refinado de la misma dimensión (512)

Entrenamiento de Red Siamesa

1

- Se calcula la similitud del coseno entre:
  - Ancla ↔ Positiva (sim_pos).
  - Ancla ↔ Negativa (sim_neg).

2

Se usa la pérdida basada en tripletes, para evaluar el modelo:

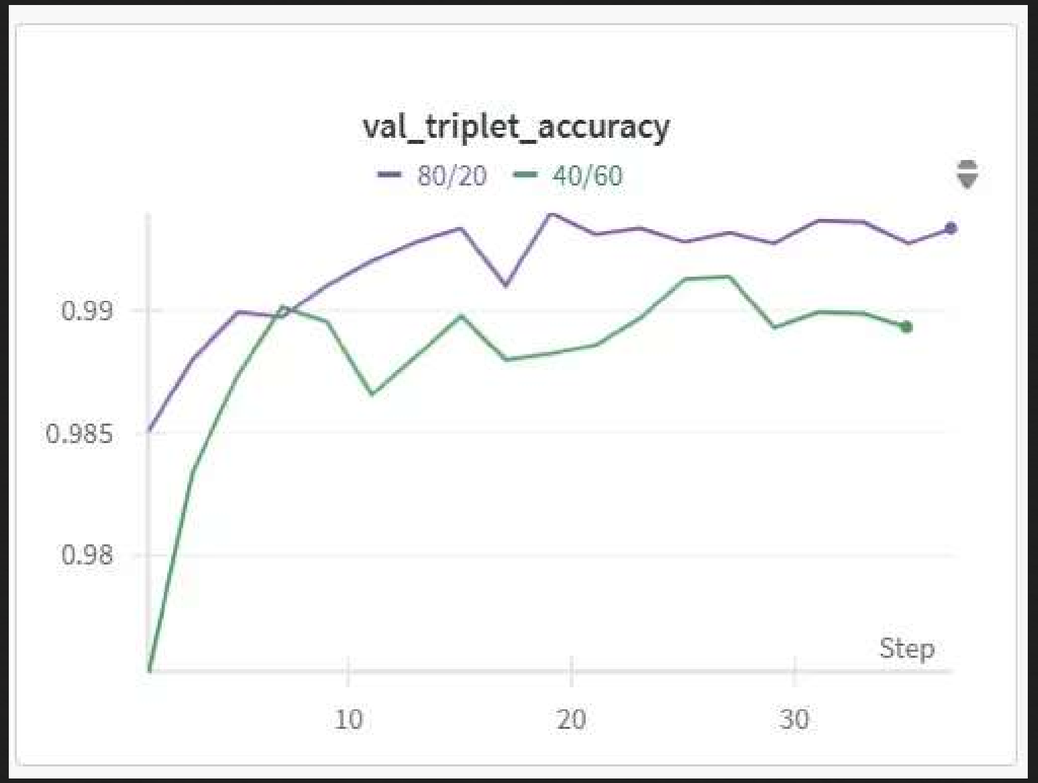
$$loss = -\log\left(\frac{\exp(sim\_pos/\tau)}{\exp(sim\_pos/\tau) + \exp(sim\_neg/\tau)}\right)$$

El objetivo es maximizar la similitud entre imágenes similares y minimizar la similitud con imágenes distintas.

Resultados de métricas de evaluación de entrenamiento

Precisión de tripletes

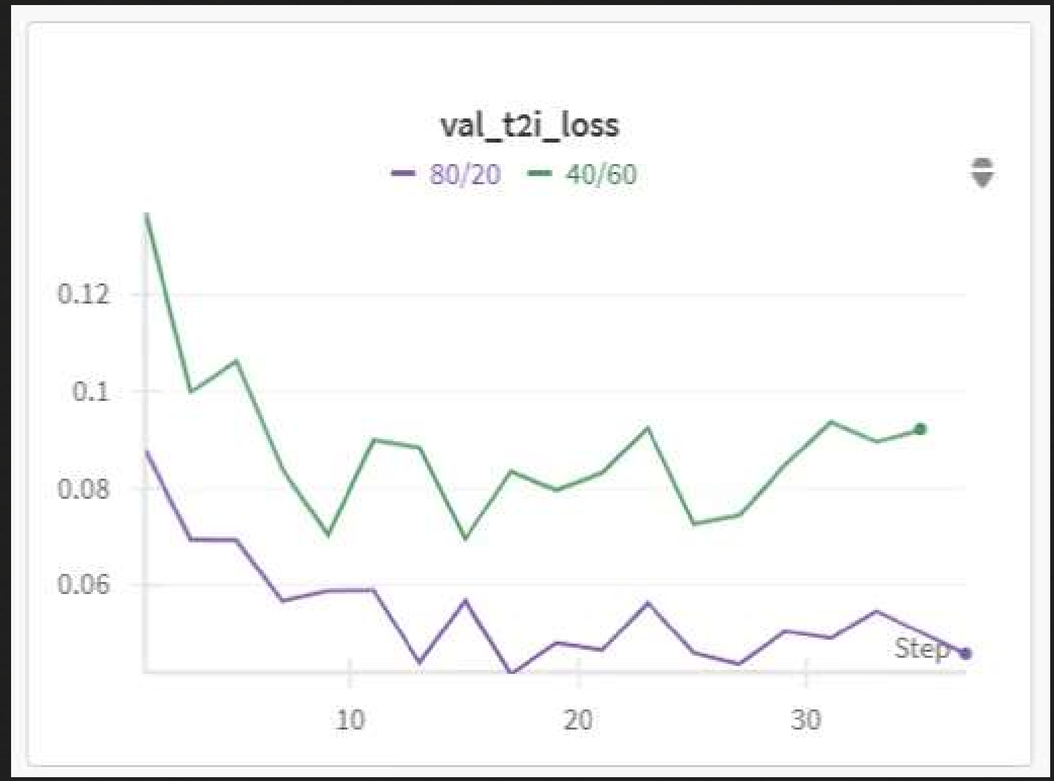
- Ambas configuraciones comienzan con una precisión cercana a 0.98.
- La línea morada alcanza una precisión máxima de casi 1.0 en algunas etapas, mientras que la línea verde presenta más variabilidad pero converge hacia 0.99.
- La configuraciones 80/20 es más estable, tiene una precisión más alta, mientras que la verde muestra fluctuaciones en los primeros pasos de entrenamiento.



Resultados de métricas de evaluación de entrenamiento

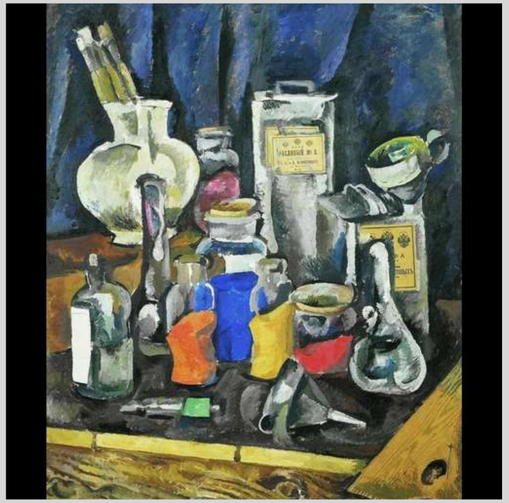
Precisiones de texto-imagen

- La pérdida disminuye significativamente en los primeros pasos, con la línea morada convergiendo a un valor más bajo (~0.046), mientras que la línea verde tiene más variabilidad antes de estabilizarse en las últimas etapas.
- En general, la configuración 80/20 tiene un rendimiento más estable.



Pruebas preliminares del modelo

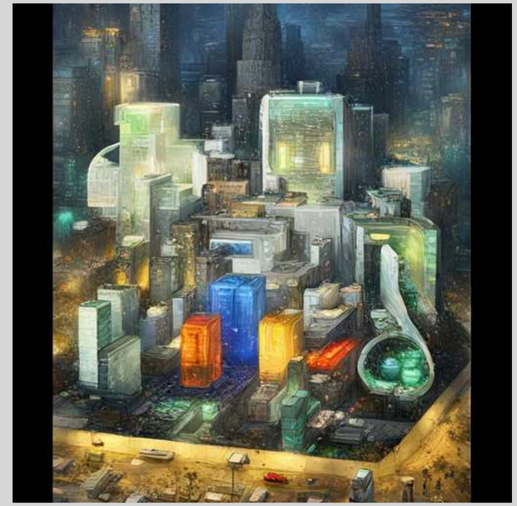
Cubism/pyotr-konchalovsky_dry-paints-1913



Descripción: a painting of a table with various items on it.

vs

Cubism/pyotr-konchalovsky_dry-paints-1913



Descripción: a painting of various antique items on a table.

Valor de similitud coseno:

0.8117

Pruebas preliminares del modelo

Minmalism/frank-stella_ifafa-i-from-the-v-series-1968

Descripción: a purple and blue print with the letter v.

vs

Minimalism/frank-stella_ifafa-ii-1967

Descripción: a green and black geometric pattern on a white background

Valor de similitud coseno:

0.7133

Pruebas preliminares del modelo

Minmalism/frank–stella_ifafa–i–from–the–v–series–1968

Descripción: a painting of a table with various items on it.

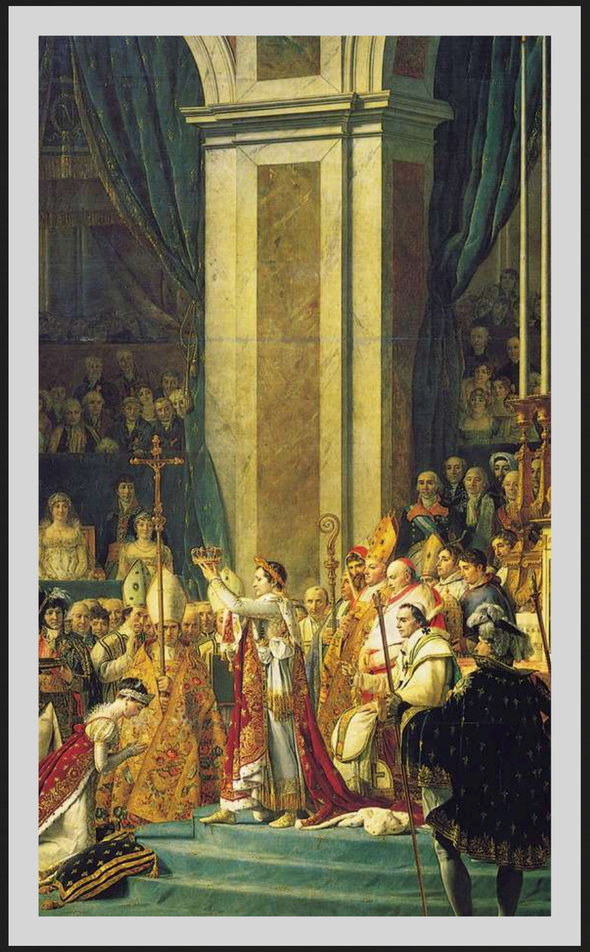
vs

Impressionism/william–merritt–chase_still–life–with–cockatoo

Descripción: a white bird is perched on a tall building.

Valor de similitud coseno:

0.1594



Gracias