CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

Análise da Geração de Imagens a Partir da Descrição em Texto de Cenas Utilizando *Stable Diffusion*

Jorge Gomes de Melo Júnior

João Pessoa, PB

2023

Jorge Gomes de Melo Júnior

Análise da Geração de Imagens a Partir da Descrição em Texto de Cenas Utilizando *Stable Diffusion*

Monografia apresentada ao curso Ciência da Computação
do Centro de Informática, da Universidade Federal da Paraíba,
como requisito para a obtenção do grau de Bacharel em Ciência da Computação

Orientadora: Thaís Gaudencio do Rêgo

Junho de 2023

Catalogação na publicação
Seção de Catalogação e Classificação

M528a Melo Júnior, Jorge Gomes de.
Análise da geração de imagens a partir da descrição em texto de cenas utilizando stable diffusion / Jorge Gomes de Melo Júnior. - João Pessoa, 2023.
42f. : il.

Orientação: Thaís Gaudencio do Rêgo, Yuri De Almeida Malheiros Barbosa.
TCC (Graduação) - UFPB/CI.

1. Modelos generativos. 2. Stable diffusion. 3. Visão computacional. 4. Inteligência artificial. I. Rêgo, Thaís Gaudêncio do. II. Barbosa, Yuri De Almeida Malheiros. III. Título.

UFPB/CI CDU 004.8

Elaborado por Michelle de Kássia Fonseca Barbosa - CRB-738

"A boniteza da vida está na certeza da incerteza e na coragem de começar tudo de novo quando se pensava que já nada podíamos fazer."

Paulo Freire

AGRADECIMENTOS

Este trabalho é fruto do apoio, carinho e orientação de diversas pessoas excepcionais. Gostaria de expressar minha profunda gratidão aos meus pais, pilares fundamentais da minha vida. Agradeço por todo o empenho e os sacrifícios realizados para que eu pudesse alcançar esta conquista. Agradeço também por cada palavra de alento, pela atenção e carinho que sempre me proporcionaram.

À minha querida irmã Natalie, um agradecimento especial. Sua presença constante e apoio, manifestados de tantas formas diferentes, foram a força que impulsionou cada passo meu. Todas as minhas conquistas também são suas, e a dívida de gratidão que tenho com você não pode ser expressa em palavras.

Aos professores Thaís Gaudencio, Leonardo Vidal e Yuri De Almeida, meu profundo agradecimento. Mais do que mentores, vocês foram faróis a me guiar nesta jornada, oferecendo ensinamentos, apoio e conselhos inestimáveis. A paciência e a segurança que vocês transmitiram foram meu porto seguro durante os desafios na universidade. O espelho do profissional e da pessoa que almejo ser reflete a imagem de vocês. Agradeço pela confiança e por sempre me incentivarem a superar minhas dificuldades. Vocês são minha inspiração!

Por fim, aos amigos que conquistei durante a universidade, meu caloroso agradecimento. Em especial à Aldo, Anderson, Arnor, Dandara e Vinicius. O companheirismo de vocês nessa jornada foi uma dádiva, tornando meus dias mais leves e agradáveis. As memórias que construímos juntos serão sempre um lembrete do quão gratificante foi esta etapa da minha vida.

Minha gratidão a todos vocês, que tornaram possível a realização deste trabalho.

RESUMO

O campo da geração de imagens sintéticas, utilizando Inteligência Artificial (IA), teve avanços significativos nos últimos anos, com contribuições notáveis de modelos como o DALL-E, *MidJourney*, *Stable Diffusion* e entre outros. Este trabalho se propõe a explorar o atual estado da arte na geração de imagens, a partir de descrições em linguagem natural, enfatizando as técnicas empregadas e abordagens adotadas na área. O *Stable Diffusion*, popular Modelo de Difusão Latente e referência no âmbito *Open Source*, que se destaca por realizar a geração de imagens sintéticas em um espaço latente, com baixo tempo de inferência e custo computacional, é o principal foco deste estudo. Os experimentos conduzidos apresentaram variações de itens a serem gerados, quantidade e cor, inicialmente com descrições textuais elementares, como "*dog*"e "*cat*", e avançando para cenários mais detalhados, como "*ten red dogs and ten blue cats*". Essa variação permitiu uma análise qualitativa aprofundada do impacto das descrições de cenas nos resultados obtidos pelo modelo *Stable Diffusion*. O estudo identificou desafios significativos na área, especialmente na otimização das descrições de cenas para os modelos generativos. A descoberta das melhores práticas para a formulação dos *prompts* é um processo em evolução, e é crucial para atingir os resultados esperados nas imagens geradas. Outra questão relevante é a limitação do modelo em produzir imagens realistas e fidedignas aos detalhes requisitados, quando apresentado com *prompts* que contém um número elevado de objetos a serem representados. A pesquisa conclui que, a seleção adequada das descrições em texto é essencial para orientar o processo de geração de imagens e alcançar os resultados desejados. Entretanto, apesar dos avanços significativos, o campo ainda demanda pesquisas adicionais para superar esses desafios e melhorar a qualidade das imagens geradas.

Palavras-chave: Modelos Generativos, Stable Diffusion, Visão Computacional, Inteligência Artificial.

ABSTRACT

The field of synthetic image generation, using Artificial Intelligence (AI), has made significant advances in recent years, with notable contributions from models such as DALL-E, *MidJourney*, *Stable Diffusion*, and others. This work proposes to explore the current state of the art in image generation from natural language descriptions, emphasizing the techniques employed and approaches adopted in the area. The *Stable Diffusion*, a popular Latent Diffusion Model and reference in the *Open Source* scope, which stands out for performing synthetic image generation in a latent space, with low inference time and computational cost, is the main focus of this study. The conducted experiments presented variations of items to be generated, quantity, and color, initially with elementary textual descriptions, such as "*dog*"and "*cat*", and advancing to more detailed scenarios, such as "*ten red dogs and ten blue cats*". This variation allowed for a deep qualitative analysis of the impact of scene descriptions on the results obtained by the *Stable Diffusion* model. The study identified significant challenges in the area, especially in optimizing scene descriptions for generative models. The discovery of best practices for formulating *prompts* is an evolving process and is crucial to achieving the expected results in the generated images. Another relevant issue is the model's limitation in producing realistic images faithful to the requested details when presented with *prompts* that contain a high number of objects to be represented. The research concludes that the appropriate selection of text descriptions is essential to guide the image generation process and achieve the desired results. However, despite significant advances, the field still demands additional research to overcome these challenges and improve the quality of the generated images.

Key-words: Generative Models, Stable Diffusion, Computer Vision, Artificial Intelligence.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

DM - *Diffusion Models* - (do português, Modelos de Difusão)

LDM - *Latent Diffusion Models* - (do português, Modelos de Difusão)

IA - Inteligência Artificial

GAN - Generative Adversarial Network - (do português, Redes Adversárias Generativas)

PLN - Processamento de Linguagem Natural

Sumário

1 INTRODUÇÃO

A Visão Computacional é um campo de estudo da Ciência da Computação voltado para o processamento e análise de imagens digitais, com o objetivo de permitir que os computadores interpretem imagens de maneira análoga aos seres humanos. Essa área é empregada em diversas aplicações, tais como reconhecimento de pessoas e objetos, vigilância por câmeras, robótica e análise de imagens médicas.

Recentemente, uma das subáreas da Visão Computacional que apresentou notável evolução foi a geração de imagens a partir de descrições textuais. Tal subárea investiga a criação de imagens sintéticas por meio de modelos generativos. Esses algoritmos de Aprendizado de Máquina são treinados com vastos conjuntos de imagens e suas respectivas descrições em linguagem natural, o que possibilita o aprendizado da estrutura e dos padrões presentes nos dados e, consequentemente, a produção de imagens sintéticas de objetos, cenários e seres a partir de texto.

A geração de imagens com o uso de modelos generativos é um campo em constante desenvolvimento, no qual novas técnicas e aplicações vêm sendo criadas. A habilidade de gerar imagens realistas automaticamente, a partir de descrições em linguagem natural, tem o potencial de transformar diversos setores, como a indústria do entretenimento, a medicina e a publicidade.

Entretanto, a geração de imagens por meio de IA ainda demanda avanços na elaboração das descrições das cenas para os modelos generativos, visto que nem todas as nuances configuráveis que podem aprimorar os resultados obtidos são conhecidas. A seleção adequada das descrições em texto (*prompts*) é essencial para orientar o processo de geração de imagens e alcançar os resultados almejados.

1.1 Contextualização

A evolução da geração de imagens sintéticas utilizando utilizando Redes Neurais teve início em 2014, com a introdução do modelo Generative Adversarial Network (GAN) [8] por Goodfellow et al. Em 2015, o engenheiro do Google Alexander Modvintsev desenvolveu o software *DeepDream*[1], que utilizava redes neurais convolucionais para gerar imagens. Embora os resultados fossem limitados, as conclusões de Modvintsev foram importantes para impulsionar a área. Em 2016, a empresa NVIDIA desenvolveu o *StyleTransfer* [12], um software capaz de gerar imagens realistas de rostos humanos. No mesmo ano, foi apresentado o BigGAN [2], uma versão aprimorada do GAN que permitia a geração de imagens de alta resolução.

[1]Disponível em: https://github.com/google/deepdream

O ano de 2021 foi marcado por importantes avanços na geração de imagens sintéticas. A instituição OpenAI divulgou os modelos DALL-E [17] e CLIP [16], que permitem a geração de imagens a partir de descrições em linguagem natural, possibilitando maior controle sobre os resultados obtidos. Além disso, surgiram novos modelos, como o *Big Sleep*[2], VQGAN + CLIP [4], *Disco Diffusion*[3] e GLIDE [14].

Em 2022, outros modelos foram desenvolvidos, como o DALL-E 2[4] e o *Stable Diffusion*[5], evidenciando o contínuo avanço na área. Embora os avanços na geração de imagens sintéticas utilizando IA tenham sido significativos, ainda há desafios a serem superados, principalmente no que diz respeito à compreensão de como utilizar a tecnologia de forma mais eficiente para obter resultados ainda melhores.

Nesse sentido, o objetivo do presente trabalho é analisar os impactos da alteração das descrições em texto utilizadas como entrada nos modelos de geração de imagens, de modo a compreender de que forma as mudanças nas descrições podem influenciar o resultado final das imagens geradas. Espera-se que os resultados obtidos possam contribuir para o aprimoramento da área e para a utilização mais eficiente dessa tecnologia tão promissora.

1.1.1 Objetivo Geral

Analisar qualitativamente a geração de imagens sintéticas a partir de descrições textuais, produzidas pelo modelo generativo *Stable Diffusion*, e examinar o impacto de alterações no texto das descrições das cenas nas imagens resultantes.

1.1.2 Objetivos Específicos

Para alcançar o objetivo geral proposto, os seguintes objetivos específicos devem ser cumpridos:

- Examinar se o acréscimo na quantidade de itens descritos na cena afeta o resultado da imagem gerada.
- Verificar se o aumento da complexidade da descrição da cena em linguagem natural influencia o resultado da imagem gerada.
- Investigar se a alteração na ordem da descrição da cena em linguagem natural impacta no resultado da imagem gerada.

[2]Disponível em: https://github.com/lucidrains/big-sleep
[3]Disponível em: https://github.com/alembics/disco-diffusion
[4]Disponível em: https://openai.com/dall-e-2
[5]Disponível em: https://github.com/CompVis/stable-diffusion

1.2 Estrutura Da Monografia

Esta monografia encontra-se organizada em seis capítulos que buscam refletir a respeito do processo adotado durante o desenvolvimento deste trabalho. No decorrer do primeiro capítulo, são apresentadas a introdução geral da monografia, a definição do problema, as premissas e hipóteses levantadas para a solução do problema e os objetivos gerais e específicos do trabalho.

Os capítulos seguintes serão estruturados da seguinte forma: no segundo capítulo serão introduzidos conceitos para fornecer um embasamento teórico sobre o problema e solução apresentados. O terceiro capítulo abordará trabalhos relacionados que serviram de base para a análise realizada. No quarto capítulo, será apresentada a metodologia aplicada no desenvolvimento e validação do estudo efetuado. No quinto capítulo, serão apresentados os resultados obtidos na pesquisa e conclusões alcançadas a partir destes. Por fim, o último capítulo apresenta as considerações finais e questionamentos suscitados a partir desta pesquisa, que podem resultar na fundamentação de trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo proporcionará o embasamento teórico essencial para a compreensão deste trabalho. Inicialmente, será abordada a arquitetura U-Net [19], conhecida pela sua eficácia na segmentação de imagens. Posteriormente, será apresentada uma descrição dos Modelos de Difusão [5] (do inglês, *Diffusion Models* - DMs), elucidando o seu processo de modelagem generativa de dados. Em sequência, serão expostos os Modelos de Difusão Latente [18] (do inglês, *Latent Diffusion Models* - LDMs), que aprimoram a estrutura dos DMs ao processar os dados em um espaço latente. Por último, será apresentado o *Stable Diffusion*[6], um popular LDM, referência no âmbito *Open Source* e considerado o estado da arte atual em geração de imagens sintéticas, a partir de uma descrição em linguagem natural.

2.1 U-Net

A arquitetura U-Net [19], desenvolvida por Olaf Ronneberger et. al (2015), foi inicialmente aplicada na segmentação de imagens biomédicas, e se destaca pela capacidade de generalizar domínios de baixa variância, ou seja, conjuntos de dados nos quais as imagens apresentam pequenas diferenças entre si. Esta notável competência permite que a U-Net capture detalhes sutis das imagens, facilitando o aprendizado mesmo com quantidades limitadas de dados, e essa característica de aprendizado a tornou amplamente aplicável em diversos campos de estudo.

A estrutura da U-Net (Figura 1) é organizada em duas vias principais, que combinam para formar um design geral em forma de "U". Inicialmente, a via de contração é empregada para capturar características contextuais e de detalhes da imagem de entrada. Essa fase de contração utiliza máscaras convolucionais e operações de discretização para gerar uma representação condensada dos dados. Ao longo deste processo, a dimensionalidade dos dados é progressivamente reduzida, e o número de canais de características é dobrado após cada passo, permitindo a captura de características mais complexas à medida que o processo se aprofunda.

[6]Disponível em: https://github.com/CompVis/stable-diffusion

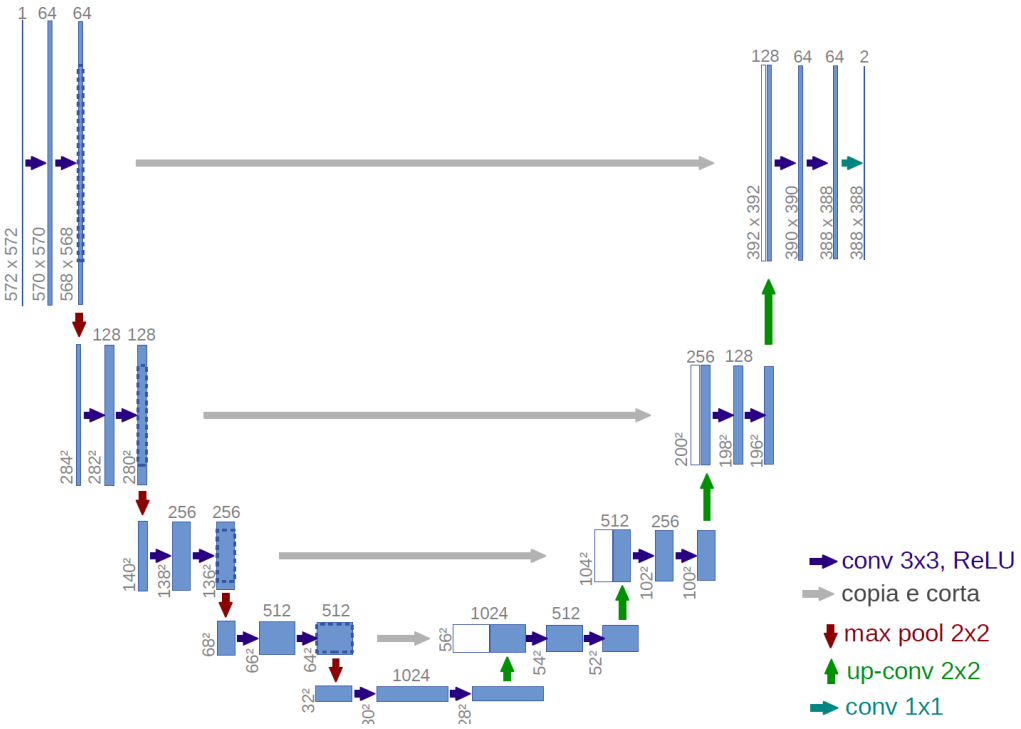


Figura 1: Arquitetura da U-Net.
Fonte: Ronneberger et al. (2015)

Na sequência, a via de expansão é aplicada de forma simétrica à contração. Esta segunda fase da U-Net é projetada para garantir a utilização dos padrões já identificados pela rede para a reconstrução precisa do resultado. Nesta etapa, as operações de *upsampling* são utilizadas para aumentar a resolução dos mapas de características. Essas operações são complementadas por convoluções, que servem para refinar a resolução espacial das saídas. Um elemento crucial na U-Net é a presença de conexões de salto (do inglês, *skip connections*) entre as camadas correspondentes na via de contração e de expansão. Essas conexões permitem que a rede propague informações de contexto e de detalhes localizados para camadas de resolução superior, facilitando a reconstituição detalhada dos padrões na saída.

A arquitetura em questão representa uma inovação significativa na segmentação de imagens, destacando-se pela habilidade de trabalhar efetivamente com conjuntos de dados de baixa variância e capturar detalhes minuciosos das imagens. Essas características a tornam uma ferramenta valiosa para uma ampla gama de aplicações, não apenas no campo da segmentação de imagens biomédicas, mas também em diversas outras áreas de pesquisa. Notavelmente, a U-Net desempenha um papel importante na estrutura dos Modelos de Difusão, contribuindo para o seu poderoso desempenho em tarefas de modelagem generativa. A aplicação e a relevância dos Modelos de Difusão será o tópico de discussão da próxima seção deste capítulo.

2.2 Modelos De Difusão

Emergindo como uma classe significativa de modelos generativos profundos, os Modelos de Difusão [5] têm adquirido crescente destaque no campo da Visão Computacional. Demonstram uma grande capacidade de geração sintética, desde a criação de detalhes meticulosos, até a diversidade na geração de amostras. Exemplares distintos desses modelos, como o Imagen [20] e os Modelos de Difusão Latentes [18], incrementaram a referência na modelagem generativa.

Essa classe de modelos é empregada em um amplo conjunto de tarefas de modelagem generativa, como a sintetização de imagens, super-resolução, preenchimento de imagem, edição e tradução de imagem para imagem. Eles também têm sido úteis em tarefas discriminativas, tais como segmentação de imagem, classificação e detecção de anomalias. Esta adaptabilidade sinaliza as diversas aplicações destes modelos de difusão.

A arquitetura dos Modelos de Difusão opera em duas etapas: a de difusão direta e a de difusão reversa. Na etapa direta, a entrada é perturbada gradativamente com a adição de ruído gaussiano em vários passos. A etapa reversa recorre a um modelo generativo que se esforça para recuperar a entrada original, a partir dos dados ruidosos, revertendo lentamente o processo de difusão.

Com a finalidade de aprender a distribuição de dados, os Modelos de Difusão degradam progressivamente uma variável de distribuição normal (Figura 4). Esta tarefa implica o aprendizado do processo reverso de uma cadeia fixa de Markov de comprimento T. Para a síntese de imagens, os modelos mais bem-sucedidos se baseiam numa variante ponderada do limite inferior variacional em $p(x)$.

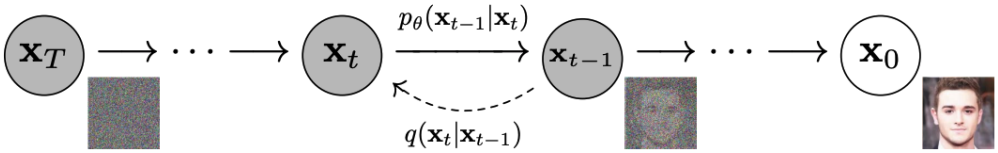


Figura 2: Processo de Difusão onde os Modelos de Difusão degradam progressivamente uma variável de distribuição normal.
Fonte: HO et al. (2020)

A introdução da arquitetura U-Net [19] nos modelos de difusão possibilitou uma aperfeiçoamento notável na qualidade das amostras geradas sobre as arquiteturas anteriores. Os Modelos de Difusão, fundamentados em probabilidade, são reconhecidos por produzir imagens de alta qualidade, fornecer cobertura de distribuição, possuir um objetivo de treinamento estacionário e escalabilidade.

Apesar de sustentarem uma referência no CIFAR-10[7], ainda ficam aquém das

[7]Disponível em: https://www.cs.toronto.edu/ kriz/cifar.html

GANs [8] em conjuntos de dados desafiadores como LSUN [23]. No entanto, eles têm demonstrado avanços consistentes com a expansão dos recursos computacionais, gerando amostras de alta qualidade mesmo em conjuntos de dados desafiadores.

Como já mencionado anteriormente, os LDMs são uma das variações dos Modelos de Difusão, que processam os dados em um espaço latente. A seguir, serão expostos como os LDMs foram desenvolvidos e seu respectivo funcionamento.

2.3 Modelos De Difusão Latente

Os Modelos de Difusão Latente [18] surgiram como uma categoria significativa no domínio dos modelos geradores, mostrando grande capacidade para uma variedade de tarefas na modelagem generativa. Em essência, esses modelos utilizam uma representação latente dos dados de entrada. Esta representação é então submetida a uma série de perturbações mediante um processo de difusão, resultando em dados que possuem um alto nível de ruído. Paralelamente, um modelo generativo atua visando restaurar a informação original, revertendo o processo de difusão em uma quantidade pré definida de passos.

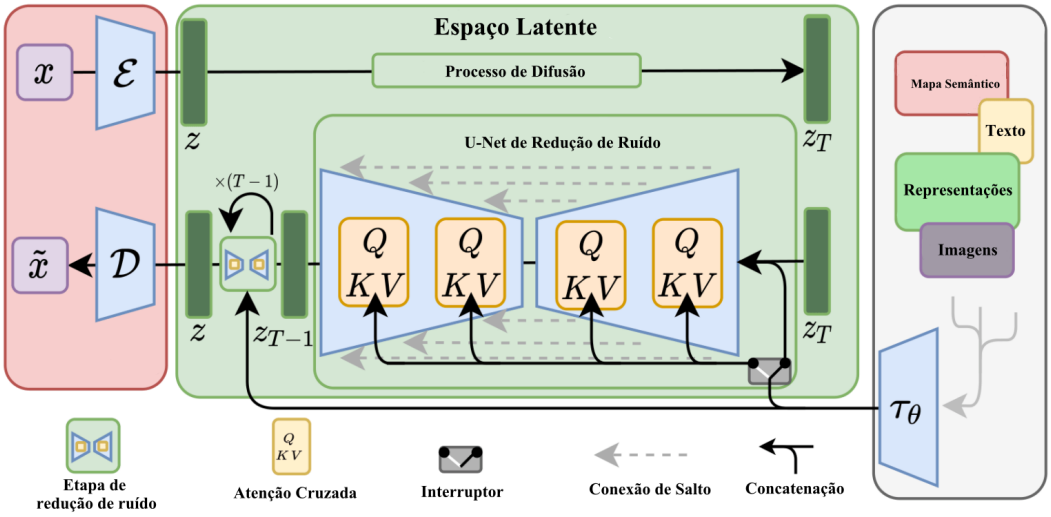


Figura 3: Arquitetura dos Modelos de Difusão Latente.
Fonte: Rombach et al. (2022)

A arquitetura dos LDMs é construída sobre dois estágios centrais: a fase de difusão direta e a fase de difusão reversa. Durante a fase de difusão direta, os dados de entrada são alterados gradualmente através da adição de ruído gaussiano em várias etapas. Neste processo, a representação latente dos dados é exposta a um mecanismo de difusão onde os detalhes da informação original são progressivamente perdidos, culminando na geração de ruído gaussiano puro. No estágio reverso, um modelo generativo é empregado para restituir os dados originais, a partir do conteúdo infestado de ruído, revertendo assim, o processo de difusão gradualmente.

A representação no espaço latente, elemento distintivo dos LDMs, em comparação a outros modelos gerativos, é aplicada tanto em tarefas generativas, quanto discriminativas. Esta representação latente é projetada para reduzir o custo computacional no processo de aprendizagem dos modelos. A avaliação e otimização dos modelos generativos no espaço de pixel enfrentam desafios, como a velocidade lenta de inferência e altos custos de treinamento, embora estratégias avançadas de amostragem e abordagens hierárquicas possam mitigar esses desafios, o treinamento com dados de imagem de alta resolução sempre implica cálculos de gradientes com alto custo.

Dessa forma, os Modelos de Difusão Latente representam uma abordagem inovadora para a modelagem generativa. Eles são reconhecidos por suas capacidades generativas e flexibilidade, sendo úteis para uma variedade de tarefas. Apesar dos desafios, como a necessidade de cálculos intensivos e a velocidade de inferência, eles têm mostrado consistência em seu desempenho, prometendo contribuir significativamente para o progresso contínuo na área de Visão Computacional e modelagem generativa. No próximo tópico, um exemplo de modelo de difusão latente, o *Stable Diffusion*, será examinado com maior profundidade.

2.4 Stable Diffusion

O *Stable Diffusion*[8], desenvolvido pela CompVis e StabilityAI em 2022, representa um marco no campo da geração de imagens por IA. Este modelo se destaca pela sua capacidade de gerar imagens detalhadas, a partir de descrições textuais. Além disso, é versátil, podendo ser aplicado a outras tarefas como preenchimento ou tradução de imagens.

O treinamento do modelo foi realizado com imagens 512×512, extraídas de um subconjunto do banco de dados LAION-5B[9]. Para condicionar o modelo a *prompts* de texto, é utilizado um codificador de texto CLIP ViT-L/14 [4] congelado. Este codificador traduz as informações de texto em uma representação numérica, que capta as ideias presentes no texto.

A estrutura do *Stable Diffusion* (Figura 4) é composta por três componentes principais. O primeiro é o codificador de texto, que traduz as informações de texto em uma representação numérica. O segundo componente, é o criador de informações de imagem, que opera no espaço latente e é responsável por sintetizar as informações de imagem. Este processo é realizado consecutivamente, adicionando mais informações relevantes a cada etapa. O último componente é o decodificador, que gera uma imagem, a partir dos dados recebidos do criador de informações. O conjunto que compõe o modelo é relativamente

[8]Disponível em: https://github.com/CompVis/stable-diffusion
[9]Disponível em: https://laion.ai/blog/laion-5b/

leve, permitindo sua execução em uma GPU com pelo menos 10GB de VRAM.

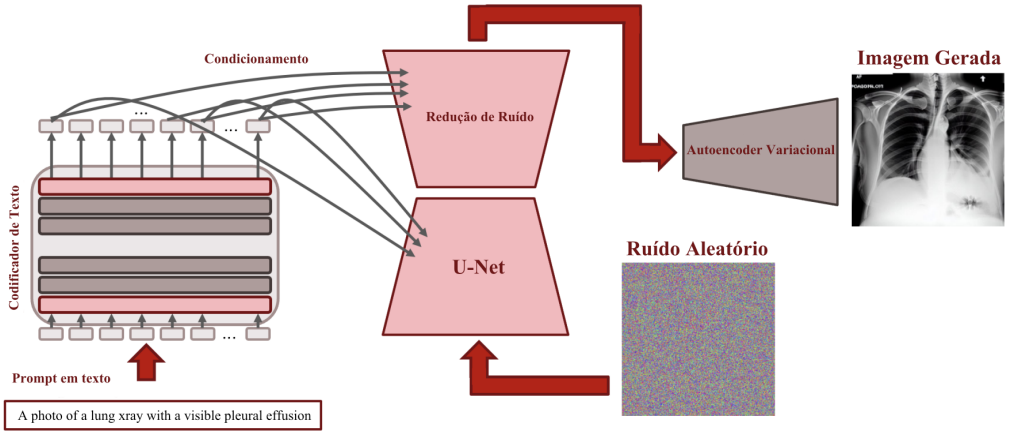


Figura 4: Arquitetura do *Stable Diffusion*.
Fonte: Chambon et al. (2022)

O modelo se destaca por sua abordagem inovadora na geração de imagens. Ao invés de processar as imagens diretamente em pixels, o processo de difusão é realizado em uma versão comprimida da imagem. Esta compressão e posterior descompressão/geração é realizada por meio de um autoencoder. Outra característica importante do *Stable Diffusion* é como o texto é incorporado ao processo de geração de imagens, para isso, o preditor de ruído é ajustado para usar o texto como entrada, adicionando uma camada de atenção entre os blocos ResNet. Ao examinar o funcionamento e a complexidade do modelo *Stable Diffusion*, fica evidente a necessidade de métricas apropriadas para avaliar adequadamente o seu desempenho. Ao examinar o funcionamento e a complexidade do modelo *Stable Diffusion*, fica evidente a necessidade de métricas apropriadas para avaliar adequadamente o seu desempenho. No próximo tópico, serão expostas as métricas FID (*Frechet Inception Distance*), Precisão e *Recall*.

2.5 Métricas De Avaliação

Ao avaliar um sistema de Aprendizado de Máquina como o *Stable Diffusion*, é crucial estabelecer uma metodologia de avaliação com métricas coerentes, as mesmas devem proporcionar uma visão qualificada do comportamento e eficiência do modelo em análise. Nesse contexto, são popularmente utilizadas a FID, Precisão e *Recall*. Cada uma dessas métricas oferece uma perspectiva única do sistema, resultando em uma análise global e robusta.

2.5.1 FID (*Frechet Inception Distance*)

A *Frechet Inception Distance* (FID) [10], proposta por Martin Heusel et al. (2017) é uma métrica que compara a distância estatística entre as distribuições de características

das imagens geradas e reais. A FID é calculada usando a distância *Fréchet*, uma medida de similaridade entre duas distribuições multivariadas. Esta distância é calculada no espaço das características extraídas pela rede *Inception*, e considera tanto a média, quanto a covariância dessas características.

Uma FID baixa indica que as duas distribuições são semelhantes, o que implica que as imagens geradas são semelhantes às imagens reais, em termos de características de alto nível. A FID é especialmente útil para avaliar a qualidade e a diversidade das imagens geradas por modelos generativos, ao considerar tanto a qualidade das imagens individuais, quanto a variedade de imagens diferentes produzidas pelo modelo.

2.5.2 Precisão E *Recall*

No trabalho "*Improved Precision and Recall Metric for Assessing Generative Models*"[13] foi apresentada uma nova métrica de avaliação da qualidade e a cobertura de amostras produzidas por modelos generativos em tarefas de geração de imagens. Os autores argumentam que as métricas existentes agrupam qualidade e variação sem uma compensação clara, dificultando o diagnóstico do desempenho do modelo.

Os autores propõem uma nova métrica que expressa a qualidade das amostras geradas usando dois componentes separados: precisão e *recall*. A precisão é definida como a probabilidade de que uma imagem aleatória da distribuição gerada esteja dentro do suporte da distribuição da imagem real. Por outro lado, o *recall* corresponde à probabilidade de que uma imagem aleatória da distribuição da imagem real esteja dentro do suporte da distribuição da imagem gerada. Os conceitos de precisão e *recall* são emprestados do campo de recuperação de informações, tendo sido aplicados ao estudo de modelos generativos. Do ponto de vista clássico, a precisão denota a fração de imagens geradas que são realistas, e o *recall* mede a fração do coletor de dados de treinamento coberto pelo gerador. Ambos são calculados como expectativas de associação de conjuntos binários em uma distribuição, ou seja, medindo a probabilidade de uma imagem extraída de uma distribuição ser classificada como pertencente ao suporte da outra distribuição.

Essas métricas de avaliação, quando utilizadas em conjunto, fornecem uma visão abrangente do desempenho de um modelo generativo. Elas permitem avaliar tanto a qualidade das amostras geradas (precisão), quanto a diversidade dessas amostras (*recall*), além de comparar a similaridade entre as imagens geradas e as reais (FID). Além disso, essas métricas podem ser usadas para otimizar os modelos durante o treinamento, ajustando os parâmetros do modelo para maximizar a precisão e o *recall* e minimizar a FID.

3 TRABALHOS RELACIONADOS

A geração de imagens sintéticas por meio de modelos generativos é um campo em rápida expansão, que vivencia um progresso notável nos últimos anos. Este progresso é evidenciado por uma série de avanços significativos e desenvolvimentos de novas técnicas em Aprendizado Profundo, Processamento de Linguagem Natural (PLN) e Visão Computacional. Novas investigações e modelos são continuamente introduzidos por empresas privadas e projetos de código aberto. Algumas das principais contribuições incluem modelos como o DALL-E da OpenAI [17], o MidJourney [10], da companhia homônima e o *Stable Diffusion*[11], fruto de uma colaboração entre a Stability.AI e a Runway. Além desses, outros modelos, como VQGAN + CLIP [4] , GLIDE [14] e Big Sleep[12], também têm desempenhado um papel importante na evolução do campo.

Este capítulo visa contextualizar e discutir o estado da arte da geração de imagens sintéticas, a partir de descrições em linguagem natural, fornecendo uma visão geral das principais técnicas, abordagens e desafios enfrentados na área. Os trabalhos foram selecionados a partir da plataforma de pesquisa acadêmica Google Scholar[13], que dispõe de um extenso acervo de publicações científicas, e do mecanismo de busca de artigos populares em Aprendizado de Máquina da organização Labml.ai[14]. As palavras-chave empregadas na pesquisa abrangeram "image generation", "generative models", "diffusion models"e "image to text". Ao longo deste capítulo, serão examinados os principais trabalhos e suas contribuições, bem como as principais limitações e desafios enfrentados no campo da geração de imagens sintéticas a partir de descrições textuais.

Rombach et. al (2022) produziram o trabalho "*High-Resolution Image Synthesis with Latent Diffusion Models*". No artigo, os autores propõem uma nova abordagem para aprimorar a eficiência de treinamento e inferência de DMs, conhecidos por apresentar resultados de síntese de última geração em dados de imagem e além. Através da aplicação de DMs no espaço latente de *autoencoders* pré-treinados, os autores desenvolvem LDMs, que alcançam resultados significantes em tarefas como preenchimento de imagem e síntese de imagem condicionada à classe, além de apresentar desempenho competitivo em várias outras tarefas. Os LDMs reduzem significativamente os requisitos computacionais em comparação com os DMs baseados em pixels, possibilitando treinamento e inferência mais eficientes.

A abordagem proposta consiste em dividir o treinamento em duas etapas: primeiro, um autoencoder é treinado para fornecer um espaço representacional de menor dimensão

[10]Disponível em: https://www.midjourney.com/home/
[11]Disponível em: https://github.com/CompVis/stable-diffusion
[12]Disponível em: https://github.com/lucidrains/big-sleep
[13]Disponível em: https://scholar.google.com/
[14]Disponível em: https://papers.labml.ai/papers/

e eficiente, que seja perceptualmente equivalente ao espaço de dados. Em seguida, os DMs são treinados no espaço latente aprendido, permitindo uma redução ótima entre a redução de complexidade e a preservação de detalhes, melhorando a fidelidade visual. A arquitetura dos LDMs inclui camadas de atenção cruzada, tornando-os geradores flexíveis para entradas condicionais gerais, como texto ou caixas delimitadoras, e permitindo a síntese de alta resolução convolucionalmente.

Além disso, os autores destacam várias vantagens dessa abordagem. O treinamento do estágio de autoencoder universal precisa ser feito apenas uma vez, permitindo seu reuso em múltiplos treinamentos de DMs, ou na exploração de diferentes tarefas. A abordagem proposta escala para dados de maior dimensão, proporcionando reconstruções mais fiéis e detalhadas, em comparação com trabalhos anteriores, e possibilitando a síntese eficiente de imagens de alta resolução. Os LDMs também apresentam um mecanismo de condicionamento de propósito geral baseado em atenção cruzada, permitindo treinamento multimodal e aplicação em tarefas como síntese de imagem condicionada a texto e layout. A pesquisa de Rombach et. al (2022) foi imprescindível para este trabalho, à medida que através dele foi possível a construção do modelo *Stable Diffusion*, e a disseminação do conteúdo sobre os modelos de difusão latente.

No trabalho intitulado "*Adapting Pretrained Vision-Language Foundational Models to Medical Imaging Domains*", Pierre et al. (2022) argumentaram que os modelos generativos multimodais, embora treinados com milhões de pares de imagens e suas respectivas descrições em texto, frequentemente falham em generalizar para domínios específicos, como as imagens médicas. Outro ponto destacado é a escassez de conjuntos de dados de condições clínicas, como derrame pleural no pulmão. Para superar essas limitações, os autores propuseram realizar o fine tuning do modelo generativo *Stable Diffusion* com imagens médicas de dois grandes conjuntos de dados de raios-X de tórax (CheXpert e MIMIC-CXR).

O *fine tuning* do modelo permitiu a geração de imagens de radiografia de alta fidelidade, sendo capaz de inserir uma anormalidade de aparência realista em uma imagem radiológica sintética. Os resultados foram validados por meio de métricas quantitativas de qualidade de imagem, como a *Fréchet Inception Distance* (FID), e avaliações qualitativas conduzidas por um radiologista torácico. No geral, o novo modelo gerado conseguiu melhorar os resultados do *Stable Diffusion* pré-treinado.

Em suma, o artigo demonstra que o modelo generativo pré-treinado pode ser adaptado para representar conceitos médicos e gerar imagens médicas sintéticas de alta qualidade a partir de descrições em texto. Essa abordagem tem potencial para mitigar a escassez de conjuntos de dados médicos rotulados e melhorar o treinamento e a análise de algoritmos de Aprendizado de Máquina no campo da imagem médica. O trabalho em questão serviu como referência para validar a viabilidade de análises qualitativas do *Stable*

Diffusion e mostrar que ainda há espaços para melhorias, apesar dos excelentes resultados atuais do modelo.

Ali Borji (2022) desenvolveu o trabalho "*Generated Faces in the Wild: Quantitative Comparison of Stable Diffusion*, Midjourney and DALL-E 2"[1], onde realiza uma comparação quantitativa entre três sistemas populares de geração de imagens: *Stable Diffusion*, Midjourney e DALL-E 2, em sua capacidade de gerar rostos fotorealistas em ambientes naturais. Segundo o estudo, o *Stable Diffusion* gera rostos melhores do que os outros sistemas, conforme a pontuação FID. O autor também apresenta um conjunto de dados chamado *Generated Faces in the Wild*, que inclui um total de 15.076 rostos gerados. O estudo espera incentivar pesquisas futuras na avaliação de modelos geradores e na melhoria desses modelos. Os rostos gerados e reais foram obtidos utilizando-se o conjunto de dados COCO e o conjunto *Labeled Faces in the Wild*. Em seguida, foi empregada a métrica FID para avaliar a qualidade dos rostos gerados em comparação a um conjunto de rostos reais.

Os resultados mostraram que o *Stable Diffusion* apresentou uma pontuação FID menor, gerando rostos melhores do que os outros dois modelos. No entanto, a qualidade dos rostos gerados ainda é muito inferior à dos rostos reais. A pesquisa sugere que trabalhos futuros explorem conjuntos maiores de rostos gerados e investiguem detalhes faciais mais refinados, como expressões, idade e especificações de pontos de vista. Além disso, os modelos podem ser comparados a outras categorias interessantes, como humanos, gatos, cães, carros e quartos.

O trabalho de Ali Borji (2022) foi utilizado como referência para realizar análises qualitativas, e comparativas entre diferentes modelos de geração de imagens, destacando as vantagens e limitações de cada abordagem. A análise meticulosa do desempenho do *Stable Diffusion*, em comparação com os outros sistemas, contribuiu para o entendimento de suas características e a identificação de áreas que necessitam de melhorias.

O Quadro 1 abaixo fornece um resumo abrangente e comparativo dos trabalhos relacionados à pesquisa. Cada linha representa um trabalho distinto, identificado pelo seu título, os autores e o ano de publicação (primeira coluna). A segunda coluna detalha o método principal utilizado em cada um desses estudos, abrangendo desde a geração de imagens com Modelos de Difusão operando no espaço latente, até a aplicação de *fine tuning* no modelo *Stable Diffusion* para criação de imagens médicas específicas.

A terceira coluna fornece informações sobre os conjuntos de dados empregados nos respectivos trabalhos. Isso inclui, por exemplo, o uso do LAION-5B, um conjunto de dados diversificado, e também de conjuntos mais específicos, como CheXpert e MIMIC-CXR, focados em imagens médicas.

Por fim, a coluna "Avaliação dos Resultados"relata as métricas e abordagens utili-

zadas para avaliar a eficácia dos métodos aplicados em cada estudo. Aqui, métricas como FID, Precisão e *Recall* são citadas, assim como a análise qualitativa humana empregada neste trabalho.

Tabela 1: Descrição dos artigos estudados

Trabalho	**Método**	**Dados**	**Avaliação dos Resultados**
"High-Resolution Image Synthesis with Latent Diffusion Models", Rombach et. al, 2019	Geração de imagens com DMs operando no espaço latente	LAION-400M	FID Precisão e *Recall*
"Adapting Pretrained Vision-Language Foundational Models to Medical Imaging Domains", Pierre et. al, 2022	Fine tune do Stable Diffusion para gerar imagens de condições médicas específicas	LAION-5B, CheXpert e MIMIC-CXR	FID
"Generated Faces in the Wild: Quantitative Comparison of Stable Diffusion, Midjourney and DALL-E 2", Ali Borji, 2022	Realiza uma comparação quantitativa entre três sistemas populares de geração de imagens Stable Diffusion, Midjourney e DALL-E 2 em sua capacidade de gerar rostos fotorealistas	LAION-5B	FID
Este trabalho	Análise da geração de imagens sintéticas a partir de descrições textuais, produzidas pelo modelo Stable Diffusion	LAION-5B	Análise qualitativa humana

Fonte: Autoria própria

Este trabalho investiga a geração de imagens sintéticas a partir de descrições textuais, utilizando o modelo generativo *Stable Diffusion*. Embora compartilhe a temática principal com os estudos correlatos mencionados, ele se diferencia por sua ênfase em analisar qualitativamente.

O estudo "High-Resolution Image Synthesis with Latent Diffusion Models"de Rombach et al. (2022) centra-se em aprimorar a eficiência de treinamento e inferência dos MDs, introduzindo os MDLs, fornecendo uma abordagem eficiente para a geração de imagens, particularmente na redução dos requisitos computacionais associados aos MDs baseados

em pixels. Esta pesquisa difere do trabalho de Rombach et al. (2022) ao focar na aplicação direta do *Stable Diffusion* para a síntese de imagens a partir de descrições textuais, ao invés de explorar alterações estruturais no modelo para melhorar a eficiência.

Em "Adapting Pretrained Vision-Language Foundational Models to Medical Imaging Domains", Pierre et al. (2022) exploram a aplicação dos modelos generativos multimodais ao domínio médico, especificamente para a geração de imagens radiográficas sintéticas a partir de descrições textuais. Assim como a presente pesquisa, o trabalho de Pierre et al. (2022) emprega o modelo *Stable Diffusion*. Contudo, enquanto Pierre et al. (2022) adaptam o modelo para um domínio altamente especializado (imagens médicas), esta pesquisa se propõe a investigar a capacidade do *Stable Diffusion* em um contexto mais amplo e não especializado, abrangendo uma variedade de descrições textuais.

O estudo de Ali Borji (2022), "Generated Faces in the Wild: Quantitative Comparison of *Stable Diffusion*, Midjourney and DALL-E 2", conduz uma análise comparativa de três sistemas populares de geração de imagens. Ao contrário da presente pesquisa, o trabalho de Ali Borji (2022) foca principalmente em uma avaliação quantitativa, comparando a capacidade desses sistemas em gerar rostos fotorealistas. Este trabalho, por outro lado, visa investigar a geração de imagens de uma variedade mais ampla de descrições textuais, não se limitando a rostos ou a qualquer outra categoria específica de imagens.

Em termos de contribuição, esta pesquisa pode enriquecer o campo da geração de imagens sintéticas, a partir de descrições textuais, ao examinar a versatilidade e aplicabilidade do modelo *Stable Diffusion* em um contexto diversificado. Além disso, ao focar na análise qualitativa, este trabalho tem potencial para complementar as análises quantitativas dos trabalhos mencionados, oferecendo uma visão mais ampla dos modelos generativos na prática. Este trabalho também pode contribuir para o entendimento dos desafios e limitações do *Stable Diffusion* em um cenário mais generalista, apontando direções para futuras pesquisas e melhorias.

4 METODOLOGIA

Neste capítulo, será abordada a metodologia empregada para atingir os objetivos específicos estabelecidos no Capítulo 1. Serão descritos detalhadamente os métodos aplicados, as decisões tomadas ao longo do desenvolvimento e a organização do ambiente experimental. Também será exposto sobre a seleção do modelo generativo adotado e as descrições das cenas de teste empregadas para avaliar o desempenho do referido modelo. Finalmente, serão detalhados os procedimentos de execução do modelo e a análise dos resultados obtidos, visando fornecer um entendimento claro e abrangente de todo o processo metodológico adotado.

4.1 Configuração Do Ambiente E Softwares Utilizados

Para desenvolver este trabalho, foi utilizada a linguagem de programação Python (versão 3.9.12) em conjunto com diversos pacotes, como *PyTorch*, que foi empregado para carregar o modelo e funções auxiliares necessárias. A biblioteca *Venv* permitiu criar um ambiente virtual, garantindo a independência do ambiente de desenvolvimento. Outras bibliotecas importantes também foram empregadas, como *cv2*, que oferece diversas funções para processamento de imagens; *Numpy*, fundamental para computação científica em Python; *Pytorch-lightning*, que facilita o treinamento de modelos em *PyTorch*, de maneira mais simples e escalável; e *Transformers*, uma biblioteca para PLN baseada no *PyTorch*. As dependências necessárias para executar o código estão listadas no arquivo "*requirements*", que inclui versões específicas de pacotes como *Numpy*, *PyTorch*, *cv2*, *Pytorch-lightning*, *Transformers*, entre outros. O modelo generativo selecionado para a realização dos testes propostos foi o *Stable Diffusion*. Na próxima seção será descrita a base de dados utilizada em seu treinamento.

4.2 Base De Dados

O LAION5B é um conjunto de dados com 5,85 bilhões de pares de imagens e textos filtrados pelo CLIP. Ele é dividido em três partes: pares com textos em inglês, pares com textos em outras línguas e pares cuja língua não pôde ser identificada. Além disso, o banco de dados inclui informações como URL, texto, largura e altura da imagem, idioma e probabilidade de ser uma imagem com marca d'água ou inadequada.

Esse conjunto de dados foi criado para possibilitar o treinamento de modelos de linguagem e visão em larga escala, como *Stable Diffusion*, DALL-E [17], GLIDE [14] e CLIP [4], já que, até então, não haviam conjuntos de dados em escala bilionária disponíveis publicamente. Com o LAION5B, pesquisadores de todo o mundo podem treinar modelos

de última geração e ampliar as possibilidades de pesquisa em modelos de linguagem e visão em várias línguas.

4.3 Descrições Dos Testes

Nesta seção, serão detalhadas as descrições textuais empregadas nos experimentos para alcançar os objetivos propostos. As descrições foram elaboradas em inglês, dado que o modelo *Stable Diffusion* exibe melhores resultados nesse idioma, devido à maioria dos dados de treinamento estarem em inglês.

Inicialmente, as descrições textuais utilizadas para a geração das imagens sintéticas continham palavras simples, como "*dog*"e "*cat*". A complexidade dessas descrições foi incrementada progressivamente, com frases como "*one dog*"e "*one cat*", seguidas por "*two dogs*"e "*two cats*", repetindo-se as interações até "*twenty dogs*"e "*twenty cats*". Subsequentemente, cores foram introduzidas nas descrições, originando cenários como "*red dog*"e "*blue cat*". A próxima etapa consistiu em incrementar a quantidade de gatos azuis e cachorros vermelhos nas cenas.

Finalmente, as descrições textuais foram combinadas, empregando frases como "*dog and cat*", "*one dog and one cat*", além da inclusão das cores. Assim, exploramos a variação e a complexidade das descrições textuais das cenas de teste.

É importante ressaltar que para cada *prompt* proposto para teste, foi criada uma pequena amostra, resultando na geração de cinco imagens. Este procedimento permite uma análise qualitativa robusta dos resultados obtidos pelo modelo generativo *Stable Diffusion*, contribuindo para uma compreensão mais aprofundada do impacto das modificações textuais nas descrições das cenas nas imagens resultantes.

O Quadro 2 a seguir fornece uma descrição geral dos experimentos realizados. Cada linha representa um experimento distinto, e a segunda coluna informa os *prompts* utilizados na execução dos testes.

Tabela 2: Descrição dos testes realizados

Experimentos	Prompts correspondentes
A: Impacto da alteração da ordem na descrição da cena	*Dog and cat*, [*one-twenty*] *red dogs and* [*one-twenty*] *blue cats.* *Cat and dog*, [*one-twenty*] *blue cats and* [*one-twenty*] *red dogs.*
B: Influência do aumento da complexidade da descrição da cena	*Dog*, [*one-twenty*] *dogs.* *Cat*, [*one-twenty*] *cats.* *Red dog*, [*one-twenty*] *red dogs.* *Blue cat*, [*one-twenty*] *blue cats.* *Red dog and blue cat*, [*one-twenty*] *red dogs and* [*one-twenty*] *blue cats.*
C: Consequências do acréscimo da quantidade de itens descritos na cena	*Dog*, [*one-twenty*] *dogs.* *Cat*, [*one-twenty*] *cats.* *Dog and cat*, [*one-twenty*] *red dogs and* [*one-twenty*] *blue cats.*

Fonte: Autoria própria

4.4 Pipeline De Execução

O modelo *Stable Diffusion* de difusão latente foi selecionado para a realização dos experimentos por ser considerado o estado da arte entre os modelos de código aberto para geração de imagens sintéticas, a partir de uma descrição em texto. Para executar o modelo *Stable Diffusion*, foi necessário instalar os pacotes do repositório do *Stable Diffusion*[1], viabilizado pela colaboração entre a Stability AI e Runway, e fazer o download dos pesos do modelo já treinado. Os experimentos foram realizados em um MacBook Air com o chip M1, com 16 GB de memória RAM, CPU de 8 núcleos e GPU de 7 núcleos.

Foi desenvolvida uma rotina em Python que permitiu a execução dos testes, nela são especificados como parâmetros: a descrição em texto que deseja-se gerar a imagem sintética, o número de instâncias que devem ser geradas e o número de iterações realizadas pelo modelo.

A partir do texto fornecido como entrada, o código carrega e executa o *Stable Diffusion* pré-treinado. Ele efetua a amostragem das imagens e analisa a presença de

conteúdo inadequado. Caso identifique algum elemento inapropriado, a imagem é trocada por uma padrão. Adicionalmente, o código insere uma marca d'água invisível nas imagens produzidas, assegurando a autenticidade da fonte.

4.5 Métricas De Avaliação

Nesta seção do trabalho, o objetivo é apresentar a metodologia empregada para avaliar a qualidade das imagens sintéticas geradas em relação às descrições textuais fornecidas. A avaliação das imagens foi conduzida por meio de uma análise qualitativa, que visou determinar se os resultados das imagens geradas estavam conforme as descrições textuais empregadas nos testes. Para alcançar uma avaliação abrangente e criteriosa, o pesquisador responsável examinou cuidadosamente as imagens geradas, comparando-as com as descrições textuais correspondentes.

Esta metodologia baseada em critérios subjetivos e humanos foi escolhida devido à complexidade e à natureza intrínseca das imagens sintéticas, que muitas vezes não podem ser avaliadas adequadamente por métricas quantitativas. A análise qualitativa proporcionou uma apreciação mais aprofundada e contextualizada dos resultados, permitindo identificar aspectos sutis e nuances que poderiam ser ignorados por uma avaliação puramente numérica.

Além disso, a análise qualitativa possibilitou a identificação de tendências e padrões nos resultados, contribuindo para uma melhor compreensão do funcionamento do modelo generativo *Stable Diffusion* e seu desempenho na geração de imagens a partir de descrições textuais. Dessa forma, a avaliação qualitativa realizada pelo pesquisador proporcionou uma base sólida para a interpretação dos resultados e a elaboração de conclusões relevantes para o campo de estudo em questão.

5 RESULTADOS E DISCUSSÕES

Este capítulo tem por finalidade apresentar os resultados científicos decorrentes da aplicação dos experimentos definidos na metodologia, conduzidos utilizando o modelo generativo *Stable Diffusion*. A análise dos resultados foi pautada pelos três objetivos específicos estabelecidos anteriormente. Para cada caso de teste, uma amostra de cinco imagens foi gerada a partir de cada *prompt* selecionado, visando avaliar a robustez e a consistência do modelo.

5.1 Impacto Da Alteração Da Ordem Na Descrição Da Cena

Com base na análise do conjunto de amostras coletadas, foi validada a hipótese de que a ordem dos itens descritos nos *prompts*, influencia de maneira expressiva o resultado das imagens geradas pelo modelo *Stable Diffusion*. Essa constatação foi observada consistentemente em todas as amostras examinadas - um total de cinco para cada *prompt*.

Por exemplo, ao executar o modelo informando o prompt "*five dogs and five cats*"(Figura 5 (a)), foi observado que a imagem gerada continha uma predominância numérica de cães em comparação aos gatos. No entanto, ao inverter a sequência descritiva para "*five cats and five dogs*"(Figura 5 (b)), a representação gerada pelo modelo alterou a priorização, destacando de forma mais acentuada a presença dos gatos, seguidos pelos cães.

(a) *Prompt*: "*five dogs and five cats*" **(b) *Prompt*: "*five cats and five dogs*"**

Figura 5: Imagens resultantes dos prompts relacionados ao teste de alteração da ordem na descrição da cena.
Fonte: Autoria Própria

Este fenômeno evidencia que o modelo *Stable Diffusion* processa os *prompts* em uma sequência definida, refletindo tal ordenação na construção das imagens geradas. Por-

tanto, a ordem na qual os elementos são descritos nos *prompts* tem um impacto significativo na distribuição quantitativa dos elementos representados visualmente. Essa descoberta enfatiza a importância do arranjo sequencial das informações fornecidas ao modelo, especialmente em contextos onde a precisão na representação de diferentes elementos é de importância crítica.

5.2 Influência Do Aumento Da Complexidade Da Descrição Da Cena

A evolução da complexidade nos *prompts* resultou em modificações correspondentes nas imagens geradas pelo modelo *Stable Diffusion*. Uma descrição elementar, como "*one dog*"(Figura 6 (a)), conduziu à geração de uma representação única de um cão. Contudo, ao adicionar mais características à descrição, exemplificado pelo *prompt* "*one red dog*"(Figura 6 (b)), o modelo conseguiu integrar esse atributo adicional na imagem produzida.

(a) *Prompt*: "*one dog*" **(b) *Prompt*: "*one red dog*"**

Figura 6: Imagens resultantes dos *prompts* relacionados ao teste do aumento da complexidade da descrição da cena.
Fonte: Autoria Própria

Foi explorada a resposta do modelo *Stable Diffusion* ao aumento progressivo da complexidade na descrição da cena, introduzindo características de cor e aumentando a quantidade de itens descritos. Para *prompts* com a adição de uma característica cromática, como "*blue cat*", "*one blue cat*", "*two blue cats*", "*three blue cats*", "*red dog*", "*one red dog*", "*two red dogs*", "*three red dogs*", as imagens geradas, em cada conjunto amostral composto por cinco imagens para cada teste, exibiram uma representação fidedigna tanto na cor, como na quantidade de itens descritos.

No entanto, a complexidade adicional em descrições como "*four red dogs*" e "*four blue cats*" desafiou a precisão do modelo. Nesses casos, apenas uma das cinco imagens

geradas para cada *prompt* conseguiu representar corretamente a quantidade e a cor dos elementos descritos. À medida que a complexidade continuou a aumentar, exemplificada pelos *prompts* "*six red dogs*" e "*six blue cats*", embora a coloração tenha sido representada corretamente em todas as amostras, a precisão na quantidade de itens representados diminuiu.

Nos *prompts* em que foram descritos itens com diferentes cores simultaneamente, como "*one red dog and one blue cat*" e "*two red dogs and two blue cats*", observou-se um padrão similar. Quatro das cinco amostras do *prompt* "*one red dog and one blue cat*" representaram corretamente a quantidade e a cor dos itens. Porém, a partir do *prompt* "*two red dogs and two blue cats*", a quantidade de itens representados corretamente começou a diminuir, ainda que a coloração fosse retratada de forma acurada.

Ao elevar ainda mais a complexidade da descrição, como exemplificado pelo *prompt* "*ten red dogs and ten blue cats*"(Figura 7), o modelo *Stable Diffusion* enfrentou obstáculos na representação precisa da quantidade de elementos descritos, assim como na manutenção de uma representação realista dos objetos na cena. Tal dificuldade pode ser atribuída à limitação do modelo na interpretação de descrições contendo um elevado número de elementos, que pode estar associada à capacidade de processamento do modelo e à natureza das representações dos objetos na base de dados utilizada para o treinamento do modelo.

(a) Instância 1 **(b) Instância 2**

Figura 7: Imagens resultantes do *prompt* "*ten red dogs and ten blue cats*"
Fonte: Autoria Própria

5.3 Consequências Do Acréscimo Da Quantidade De Itens Descritos Na Cena

Os achados apontam para uma implicação direta entre o acréscimo na quantidade de elementos descritos na cena e a degradação da precisão na representação desses elementos e na qualidade de realismo nas imagens geradas pelo modelo *Stable Diffusion*. Para

ilustrar, em descrições de cena que continham até quatro elementos, tal como observado nos *prompts* "*one cat*", "*two cats*", "*three cats*", "*four cats*", "*one dog*"(Figura 8 (a)), "*two dogs*", "*three dogs*", "*four dogs*", o modelo apresentou um desempenho notável. Em todas as cinco amostras geradas para cada um destes *prompts*, as representações geradas eram tanto precisas na quantidade de elementos descritos, quanto realistas em sua apresentação.

Contudo, ao introduzir uma quinta entidade nos *prompts*, como em "*five cats*"e "*five dogs*", foi constatada uma diminuição substancial na precisão das representações geradas pelo modelo. De fato, em tais circunstâncias, apenas uma das cinco amostras geradas para cada *prompt* foi capaz de reproduzir corretamente o número de entidades descritas, mesmo que a qualidade do realismo ainda permanecesse aceitável.

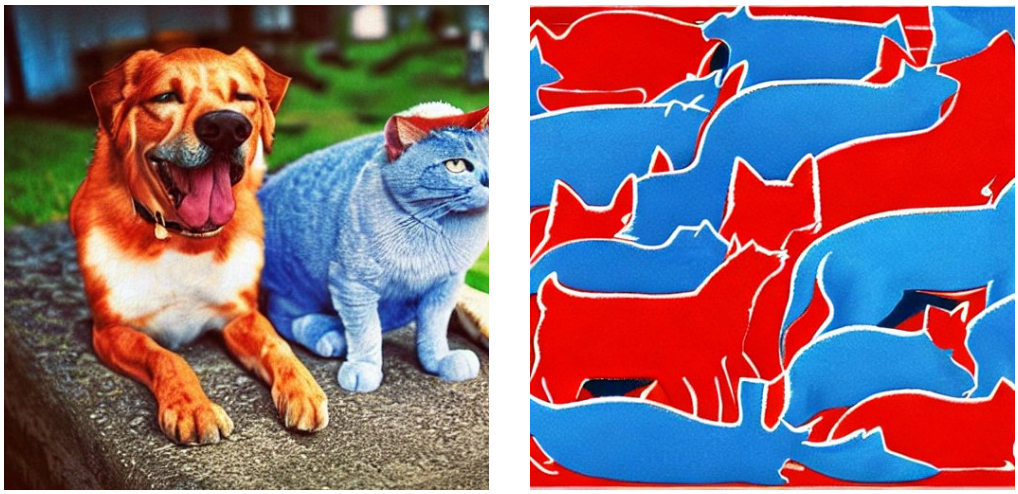
O comportamento persistiu quando a descrição textual incluiu seis ou mais entidades, exemplificado nos *prompts* "*six cats*", "*ten cats*", "*six dogs*", e "*twenty dogs*"(Figura 8 (b)). Em todos esses casos, as amostras geradas pelo modelo falharam em representar a quantidade correta de elementos descritos, demonstrando uma diminuição ainda mais acentuada na precisão da representação. Além disso, com a introdução de dez elementos nas descrições, a qualidade do realismo nas imagens geradas também diminuiu.

(a) *Prompt*: "*one dog*" **(b) *Prompt*: "*twenty dogs*"**

Figura 8: Imagens resultantes dos prompts relacionados ao teste do acréscimo da quantidade de itens descritos na cena.
Fonte: Autoria Própria

Ao investigar *prompts* que englobam descrições combinadas de cães e gatos (Figura 9), uma tendência semelhante foi observada. Para amostras geradas para *prompts* até "*three dogs and three cats*", todas as representações obtidas conseguiram retratar corretamente a quantidade de itens descritos. Contudo, ao aumentar a complexidade para "*four dogs and four cats*", apenas três entre as cinco amostras geradas puderam retratar corretamente a quantidade de entidades descritas. Quando o *prompt* foi "*five dogs and*

five cats", apenas uma única amostra conseguiu representar corretamente a quantidade de itens descritos. Em *prompts* subsequentes, todas as amostras falharam em retratar corretamente a quantidade de itens.



(a) ***Prompt*: "*one red dog and one blue cat*"**

(b) ***Prompt*: "*ten red dogs and ten blue cats*"**

Figura 9: Imagens resultantes dos prompts relacionados ao teste do acréscimo da quantidade de itens descritos na cena, incluindo a variância de cor.
Fonte: Autoria Própria

Este conjunto de evidências retrata a capacidade do modelo *Stable Diffusion* em lidar com descrições textuais complexas. Embora o modelo tenha demonstrado competência na decodificação de descrições textuais e na reprodução na imagem da ideia de pluralidade, contida em descrições com um número maior de elementos, a precisão na representação exata do número de elementos muitas vezes não foi atingida. Este comportamento tornou-se mais evidente à medida que a quantidade de itens descritos na cena se ampliava, colocando um desafio adicional para o modelo no que concerne à organização espacial dos elementos na imagem, o que, em certos casos, resultou em elementos sobrepostos ou agrupados.

Dessa forma, os achados sinalizam que o modelo *Stable Diffusion* decodifica as descrições de cenas expressas em linguagem natural de maneira sequencial e contextualizada, o que repercute diretamente na constituição visual dos elementos na imagem gerada. A complexidade ampliada na descrição textual, seja no que tange ao incremento de detalhes ou ao acréscimo na quantidade de elementos descritos, interfere de maneira significativa nos resultados obtidos nas imagens geradas. É preciso, no entanto, ressaltar que este modelo apresenta algumas restrições na precisão de suas representações, em especial quando confrontado com descrições de alta complexidade. A evolução futura desta classe de modelos de Inteligência Artificial passa necessariamente pela mitigação destas limitações, visando o aperfeiçoamento de sua capacidade interpretativa e representativa, independentemente do grau de complexidade das descrições a serem processadas.

6 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

O presente trabalho se propôs a analisar a geração de imagens sintéticas, a partir de descrições textuais, utilizando o modelo generativo *Stable Diffusion*, buscando examinar qualitativamente a influência de alterações nas descrições de cena nas imagens resultantes. O foco foi no impacto da alteração na ordem das descrições, no aumento da complexidade das descrições e no acréscimo na quantidade de elementos descritos na cena.

Neste contexto, a análise realizada ao longo deste trabalho permitiu observar que o modelo *Stable Diffusion* interpreta a linguagem natural de maneira sequencial e contextualizada, dando origem a imagens que, em grande medida, correspondem às descrições fornecidas. Observou-se que a alteração na ordem das descrições das cenas pode resultar em diferenças significativas nas imagens geradas, validando a hipótese de que a interpretação do modelo é sequencial.

Quanto à complexidade das descrições, constatou-se que o modelo é capaz de incorporar elementos de maior sofisticação à cena, incluindo a distinção de cores para os objetos descritos. No entanto, a precisão e o detalhamento das imagens geradas são desafiados quando as descrições ascendem a patamares mais elevados de minuciosidade, particularmente ao representar múltiplos objetos (como cães e gatos) em variações quantitativas e com diversas colorações simultaneamente.

O aumento no número de elementos descritos na cena também demonstrou influenciar as imagens resultantes. O modelo mostrou-se capaz de produzir imagens com múltiplos elementos, ainda que a representação exata do número de itens descritos não fosse sempre atingida, especialmente em *prompts* com mais de cinco repetições por objeto a ser representado. Esta constatação pode apontar para limitações inerentes ao modelo em lidar com descrições de alto número de itens na cena.

A despeito das limitações identificadas, os resultados alcançados neste estudo confirmaram o potencial do *Stable Diffusion*, enquanto ferramenta para geração de imagens sintéticas, a partir de descrições em linguagem natural. A capacidade do modelo de interpretar e traduzir descrições textuais de maneira geralmente fiel, mesmo incrementando a complexidade da descrição e no número de elementos, é uma indicação promissora de sua aplicabilidade em uma grande diversidade de situações e contextos práticos.

Este trabalho, portanto, contribui para a compreensão das nuances de funcionamento do modelo *Stable Diffusion* e das influências de variações na descrição textual sobre a geração de imagens sintéticas. Espera-se que os achados possam fomentar futuras pesquisas na área e inspirar o desenvolvimento de modelos generativos ainda mais precisos e versáteis.

Como perspectivas futuras, sugere-se a realização de estudos adicionais que con-

templem a utilização de descrições textuais de maior complexidade e diversidade, e que explorem outros modelos generativos de imagens, com o intuito de proporcionar uma comparação mais aprofundada e abrangente das capacidades e limitações desses sistemas.

Outro aspecto a ser considerado em investigações futuras é a implementação e avaliação de um processo de ajuste fino do modelo *Stable Diffusion*. A reconfiguração de um modelo pré-treinado como o *Stable Diffusion* para ajustá-lo a tarefas ou contextos específicos pode potencialmente melhorar a precisão e relevância das imagens geradas, a partir das descrições textuais, expandindo ainda mais as capacidades do modelo.

Posteriormente a essa etapa de ajuste fino, a replicação dos testes apresentados neste estudo permitiria uma avaliação comparativa da performance do modelo antes e após o ajuste. Tais resultados poderiam fornecer *insights* adicionais sobre as possibilidades de aperfeiçoamento do *Stable Diffusion* e sobre a eficácia do ajuste fino como estratégia para o incremento da performance dos modelos generativos de imagem.

Adicionalmente, sugere-se a realização de estudos voltados à exploração da aplicação prática e usabilidade de tais modelos em contextos do mundo real, como design gráfico, publicidade, ensino e aprendizagem. Acredita-se que essas linhas de pesquisa podem enriquecer ainda mais o campo de estudo, trazendo contribuições valiosas e práticas para o universo da geração de imagens sintéticas a partir de descrições textuais.

REFERÊNCIAS

[1] BORJI, Ali. Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2. **arXiv preprint arXiv:2210.00586**, 2022.

[2] BROCK, Andrew; DONAHUE, Jeff; SIMONYAN, Karen. Large scale GAN training for high fidelity natural image synthesis. **arXiv preprint arXiv:1809.11096**, 2018.

[3] CHAMBON, Pierre et al. Adapting pretrained vision-language foundational models to medical imaging domains. **arXiv preprint arXiv:2210.04133**, 2022.

[4] CROWSON, Katherine et al. Vqgan-clip: Open domain image generation and editing with natural language guidance. **Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII.**, Cham: Springer Nature Switzerland, p. 88-105, 2022.

[5] CROITORU, Florinel-Alin et al. Diffusion models in vision: A survey. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 2023.

[6] DING, Ming et al. Cogview: Mastering text-to-image generation via transformers.**Advances in Neural Information Processing Systems**, v. 34, p. 19822-19835, 2021.

[7] DING, Ming et al. Cogview2: Faster and better text-to-image generation via hierarchical transformers. **arXiv preprint arXiv:2204.14217**, 2022.

[8] GOODFELLOW, Ian et al. Generative adversarial networks.**Communications of the ACM**, v. 63, n. 11, p. 139-144, 2020.

[9] GATYS, Leon A.; ECKER, Alexander S.; BETHGE, Matthias. Image style transfer using convolutional neural networks. **Proceedings of the IEEE conference on computer vision and pattern recognition.**, p. 2414-2423, 2016.

[10] HEUSEL, Martin et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. **Advances in neural information processing systems**, v. 30, 2017.

[11] HO, Jonathan; JAIN, Ajay; ABBEEL, Pieter. Denoising diffusion probabilistic models. **Advances in Neural Information Processing Systems**, v. 33, p. 6840-6851, 2020.

[12] KARRAS, Tero; LAINE, Samuli; AILA, Timo. A style-based generator architecture for generative adversarial networks. **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, p. 4401-4410, 2019.

[13] KYNKÄÄNNIEMI, Tuomas et al. Improved precision and recall metric for assessing generative models. **Advances in Neural Information Processing Systems**, v. 32, 2019.

[14] NICHOL, Alex et al. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. **arXiv preprint arXiv:2112.10741**, 2021.

[15] PARK, Taesung et al. Semantic image synthesis with spatially-adaptive normalization. **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.**, p. 2337-2346, 2019.

[16] RADFORD, Alec et al. Learning transferable visual models from natural language supervision. **International Conference on Machine Learning**, PMLR, p. 8748-8763, 2021.

[17] RAMESH, Aditya et al. Zero-shot text-to-image generation. **International Conference on Machine Learning**, PMLR, p. 8821-8831, 2021.

[18] ROMBACH, Robin et al. High-resolution image synthesis with latent diffusion models. **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.**, p. 10684-10695, 2022.

[19] RONNEBERGER, Olaf; FISCHER, Philipp; BROX, Thomas. U-net: Convolutional networks for biomedical image segmentation. **Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18**, Springer International Publishing, p. 234-241, 2015.

[20] SAHARIA, Chitwan et al. Photorealistic text-to-image diffusion models with deep language understanding.**Advances in Neural Information Processing Systems**, v. 35, p. 36479-36494, 2022.

[21] VIAZOVETSKYI, Yuri; IVASHKIN, Vladimir; KASHIN, Evgeny. Stylegan2 distillation for feed-forward image manipulation. **Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16.**, Springer International Publishing, p. 170-186, 2020.

[22] WU, Chenfei et al. Nüwa: Visual synthesis pre-training for neural visual world creation. **Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII.**, Cham: Springer Nature Switzerland, p. 720-736, 2022.

[23] YU, Fisher et al. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. **arXiv preprint arXiv:1506.03365**, 2015.

[24] ZHANG, Chenshuang et al. Text-to-image diffusion model in generative ai: A survey. **arXiv preprint arXiv:2303.07909**, 2023.

[25] ZHU, Jun-Yan et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. **Proceedings of the IEEE international conference on computer vision.**, p. 2223-2232, 2017.