

Faculdade de Arquitetura e Urbanismo da Universidade de São Paulo - FAUUSP

César Valentino Nazário Zambone

**Exploração e desenvolvimento de novos métodos de animação
habilitados pelos avanços recentes no campo da Inteligência
Artificial Generativa**

São Paulo - SP

2023

Faculdade de Arquitetura e Urbanismo da Universidade de São Paulo - FAUUSP

César Valentino Nazário Zambone

**Exploração e desenvolvimento de novos métodos de animação
habilitados pelos avanços recentes no campo da Inteligência
Artificial Generativa**

Orientador: Leandro Manuel Reis Velloso

São Paulo - SP

2023

Resumo

Estamos vivendo um momento de grandes mudanças tecnológicas e artísticas com o advento das recentes inovações no campo da IA (inteligência artificial) generativa.

Este projeto busca explorar e desenvolver diferentes métodos de animação, utilizando-se dessas novas ferramentas, que podem tanto utilizar como base animações feitas por métodos mais tradicionais (como a animação quadro-a-quadro e a animação em *motion graphics*) quanto utilizar somente imagens e texto.

Os métodos demonstrados nesse projeto utilizam o modelo de difusão *Stable Diffusion* (acessado pela WebUI ComfyUI - uma interface que permite a criação de processos complexos e personalizáveis), com o auxílio de tecnologias como o *ControlNet*, *LoRA*, *AnimateDiff* e *IP-Adapter*.

O projeto tem como objetivo produzir esses métodos, junto a demonstrações das mesmas.

Palavras-chave: animação, inteligência artificial, *ControlNet*, *Stable Diffusion*, *AnimateDiff*, *IP-Adapter*

1- Introdução

1.1- Motivação

Até pouco tempo atrás, a geração de imagens via modelos de difusão se dava unicamente por meio de *prompts* de texto (passagens de texto que orientam a geração da imagem). Porém, ao final de 2022, o módulo *ControlNet* foi publicado para o *Stable Diffusion*. Esse módulo habilita um controle muito mais fino da geração, com imagens e lineart podendo ser usadas para parametrizar a geração. A publicação desse módulo em teoria permitiria a criação de animações, por meio do uso de *linearts* animados por outros meios mais tradicionais. Porém, as animações produzidas por esse método tendem a ser inconsistentes: as formas geradas seguem as linhas e cores da animação fornecida, mas os novos detalhes gerados mudam a cada quadro. Isso não necessariamente é ruim, mas é uma limitação desse tipo de técnica.

Mais recentemente, em julho de 2023, um novo módulo chamado *AnimateDiff* foi publicado. Ele busca justamente trazer consistência entre quadros de uma animação, na hora de sua geração.

O meu interesse prévio na tecnologia da geração de imagens por meio de *prompts* de texto aliado ao meu interesse e trabalho com animações quadro-a-quadro e *motion design* motivaram a criação desse projeto. Além disso, o ambiente incerto para muitos artistas, que tendem a reagir com animosidade para com essas novas tecnologias, também foi um grande motivador: apresentar as potencialidades de uma nova ferramenta pode vir a habilitar outros artistas a abraçarem a tecnologia de formas criativas.

1.2 - Justificativa

Embora o campo da arte generativa por meio de *prompts* de texto tenha avançado imensamente nos últimos 4 anos, esse avanço acabou se limitando principalmente à parte técnica, da criação dessas ferramentas. Justamente por conta dessas ferramentas serem novas, ainda há muito a ser explorado no que se tratam as diferentes possibilidades de uso das mesmas, do ponto de vista criativo.

Com a introdução do módulo *ControlNet*, ter um controle mais fino sobre a imagem a ser gerada passou a ser possível. A composição da imagem, por exemplo, passou a ser diretamente controlável pelo usuário. Com esse controle mais preciso, passa a ser possível criar imagens que usam essas técnicas com muito mais intenção e direcionamento criativo, o que pode ser muito mais interessante do ponto de vista do design gráfico do que apenas diretamente transformar passagens de texto em imagens.

Este é um projeto experimental, que servirá também para coleta de dados e formulação de novas técnicas que utilizam as ferramentas citadas de diferentes formas. Entender os modos de uso de uma ferramenta é essencial para poder usá-las de forma efetiva e criativa, e esse projeto busca demonstrar justamente isso, de forma que, possivelmente, contribua com a adoção dessas e outras técnicas similares.

1.3 - Contexto

O campo da arte imagética generativa por meio de *prompts* de texto é um campo nascente. Esse campo teve início publicamente com o lançamento do modelo *Image GPT*, divulgado em junho de 2020 (ainda que o modelo em si não tenha sido publicado) pela sua desenvolvedora, a *OpenAI*. Sua tecnologia tem base na mesma tecnologia dos *LLMs* (*Large Language Model*, ou Grande Modelo de Linguagem) que possibilitam a existência do *ChatGPT*, por exemplo, mas seu modelo foi treinado para detectar e prever relações entre pixels ao invés de palavras, como é o caso da maioria dos *LLMs*. Isso fez com que o *Image GPT* pudesse completar imagens com base apenas nos pixels presentes nela. Ainda não haviam *prompts* de texto que guiassem a geração das imagens, mas a existência do modelo provou ser possível a geração de imagens através desses modelos.

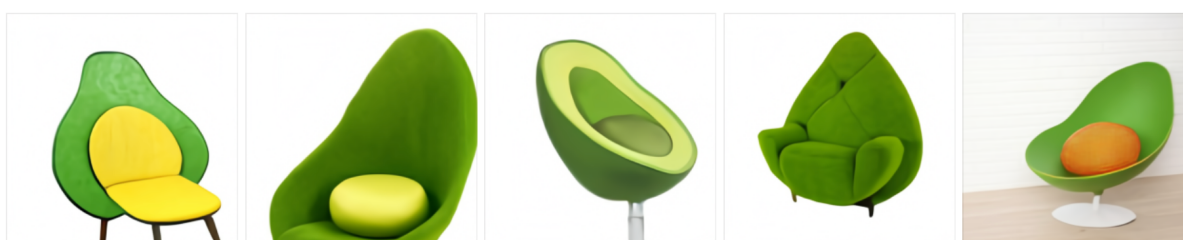
Figura 1 - Exemplos de imagens geradas pelo Image GPT¹



fonte: OpenAI (2020)²

Em janeiro de 2021, a OpenAI divulgou duas conquistas técnicas simultaneamente: o *Dall-E* e o *CLIP*. O *Dall-E* foi o primeiro modelo de geração de imagens por meio de prompts de texto divulgado publicamente (ainda que o acesso a ele fosse restrito apenas aos funcionários da empresa). Pela primeira vez, foi demonstrada a possibilidade de se gerar imagens a partir de qualquer frase escrita por um usuário. O *CLIP* (*Contrastive Language–Image Pre-training*, ou Pré-treino Comparativo de Linguagem-Imagem) é uma das tecnologias que permitiram a criação do *Dall-E*, pois é ele que determina a correlação semântica de palavras e frases com imagens. Em suma, o *CLIP* é um modelo que avalia a proximidade entre uma descrição e uma imagem. Por meio dele que a geração das imagens do modelo *Dall-E* é orientada. Diferentemente do modelo de geração de imagens, porém, o *CLIP* foi publicado como código-aberto, habilitando qualquer pessoa a utilizar o código.

Figura 2 - Imagens geradas pelo modelo Dall-E, a partir do prompt “an armchair in the shape of an avocado” (em português, “uma poltrona na forma de um abacate”)



fonte: OpenAI (2021)³

¹ A primeira imagem, da esquerda para a direita, é a imagem que foi alimentada ao modelo, com a metade inferior dela estando oculta para que o modelo possa completá-la. A última imagem é a imagem original completa. As imagens do meio são produto do modelo.

² Disponível em: <<https://openai.com/research/image-gpt>>. Acesso em 03 dez. 2023

³ Disponível em: <<https://openai.com/research/dall-e>>. Acesso em 03 dez. 2023

Poucos dias depois da publicação do *CLIP*, o usuário do twitter @advadnoun utilizou o código recém publicado e, com ele, criou e publicou abertamente um programa chamado *BigSleep* (Grande Sono), que utilizava o *CLIP* juntamente a um modelo criado pela *DeepMind* chamado *BigGAN* (*Big Generative Adversarial Network*, ou em tradução livre, Grande Rede Generativa Adversarial) para que se possa gerar imagens a partir de um *prompt* de texto. Esse foi o primeiro modelo de geração de imagens por meio de *prompts* de texto acessível ao público geral.

Figura 3 - Imagens geradas pelo *BigSleep*.⁴



fonte: Página do *BigSleep* no GitHub (2020)⁵

Por volta de fevereiro de 2021, tomei conhecimento desse modelo e comecei a realizar experimentos com ele. As imagens geradas pelo *BigSleep* são muito menos coerentes que as geradas pelo *Dall-E* (quanto mais os modelos de hoje), mas o seu uso já permitia vislumbrar as possibilidades que um modelo mais avançado poderia fornecer. Foram por meio destas experimentações que meu interesse ativo na tecnologia de geração de imagens por meio de texto surgiu.

Em março do mesmo ano, @advadnoun inaugurou uma nova técnica, unindo o CLIP com um outro modelo de geração de imagens, o *VQGAN* (*Vector Quantized Generative Adversarial Networks*, ou “Redes Adversariais Generativas Quantizadas por Vetores”). Em abril, ele publicou para os assinantes de seu Patreon (uma plataforma de financiamento de artistas e criadores independentes) uma versão mais refinada, chamada *LatentVisions*. Foi por conta dessa versão mais refinada que eu aprofundi meu interesse no assunto, e com ela que eu comecei a produzir

⁴ À esquerda, uma imagem gerada a partir do *prompt* “a lonely house in the woods” (uma casa solitária na floresta). No centro, uma imagem gerada a partir do *prompt* “the tragic intimacy of the eternal conversation with oneself” (a intimidade trágica da eterna conversa consigo mesmo). À direita, uma imagem gerada a partir do *prompt* “artificial intelligence” (inteligência artificial).

⁵ Disponível em: <<https://github.com/lucidrains/big-sleep>>. Acesso em 03 dez. 2023

imagens em volumes maiores. Isso se deu por conta não só do aumento da capacidade técnica do modelo (que ainda assim é bem inferior a do *Dall-E* no quesito da coerência geral da imagem), como também por conta das diferentes formas que o modelo interpreta os *prompts*, além dos detalhes aparentemente aleatórios que ele gera (como a textura metálica de plástico presente na **Figura 4**).

Figura 4 - Imagens geradas com o *LatentVisions* a partir do prompt “human plant hybrid” (em português, “híbrido entre homem e planta”).



fonte: *Própria* (2021)

Em Janeiro de 2022, a OpenAI divulgou a sequência de seu modelo original, com o nome *Dall-E 2*. O modo de uso dos dois era idêntico, mas a resolução e a coerência das imagens geradas pelo novo modelo eram muito melhores. Inicialmente o modelo era acessível somente aos funcionários da empresa, assim como ocorreu com a primeira iteração do modelo um ano antes, mas em setembro o acesso ao modelo foi aberto ao público, e pela primeira vez um modelo de geração de imagens de alta fidelidade estava disponível ao público (ainda que mediante à compra de créditos a serem gastos a cada geração).

Figura 5 - Imagens geradas pelo *Dall-E 2* a partir do prompt “An astronaut riding a horse in photorealistic style” (em português, “Um astronauta andando a cavalo em estilo fotorealista”)



fonte: *OpenAI* (2022)⁶

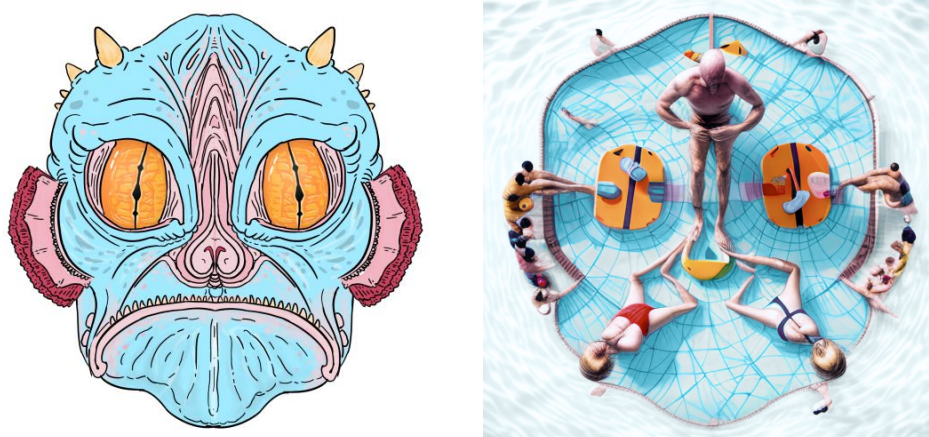
⁶ Disponível em: <<https://openai.com/dall-e-2>>. Acesso em 03 dez. 2023

Hoje em dia os principais modelos são, em larga parte, o *Stable Diffusion* (um modelo de código aberto produzido pela *Stability AI*), o *MidJourney* (um modelo de código fechado produzido pelo laboratório de pesquisa *MidJourney, Inc.*) e o Dall-E 3. Todos os modelos possuem a capacidade de gerar imagens de alta fidelidade, com qualidade superior à do *Dall-E 2*. O *MidJourney* e o Dall-E 3 possuem uma barreira de entrada muito menor, já que possuem menos parâmetros ajustáveis, não exigem uma máquina própria para rodar o programa e tendem a gerar imagens com mais verossimilhança, com o Dall-E 3 em específico tendo muito mais atenção aos detalhes descritos no texto que os outros modelos disponíveis. Porém, o *Stable Diffusion* é muito mais personalizável, com acesso à diversos módulos com funções ajustáveis, e com mais capacidade de ajustes em todos os seus parâmetros. Por esses motivos, esse será o modelo usado ao longo desse projeto.

O *Stable Diffusion* possui dezenas de módulos com funções que vão desde a usabilidade do aplicativo a até o modo de geração das imagens, mas para os propósitos desse projeto, os principais módulos adicionais são o *ControlNet*, o *LoRA*, o *AnimateDiff* e o *IP-Adapter*, além da função nativa de nome *img2img*.

A função *img2img* permite utilizar uma imagem como ponto de partida para a geração da imagem. Desse modo, é possível orientar a forma geral (ainda que de forma menos precisa que o *ControlNet*) e as cores da imagem a ser gerada.

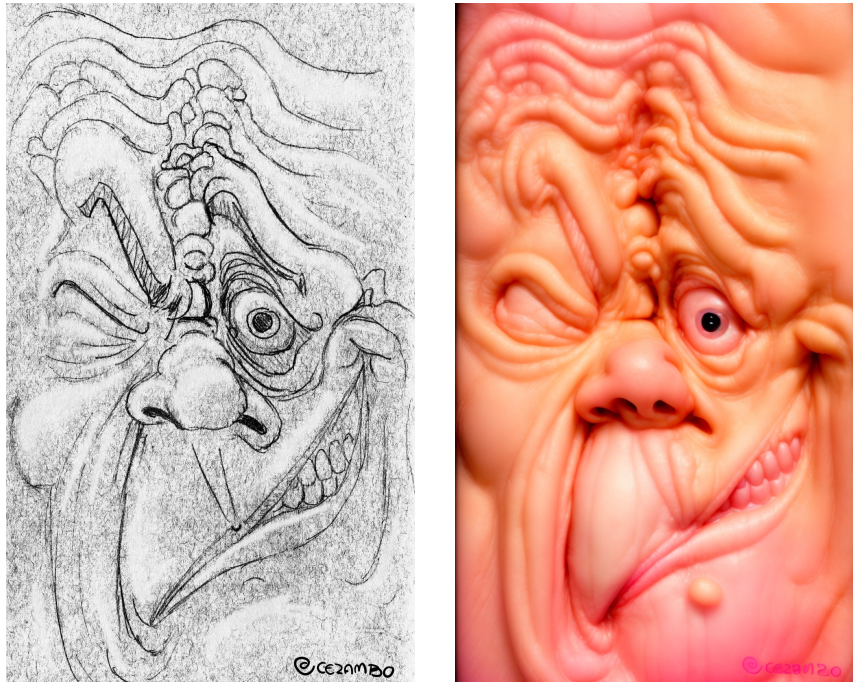
Figura 5 - À esquerda, um desenho feito por mim. À direita, uma imagem feita no *Stable Diffusion* no modo *img2img*, usando a imagem à esquerda e o *prompt* “people playing on the pool, aerial photo”.



fonte: Própria (2023)

O *ControlNet*, que é o módulo que habilita os experimentos realizados nesse projeto, permite orientar a geração das imagens por meio de uma outra imagem, que serve como “estrutura”.

Figura 6 - Exemplo de uso do *ControlNet*⁷



fonte: *Própria* (2023)

Os módulos *LoRA* (*Low-Rank Adaptation*, ou “Adaptação de Baixo Escalão”) são pequenos modelos treinados em conceitos específicos, como pessoas, objetos ou características, que podem ser usados em conjunto com os modelos maiores. O modelo *LoRA* mais usado ao longo desse projeto é o chamado *add_detail*, que como o nome sugere, permite adicionar detalhes à imagem gerada

⁷ À esquerda, um desenho feito por mim. À direita, uma imagem gerada no *Stable Diffusion* com o uso do *ControlNet*, com a imagem à esquerda como *input*, com o *prompt* positivo “face closeup, distorted face, half sad, half happy, skin folds, skin texture, analog photo, great lighting, weird, pink skin, pores” e o *prompt* negativo “illustration, drawing, bad, blurry, undefined, grey, neck, limbs, multiple faces”, além do *LoRA add_detail*.

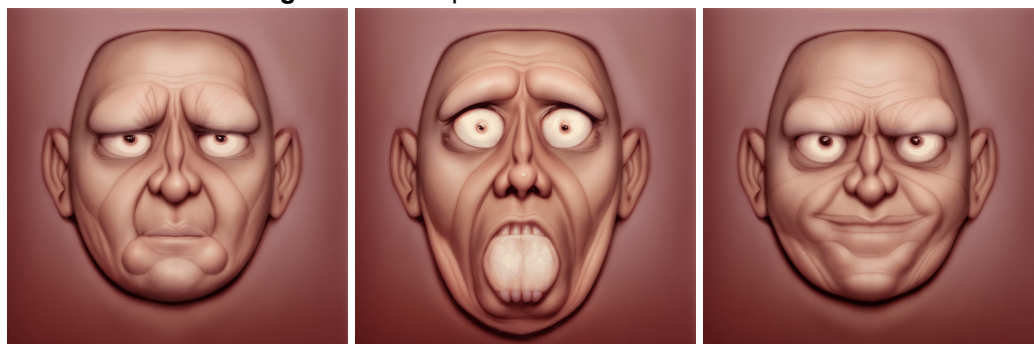
Figura 7 - Exemplo de uso do LoRA “add_detail”⁸



fonte: Própria (2023)

O módulo *AnimateDiff* é um módulo treinado com clipes de vídeo, de modo a criar um “modelo de movimento”, que pode então ser adicionado a modelos de difusão tradicionais. Desse modo, os modelos passam a ter a capacidade de criar múltiplas imagens em sequência, todas consistentes entre si, criando a ilusão de movimento. Esse módulo pode ser usado em conjunto com o *ControlNet*, possibilitando assim a geração de animações com quadros consistentes entre si, seguindo os quadros desenhados pelo usuário.

Figura 8 - Exemplo de uso do *AnimateDiff*⁹



fonte: Própria (2023).

⁸ Imagens geradas no *Stable Diffusion* com o uso do LoRA “add_detail”, com o *prompt* “closeup of a couple in new york, RAW, 8K, UHD, natural lighting, perfect day, amateur, candid, extreme close up”. Na imagem da esquerda, o LoRA estava com peso 0. Na da direita, com peso 1. Um ponto interessante de se notar é que apesar da pele, cabelos e roupas terem adquirido detalhes mais visíveis, as feições e os objetos presentes nas imagens mudaram um pouco. É extremamente difícil treinar um modelo LoRA de modo que a imagem na qual o modelo é aplicado não mude de formas imprevisíveis. Vale dizer inclusive que o LoRA usado nessas imagens é um exemplo de modelo que afeta bem pouco a imagem gerada de formas inesperadas, se comparado com outros.

⁹ Imagens geradas por uma das técnicas que virão a ser detalhadas nos próximos capítulos. A técnica utiliza, entre outros módulos, o módulo *AnimateDiff*, de modo a gerar quadros consistentes entre si

O módulo *IP-Adapter* (*Image Prompt Adapter*, ou “adaptador de prompt de imagem”) é um módulo que permite a utilização de imagens como prompts para modelos de difusão. Isso habilita um controle muito mais preciso da geração de imagens, afinal a linguagem textual muitas vezes é insuficiente pra se descrever uma imagem de forma satisfatória e completa.

Figura 9 - Exemplo de uso do *IP-Adapter*¹⁰



fonte: *Própria* (2023)

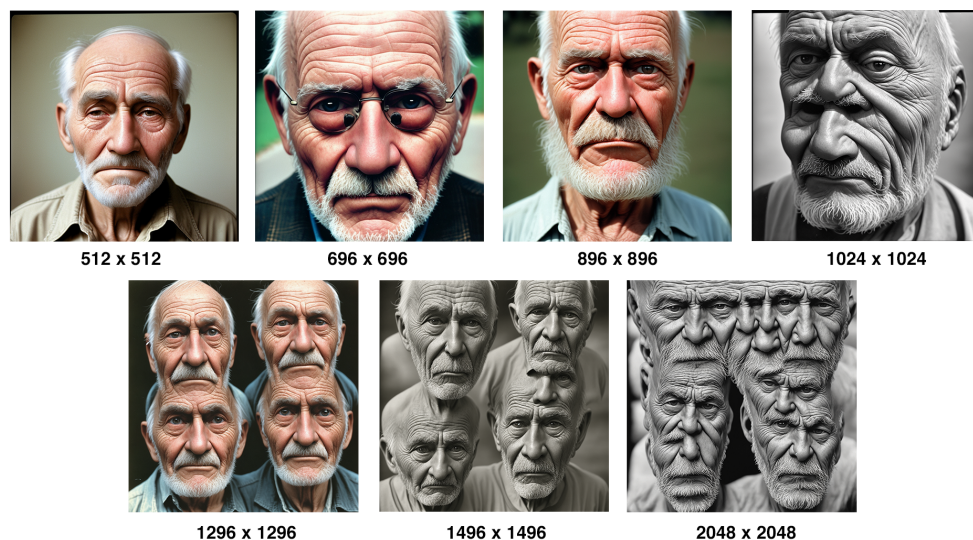
O Stable Diffusion possui diversos parâmetros, que devem ser levados em consideração na hora de se usar essa ferramenta. Esses parâmetros são o tamanho da imagem, o número de *steps*, a *seed*, o *sampler*, o *CFG*, o *denoise* e o *prompt*.

O tamanho da imagem inicial não pode variar mais de 50% na altura ou na largura do tamanho padrão de 512x512. Isso se dá porque os modelos são em sua maior parte treinados para gerarem imagens nessas dimensões, e caso o tamanho fuja muito disso, principalmente em imagens de alta resolução, detalhes “fractais”, com elementos que se repetem, tendem a aparecer.

Se for necessário gerar imagens maiores que esse limite, é necessário gerar a imagem inicialmente num tamanho que não o ultrapasse. Então, a imagem deve ser aumentada e reinserida no modelo, no modo *img2img*. Desse modo, a imagem inicial serve de estrutura pros detalhes a serem adicionados na segunda etapa.

¹⁰ À esquerda, um desenho feito por mim. Ao centro, um quadro de El Lissitzky. À direita, uma imagem feita no *Stable Diffusion* no modo *img2img*, usando a imagem à esquerda como base e a imagem do centro como *prompt*, por meio do módulo *IP-Adapter*.

Figura 10 - Imagens com *seed* travada e *prompts* idênticos, com variação na resolução, em pixels¹¹



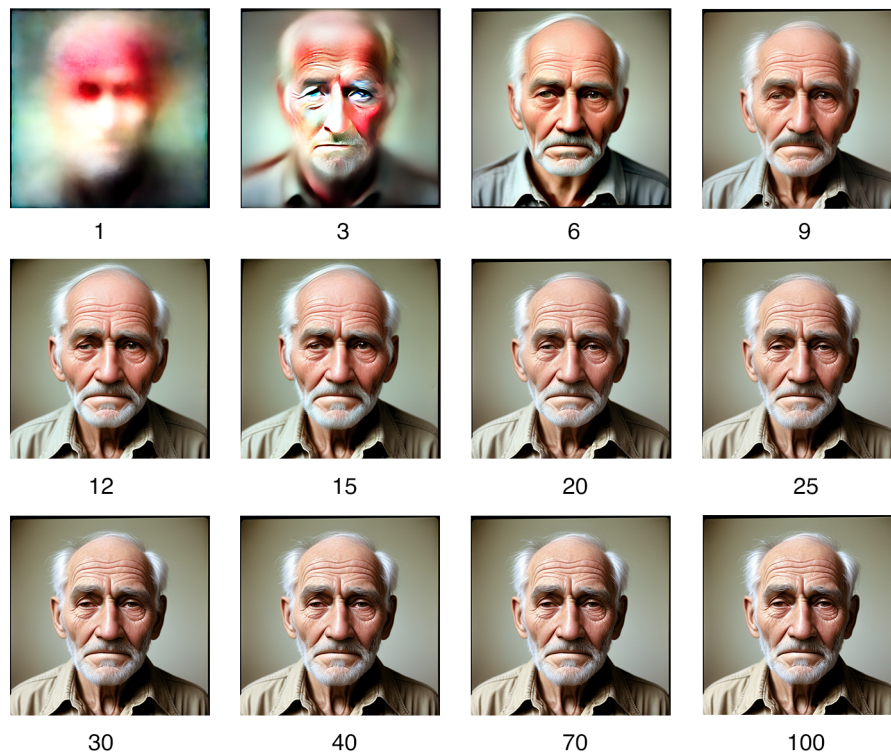
fonte: *Própria* (2023)

O *sampler* é o método pelo qual a imagem será gerada a partir de ruído. Existem dezenas de tipos de *samplers*, que se dividem em linhas gerais em 2 tipos: os *estocásticos* (que tendem a variar bastante entre os *steps*) e os *convergentes* (que tendem a convergir em uma imagem específica na medida em que os *steps* passam).

Os *steps* são o número de rodadas de remoção de ruído aplicados pelo *sampler* selecionado. Quanto maior o número, maior a qualidade e o grau de detalhes (se o *sampler* for convergente). Porém, após certo ponto, aumentar o número de *steps* tende a trazer rendimentos decrescentes, e como o número de *steps* é um dos principais fatores contribuintes com o tempo de geração, é importante achar um equilíbrio - uma quantidade de *steps* que gere resultados bons, mas que não acarrete em um tempo de geração muito grande.

¹¹ Mesmo que a *seed* seja a mesma, a imagem muda completamente quando se muda o seu tamanho. A partir do 4º exemplo, elementos começam a se repetir.

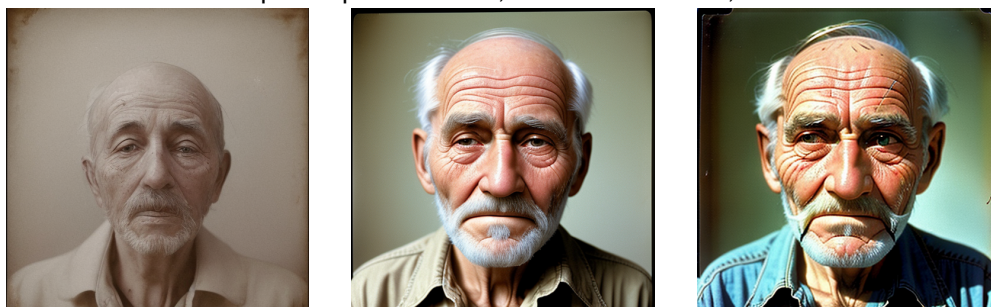
Figura 11 - Imagens com *seed* travada e *prompts* idênticos, com variação no número de *steps*



fonte: *Própria* (2023)

O parâmetro CFG (*Classifier-Free Guidance*, ou Orientação Livre de Classificadores) determina o quão “fiel” ao *prompt* de texto a imagem final será. Valores mais baixos tendem a gerar imagens que não seguem tão bem o *prompt*, e que possuem cores mais esmaecidas. Valores mais altos tendem a gerar imagens mais próximas das descritas pelo *prompt*, além de tenderem a ter cores mais vibrantes e com mais contraste, mas tendem a exigir *prompts* mais detalhados para se alcançar um resultado interessante. Valores altos demais, porém, acabam gerando artefatos visuais e diminuindo a fidelidade da imagem.

Figura 12 - Imagens com o mesmo *prompt* e *seed*, mas com CFGs variados. Da esquerda para a direita, os CFGs são de 1, 8 e 20.



fonte: *Própria* (2023)

A *seed* é o valor numérico através da qual os ruídos necessários para a geração são gerados. O valor da *seed* tem impacto direto na imagem final, e valores diferentes podem gerar composições e estilos completamente distintos, mesmo que todos os outros parâmetros continuem iguais, enquanto *seeds* iguais com parâmetros iguais geram imagens idênticas. Por padrão, a *seed* costuma mudar de forma aleatória após cada geração, de modo a produzir imagens diferentes. Caso um estilo pareça promissor, a *seed* pode ser configurada de modo a ficar fixa, e os outros parâmetros devem ser ajustados e testados até que a imagem gerada esteja aceitável.

Figura 13 - Imagens com *seeds* diferentes, geradas a partir do *prompt* “An old man face, great analog photography”



fonte: Própria (2023)

O parâmetro *denoise* é relacionado ao modo *img2img*, e determina o quão distante da imagem base a imagem gerada será. O parâmetro pode possuir qualquer valor entre 0 e 1. Se o *denoise* for 0, a imagem produzida será idêntica à imagem base. Se o valor for 1, a imagem base deixa de ter qualquer influência sobre a imagem gerada - nesse caso, a geração acaba operando de forma idêntica ao modo padrão, de imagens geradas baseadas somente num *prompt* de texto.

Figura 14 - Imagens demonstrando o parâmetro *denoise* ¹²



fonte: Própria (2023)

¹² Da esquerda para a direita: uma imagem gerada por mim, usada como imagem base para as imagens seguintes; uma imagem gerada com *denoise* de 0.5; uma imagem gerada com *denoise* de 0,7 e uma imagem gerada com *denoise* de 1. Todas as imagens usam o mesmo *prompt* de texto: “human plant hybrid” (em português, “híbrido entre homem e planta”).

Com ferramentas com a capacidade de gerar imagens realísticas como essas estando tão facilmente acessíveis ao público, tornou-se quase inevitável que uma imagem gerada por esses métodos viria a causar algum tipo de confusão. Um possível primeiro exemplo de ocorrência desse fenômeno pode ter sido o caso da foto do “Papa estiloso” - uma imagem gerada por meio do *MidJourney* do Papa Francisco usando um casaco estiloso, que acabou viralizando nas redes sociais e levando muita gente a acreditar que a imagem se tratava de uma foto real. Ainda que esse primeiro caso de confusão em grande escala tenha sido relativamente benigno, ele pôde demonstrar ao público geral a capacidade de criação dessas ferramentas, e como elas poderiam ser usadas para fins maliciosos, como geração de imagens com o fim de difamar um oponente político, por exemplo.

Figura 15 - Imagem do Papa Francisco usando um casaco estiloso, gerada por meio do *MidJourney*.



fonte: Postagem no Reddit (2023)¹³

Uma outra questão polêmica que surgiu sobre esses modelos de geração de imagens origina do modo como eles são treinados. Para que esses modelos mais recentes obtenham um entendimento sobre as formas a serem geradas, eles precisam ser treinados usando imagens com descrições de seus conteúdos, num

¹³ Disponível em: <https://www.reddit.com/r/midjourney/comments/120vhdc/the_pope_drip/>. Acesso em 03 dez. 2023

formato compatível com o CLIP (De forma resumida: nesses modelos mais recentes, o CLIP é usado principalmente para converter as palavras de um *prompt* em *embeddings*: unidades numéricas que possuem algum significado que pode ser representado tanto em forma de texto quanto em forma imagética). Essas imagens muitas vezes são coletadas diretamente da internet, sem qualquer consulta prévia dos possíveis detentores de direitos intelectuais das imagens, sendo que muitos dos criadores das imagens são fotógrafos e artistas independentes que dependem de seu próprio trabalho para sobreviver. O *dataset* usado para o treino do modelo *Stable Diffusion 1.5*, por exemplo, foi o LAION-5B - um dataset que contém 5,85 bilhões de imagens coletadas da internet, com descrições em texto filtradas por meio do CLIP.

Ao passo que sim, muitas dessas imagens usadas foram feitas por artistas independentes que não receberam compensação alguma, é importante notar alguns detalhes:

- Os modelos criados por meio desses datasets não os contém, e portanto não contém as imagens usadas neles
- As imagens coletadas eram todas disponibilizadas abertamente na internet pelos próprios artistas. Nenhuma das imagens foi coletada de coleções privadas sem o consentimento dos detentores de direitos das mesmas.
- No caso do LAION-5B, as imagens sequer foram propriamente coletadas. O dataset em questão é apenas uma lista de URLs de imagens pareadas com descrições em texto correspondentes às imagens em questão. Ou seja: as legendas das imagens são ligadas à links que levam aos locais onde as imagens foram originalmente postadas.
- O processo de treino não consome as imagens usadas nem as armazena, apenas aprende padrões visuais e relações entre imagem e texto a partir delas - um processo análogo à forma como artistas humanos aprendem. Para que um artista possa se desenvolver tecnicamente, muitas vezes é necessário buscar inspiração e referências nos trabalhos artísticos de outros (ISHIGURO, C., & OKADA, T. ,2017). Se aprender padrões e estilos através da

observação de imagens é uma violação de direitos, não seria um salto lógico muito grande dizer que ir a um museu para evoluir como artista também é uma violação de direitos.

Independente da relevância desses pontos, esses artistas profissionais independentes continuam em uma situação precária que piora a cada dia, e o futuro do mercado para artistas como esses fica mais incerto a cada avanço técnico que ocorre no campo da produção de imagens habilitada por inteligência artificial.

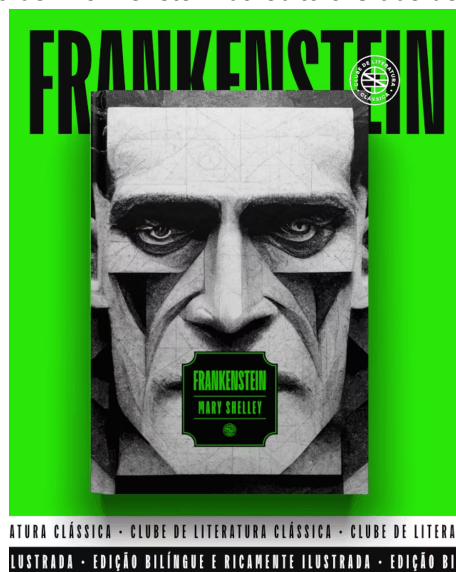
Não é preciso, porém, esperar pra que os impactos dessa tecnologia se tornem visíveis. Muitos artistas já relatam um volume menor de contratação pelos seus serviços, além de uma dificuldade maior para conseguir entrar nas indústrias audiovisuais em posições iniciantes, que hoje já estão tomadas pelo uso dessas novas tecnologias (CARTER, 2023). Como forma de mitigar os impactos dessa tecnologia, muitos artistas estadunidenses estão se unindo com o intuito de codificar a arte gerada por essas novas ferramentas de modo a impossibilitar a atribuição de copyright à elas, por meio de ações legais e audiências (GRANT, 2023).

O Escritório de Direitos Autorais dos EUA tende a concordar com a impossibilidade de se aplicar registros de direitos autorais à obras geradas por essas tecnologias. O guia oficial do escritório afirma explicitamente que "obras produzidas por uma máquina ou mero processo mecânico que opera de forma aleatória ou automática sem qualquer entrada criativa ou intervenção de um autor humano" não são elegíveis (GRANT, 2023). Claro, leis e normas estão sempre abertas à interpretação, e não seria absurdo dizer que a escolha das palavras usadas na hora da geração das imagens pode configurar intervenção artística humana, nem que o uso de imagens geradas por IA em uma obra maior excluiria a possibilidade de registro da obra.

Este segundo exemplo evoca o recente caso da desclassificação no Prêmio Jabuti da nova edição do *Frankenstein*, de Mary Shelley, publicada pelo Clube da Literatura Clássica. A obra contém cerca de 50 imagens geradas com o uso do *MidJourney*. Uma obra que faz o uso de uma ferramenta de geração de imagens por IA estar entre os finalistas do Prêmio Jabuti, a mais tradicional premiação da literatura brasileira, causou muita polêmica online, principalmente entre os artistas do Twitter. Ainda que o uso da ferramenta tenha sido anunciado dentro da obra, muitos dos jurados, como o cartunista André Dahmer e o ilustrador Eduardo

Baptistão, disseram não saber à época que as imagens tinham sido geradas por IA (GABRIEL, 2023). Em comentário para o Estadão, Baptistão completa que “O regulamento diz que o autor responde pela originalidade, autenticidade e/ou autoria do material inscrito”, e que “Na minha opinião, a IA não se enquadra nesses critérios. Mas quem vai determinar o que fazer neste caso é a CBL [Câmara Brasileira do Livro, que organiza o Prêmio Jabuti], baseada no parecer da sua assessoria jurídica” (QUEIROZ, 2023). Por outro lado, Vicente Pessoa, o designer responsável pela criação das imagens, argumenta que o processo de criação de imagens para um livro como esse é muito mais ativo do que simplesmente inserir palavras como prompt. Para o Estadão, ele diz que “Foram feitas mais de 1.200 imagens, das quais a gente aproveitou 50. Talvez fosse menos trabalhoso ilustrar na mão” (QUEIROZ, 2023). Com a capacidade de produção de imagens tão mais agilizada, o processo de curadoria tornaria-se muito mais importante que o de criação em si, nesse caso.

Figura 16 - Edição de 'Frankenstein' da editora Clube de Literatura Clássica



fonte: Estadão (2023)¹⁴

As grandes empresas produtoras desses modelos, porém, não estão completamente passivas quanto a essas reações e desenvolvimentos, principalmente no que se diz a questão dos direitos autorais, e ao pagamento dos

¹⁴ Disponível em: <https://www.estadao.com.br/cultura/literatura/premio-jabuti-livro-inteligencia-artificial-categoria-melhor-ilustracao-nprec/>. Acesso em 03 dez. 2023

criadores dos dados de treinamento desses modelos (no caso, escritores e artistas visuais). Cada uma dessas grandes empresas apresenta um argumento diferente (DAVIS, 2023):

- A Meta afirma que a compensação financeira pelos dados usados no treinamento seria tão ínfimo, levando em conta o tamanho do dataset completo e o valor gerado pelos modelos, que o trabalho imenso de identificar a fonte de cada componente do dataset não valeria a pena.
- A Google argumenta que o treino de IAs é análogo à leitura de livros, e como adquirir conhecimento através da leitura de trabalhos com registro de copyright não configura uma violação desse registro, o treino desses modelos com o uso de trabalhos registrados também não deveria configurar.
- A Microsoft argumenta que a imposição de novas regulamentações iria “esfriar” o desenvolvimento de modelos IA por empresas e grupos menores, com o desenvolvimento desses modelos se tornando acessível apenas às empresas grandes, que seriam as únicas com os recursos para comprar esses direitos e desenvolver esse modelos IA.
- A Anthropic avalia as regulamentações atuais como suficientes. A empresa afirma que as leis atuais protegem todos os interessados sem bloquear o desenvolvimento futuro das IAs. Além disso, ela também argumenta que a cópia dos dados é apenas um passo intermediário, visto que os dados usados no treinamento não são literalmente copiados e colados na hora da comunicação dos programas com os usuários.
- A Adobe utiliza o precedente legal do caso “Sega v. Accolade” como argumento. No caso, a Accolade utilizou uma cópia intermediária de um programa da Sega para realizar um processo de engenharia reversa do funcionamento de seus consoles, assim possibilitando o desenvolvimento de jogos de desenvolvedoras independentes para a plataforma. O tribunal considerou a ação da Accolade como benéfica e justa, visto que ela permitiu uma maior expressão criativa dos desenvolvedores independentes. A Adobe argumenta que, assim como no caso “Sega v. Accolade”, a cópia intermediária de conteúdo

protegido para fins de engenharia reversa ou para criar novos trabalhos (como os modelos IA) pode ser considerada *fair use*.

Enquanto que a legislação, regulamentação e ações legais contra grandes empresas são fatores importantíssimos para o futuro dessas tecnologias, a capacidade individual de afetar essas causas por aqueles mais afetados (artistas visuais e escritores) segue muito pequena. Claro, é possível se organizar e promover essas ações em certa medida (como mencionado anteriormente), mas outras ações de impacto mais imediato podem ser tomadas, como a adoção dessas novas técnicas no processo criativo dos artistas, mesmo que de forma parcial ou intermediária.

Outro impacto dessas tecnologias, que segue muito pouco discutido, é a democratização do acesso à produção artística. A arte por muito tempo foi um meio de expressão restrito, limitado àqueles virtuosos, que tinham a capacidade técnica necessária para apelar aos gostos de mecenas. A produção artística, como consequência, acabou restrita a poucos meios de representação, e a produções de baixo volume. Esse é um dos problemas da arte como a conhecemos, compreendidos por grupos vanguardistas como os Dadaístas e Situacionistas, que por diversos meios buscaram remediá-lo. Com a invenção do *Ready-Made*, por exemplo, o virtuosismo se desvincula do processo artístico (ARCURI, 2014), e com as pinturas industriais de Giuseppe Pinot-Gallizio, a relação produtiva e mercadológica das artes acaba mudando também (HOME, 2004). Porém, ainda que esses esforços tenham mudado a compreensão do que é arte, muitos dos meios de representação continuaram restritos, exigindo investimento monetário e de tempo que grande parte da população não tem como oferecer. Além disso, essas intervenções conceituais da arte, ainda que em certa medida democratizem a arte, são inerentemente limitadas: os meios e métodos oferecidos pelos grupos são sempre uma nova maneira de se fazer uma nova arte, distinta da arte produzida por meios tradicionais.

A tecnologia de geração de imagens por meio de modelos de difusão muda mais profundamente a relação dos artistas com o público, e dos artistas com a arte. Diferentemente dos métodos produzidos pelos grupos vanguardistas, esse novo método tem a capacidade de recriar diversos estilos existentes e ainda não existentes, sendo portanto imensamente mais capacitado. Tradicionalmente, há uma

barreira entre artistas (que produzem a arte e detém os meios e o virtuosismo) e o público (que passivamente observa a arte). Essa relação muda completamente quando, em teoria, através do uso dessas ferramentas, qualquer um poderia criar arte similar àquela que eles consomem. Pela primeira vez, surge uma tecnologia com a capacidade de democratizar verdadeiramente o acesso ao processo de criação artística e, portanto, mudar a própria relação do público com a arte.

1.4 - Objetivo

Produzir métodos de animação utilizando técnicas de animação tradicionais, como a animação quadro-a-quadro e o *motion graphics*, e técnicas generativas com o uso de IA, através do modelo de difusão *Stable Diffusion* (acessado pela WebUI *ComfyUI* - uma interface que permite a criação de processos complexos e personalizáveis), com o auxílio de tecnologias como o *ControlNet*, *LoRA*, *IP-Adapter* e *AnimateDiff*, e assim, familiarizar artistas com as possibilidades oferecidas por essas novas técnicas.

2 - Processo Criativo

A ideia para esse projeto iniciou-se em parte por conta da reação do público - e principalmente de artistas - nas redes sociais à liberação do acesso em larga escala a modelos de difusão - inicialmente, ao modelo *Dall-E 2*. Com esses novos modelos, tornou-se possível com apenas uma única frase como *input* criar-se peças gráficas em segundos, podendo elas serem fotorealísticas ou emularem técnicas gráficas que tradicionalmente exigiriam muito tempo de prática de um artista tradicional, por exemplo.

Essas ferramentas naturalmente causaram grandes discussões sobre a validade das peças gráficas produzidas por esses métodos como “arte de verdade”, entre outras coisas. O objetivo inicial era produzir um videoensaio cômico abordando esse tema, utilizando diversas técnicas que envolvem IA em sua produção. Antes de se escrever o roteiro, uma análise geral foi feita sobre os principais argumentos feitos por oponentes à arte feita com IA, com o roteiro sendo escrito de forma a buscar refutá-los com argumentos e exemplos históricos (como a introdução da fotografia no meio artístico, por exemplo). Além dessa análise, foi

iniciado também um processo de experimentação com essas ferramentas anteriormente citadas. Por meio dessas experimentações, foi possível se observar o potencial criativo dessas técnicas, e como elas poderiam acabar não sendo tão bem exploradas por meio do formato do videoensaio.

Por conta disso e por conta também de um conselho dado pelo meu orientador Leandro, o roteiro foi reescrito, agora com o intuito de se fazer uma animação de comédia. O novo roteiro seguia o formato das televidas, popularizado pela polishop e pelo canal Shoptime, e fala sobre um produto fictício, que busca “educar” (tanto no sentido do ensino quanto no sentido dos “bons modos”) crianças de pais sem tempo para dar atenção aos filhos. O roteiro foi também convertido em um storyboard detalhado.

Após a apresentação do TCC 1 à banca, obtive a devolutiva de que o roteiro não dialogava tanto com o conceito técnico apresentado. O roteiro acabou sendo reescrito mais uma vez, agora se tratando de um outro produto: o “raio lindificador”. Esse raio teria a capacidade de “embelezar” qualquer coisa, transferindo a beleza de um objeto sacrificial para o alvo do raio. O roteiro foi pensado ao redor de uma técnica de animação que gradualmente introduz características à uma parte da imagem - no caso do raio lindificador, o objetivo seria deixar o alvo gradualmente mais estereotipicamente bonito, de forma grotesca. Um storyboard detalhado também foi produzido, com cada técnica que seria usada também sendo especificadas abaixo de cada uma das cenas.

Infelizmente, o novo roteiro acabou ficando muito grande para ser feito apenas por uma pessoa, em apenas um semestre. Tentei cortar algumas cenas, mas senti que o resultado ficou com um ritmo muito pior. No final, optei por focar nas técnicas, e nesse meio tempo desenvolvi também mais algumas outras, em adição às previamente desenvolvidas.

Nessa segunda etapa do TCC, busquei também trocar de ferramenta. Na primeira etapa, estava utilizando a WebUI *Automatic1111* - uma interface muito mais complexa e personalizável que a maioria dos serviços de geração de imagens, como a do *MidJourney* e do *Dall-E 3*. Porém, ela possui algumas limitações graves, principalmente para a geração de muitas imagens em sequência - algo essencial para a criação de animações. A geração de cada quadro exigia múltiplos passos manuais pra serem gerados, além do preparo prévio de todos os quadros base a serem usados. Um único segundo de animação tradicionalmente contém de 12 a 24

imagens - produzir esses quadros manualmente rapidamente se demonstraria inviável.

Buscando uma ferramenta mais personalizável e automatizável, encontrei o *ComfyUI* - uma interface inspirada por fluxogramas, que habilita a criação de algoritmos complexos e automatizados. A interface apresenta diversos tipos de blocos com diferentes parâmetros e funções, que podem ser conectados de diferentes formas. Com essa nova ferramenta, todos os processos explicados no relatório parcial apresentado ao final do TCC 1 puderam ser automatizados em alguma medida.

3 - Resultados

Os métodos desenvolvidos utilizam um ou mais algoritmos desenvolvidos na WebUI *ComfyUI*, que podem ter seus diferentes parâmetros calibrados para se obter diferentes efeitos. Todos os processos usam o *LoRA* “*add_detail*”, que confere mais detalhes às imagens geradas.

Todos os vídeos gerados pelas técnicas que serão apresentadas, além de outras animações que não foram incluídas no relatório, podem ser acessados por meio do seguinte QR code:

Figura 17 - O QR code que leva às pastas com os vídeos¹⁵



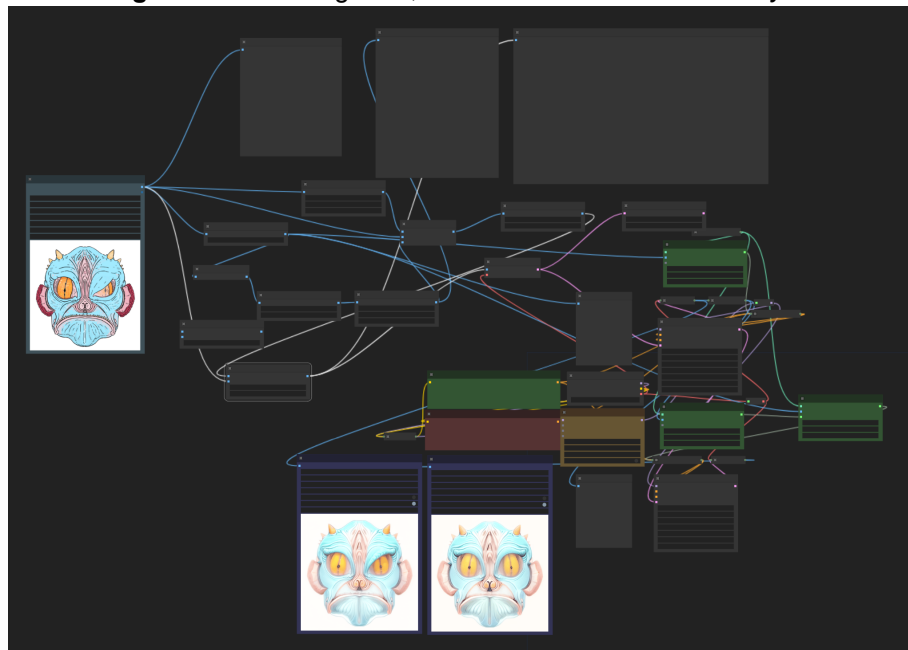
fonte: Própria (2023)

¹⁵ O link do QR code:

https://drive.google.com/drive/folders/1klAvEvaRmqtQFpaQee41w6323hd3M_P4?usp=drive_link

3.1 - Processo Base - AnimateDiff

Figura 18 - O fluxograma, como visto na interface ComfyUI.



fonte: Própria (2023)

O Processo Base tem como inputs uma animação de até 24 quadros, um prompt negativo e um prompt positivo. Ele utiliza os módulos *ControlNet* e *AnimateDiff*, além de módulos de edição de imagem. Esse processo é usado como base por alguns outros métodos que serão apresentados, mas o resultado produzido por ele pode ser usado como produto final. O algoritmo faz diversos processos automaticamente:

- Extrai as linhas da animação e as usa como input no módulo *ControlNet*, que influencia todos os próximos passos;
- Clareia as linhas presentes nos quadros da animação para o uso no modo *img2img* (isso acontece pois, na medida que fui realizando testes, percebi que as linhas pretas acabavam muito marcadas na animação final, o que pode acabar atrapalhando em animações onde o objetivo é obter um estilo mais realístico);
- Renderiza os quadros com o auxílio do módulo *AnimateDiff*, produzindo assim uma animação de quadros consistentes;

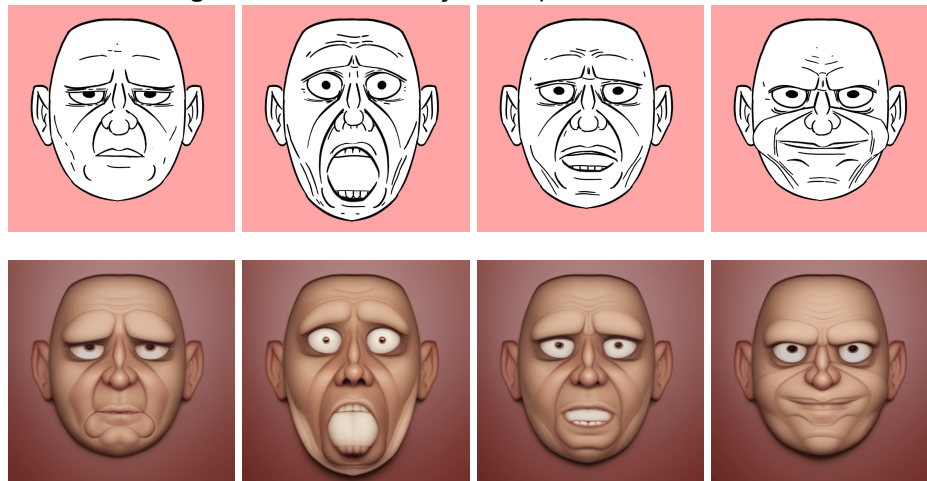
- Aumenta o tamanho dos quadros produzidos no passo anterior e os renderiza novamente, ainda utilizando o módulo AnimateDiff (desse modo, os quadros adquirem mais detalhes);
- Une todos os quadros numa nova animação.

Figura 19 - Demonstração comparativa da técnica¹⁶



fonte: Própria (2023)

Figura 20 - Demonstração comparativa da técnica¹⁷



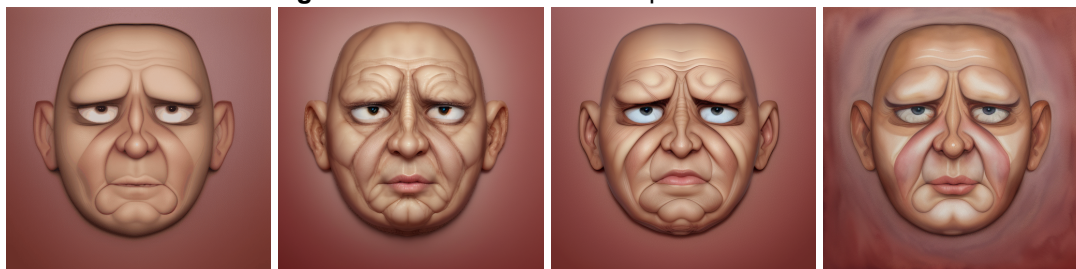
fonte: Própria (2023)

3.2 - Refinador de quadro para uso em IP-Adapter

¹⁶ Na fileira de cima, quadros da animação original. Na fileira de baixo, os quadros correspondentes da animação produzida. Nesse caso, o valor do *denoise* foi 0.5.

¹⁷ Na fileira de cima, quadros da animação original. Na fileira de baixo, os quadros correspondentes da animação produzida. Nesse caso, o valor do *denoise* foi 1.

Figura 22 - Resultados obtidos pelo método ¹⁸

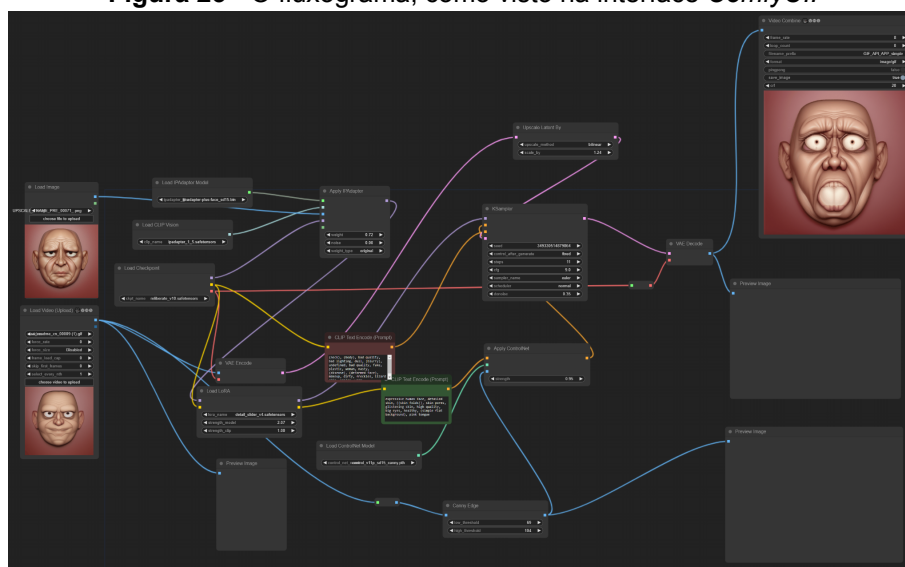


fonte: Própria (2023)

3.3 - Modificação de quadros por meio do IP-Adapter

Esse processo foi pensado com o intuito de se aplicar a imagem selecionada no método anterior aos quadros da animação produzida pelo Processo Base, por meio do módulo *IP-Adapter*. Esse método pode também, em conjunto com o módulo mencionado, ser orientado por *prompts* de texto.

Figura 23 - O fluxograma, como visto na interface *ComfyUI*.



fonte: Própria (2023)

Esse processo extrai as linhas da animação inserida, que são usadas pelo módulo *ControlNet*. Os quadros da animação então são usados no modo *img2img* (na configuração padrão, com o *denoise* entre 0,3 e 0,4), e a imagem do *IP-Adapter*

¹⁸ À esquerda, o quadro extraído da animação inserida no fluxograma. As outras imagens são diferentes resultados obtidos pelo método.

(gerada no processo anterior) somada aos os prompts de texto (quando presentes) são então aplicados. Note que, diferentemente do Processo Base, esse método não usa o módulo *AnimateDiff*, e portanto não há um modelo que coordena a coerência entre os quadros gerados.

Figura 24 - Demonstração comparativa da técnica.¹⁹



fonte: Própria (2023)

Inicialmente, essa etapa seria outra etapa intermediária - a animação gerada por esse método é usada no próximo método a ser apresentado. Porém, o resultado por esse método é bem interessante justamente por conta dessa inconsistência entre os quadros. Além disso, o módulo *AnimateDiff* costuma minar um pouco a capacidade de geração de detalhes dos modelos, e por conta disso as animações geradas por esse método possuem quadros muito mais detalhados.

Após ter essa realização, percebi que os parâmetros desse fluxograma poderiam ser ajustados pra se criar diferentes tipos de animações, e a imagem inserida no módulo IP-Adapter poderia ser qualquer imagem, e não só as geradas pelo método anterior.

¹⁹ Na fileira de cima, quadros da animação produzida pelo Processo Base, inserida nesse algoritmo. Na fileira de baixo, os quadros correspondentes da animação produzida. Note o nível de detalhe muito superior, e como as rugas do personagem mudam de forma aleatória entre os quadros.

Figura 25 - Demonstração comparativa da técnica.²⁰



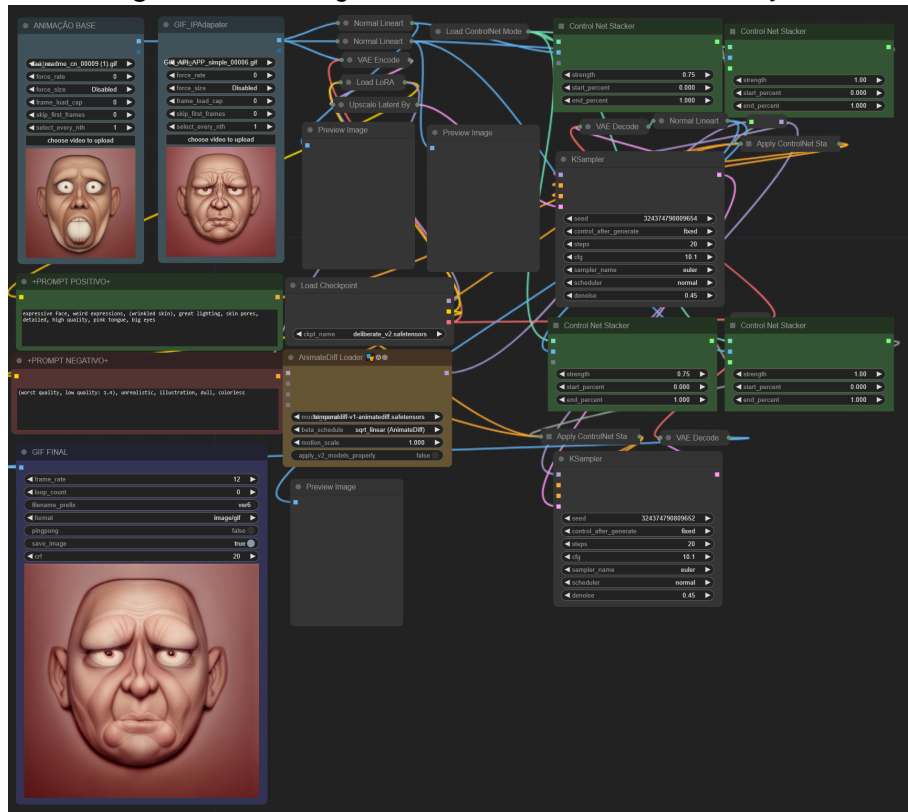
fonte: Própria (2023)

3.4 - Processo Final - AnimateDiff

Este é o último processo desenvolvido por mim para animações com base em *AnimateDiff*. Ele utiliza a animação gerada pelo método anterior como base no modo *img2img*, além da animação original, que serve para extração de linhas, que são então usadas pelo módulo *ControlNet*. Através de testes, percebi que uma parte dos detalhes gerados no processo anterior são preservados, mas agora com mais estabilidade temporal.

²⁰ Na fileira de cima, quadros da animação produzida pelo Processo Base, inserida nesse algoritmo. Na fileira do meio, os mesmos quadros, após passarem pelo processo com o *ControlNet* com peso 0 (ou seja, desligado) e com *denoise* 0.5. Na fileira de baixo, os mesmos quadros da fileira de cima, após passarem pelo processo com o *ControlNet* com peso 0.5 e com *denoise* 1 (ou seja, sem modo *img2img*). Note como os quadros da fileira do meio saem um pouco do molde dos quadros originais, mas sem perder as cores presentes neles. Note também como os quadros da última fileira são extremamente diferentes entre si, mas ainda mantém a forma geral presente nos quadros originais, ainda que com linhas mais arredondadas e geométricas. É importante notar que ambas as fileiras das imagens que passaram pelo processo demonstrado usaram a mesma imagem no módulo *IP-Adapter*: o quadro “Vence os Brancos com a Cunha Vermelha”, de El Lissitzky.

Figura 26 - O fluxograma, como visto na interface ComfyUI.

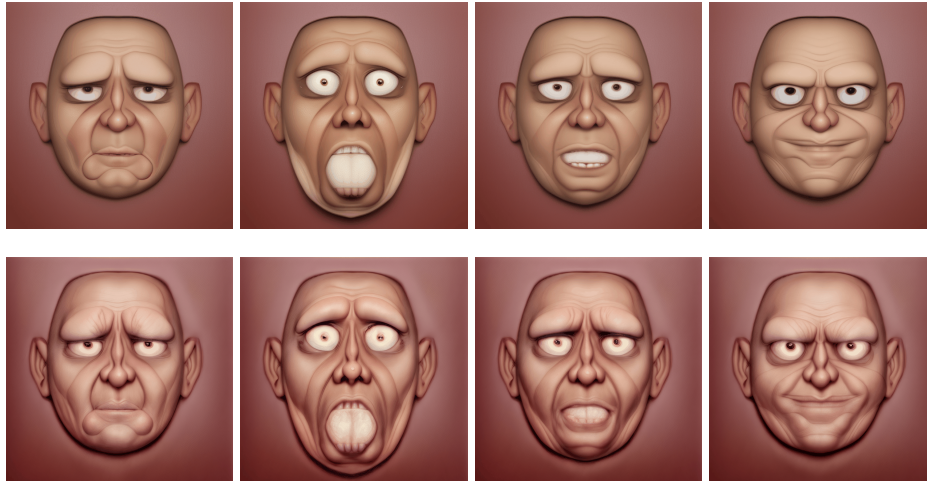


fonte: Própria (2023)

O processo ocorre em 2 etapas. Na primeira, a animação gerada pelo processo anterior é usada como base no modo *img2img*, com *denoise* em 0.45, e com 2 módulos ControlNet em conjunto: O primeiro sendo baseado nas linhas da animação do processo anterior, em peso 0.75, e o segundo sendo baseado nas linhas da animação original, em peso 1 - desse modo, a estrutura da animação original toma mais importância, mas os detalhes produzidos pelo processo anterior não são perdidos. A animação produzida fica mais temporalmente coesa, mas não totalmente.

O segundo passo repete exatamente o que ocorre no primeiro, mas agora, no lugar da animação produzida pelo processo anterior, a animação produzida no primeiro passo é utilizada. O resultado é uma animação mais detalhada que a produzida no Processo Base, mas sem perder a coesão temporal, como ocorre no processo anterior, que não utiliza o módulo *AnimateDiff*.

Figura 27 - Demonstração comparativa da técnica.²¹

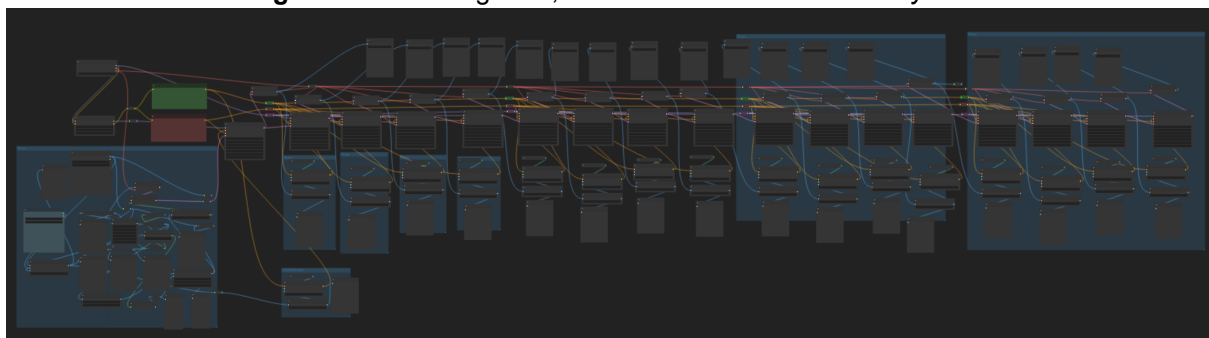


fonte: Própria (2023)

3.5 - Mudança Gradual

Esse método foi desenvolvido ao final do primeiro semestre de 2023, antes do lançamento do *AnimateDiff*, e portanto não o utiliza em seus processos. O seu funcionamento é bem diferente do funcionamento dos outros métodos até então apresentados, pois não utiliza uma animação base, mas apenas uma imagem e um prompt de texto, que deve descrever algo bem diferente do conteúdo da imagem. Por conta disso, o resultado desse método é muito mais imprevisível.

Figura 28 - O fluxograma, como visto na interface *ComfyUI*.



fonte: Própria (2023)

Com a imagem inicial e o prompt selecionados, o algoritmo automaticamente isola o fundo do objeto em destaque na imagem, de modo a afetar apenas o objeto. Então, a imagem é usada como base no modo *img2img*, com o *denoise* de 0.1, e da

²¹ Na fileira de cima, quadros da animação produzida pelo Processo Base, inserida nesse algoritmo. Na fileira de baixo, os quadros correspondentes da animação produzida.

mesma imagem são extraídas as formas, para serem usadas pelo *ControlNet*. A imagem então é salva e é repassada para o módulo seguinte, que usa o quadro recém gerado como base no modo *img2img* e para o *ControlNet*.

Esse processo se repete 16 vezes, gerando 16 quadros. O *denoise* é lentamente elevado ao longo do processo, com ele tendo o valor de 0.3 ao final dele. Esse limite de 16 quadros é arbitrário - o último quadro pode ser inserido no início do processo manualmente, de modo a gerar mais 16 quadros.

Figura 29 - Sequência produzida pela técnica²²



fonte: Própria (2023)

3.6 - Animação em Motion

Esse processo é muito mais próximo do motion design tradicional, e envolve os seguintes processos, nessa ordem:

²² Imagens produzidas pelo método descrito (seleção de 8 entre os 16 quadros gerados). No canto superior esquerdo, a imagem inicial. No canto inferior direito, a imagem final. O prompt positivo usado foi "ugly monster, great detail, disease", e o negativo foi "dog, fake, plastic, woman" - inserir o nome do objeto em foco da imagem inicial no prompt negativo ajuda na transformação.

1. O elemento a ser animado por *motion design* é desenhado, com cada parte que vai se mover sendo um desenho separado (um processo similar à fazer um boneco de papel, no sentido de que cada parte deve ser feita separadamente e posteriormente unida em um único boneco articulado)

Figura 30 - Desenho preparado, com elementos separados²³



fonte: Própria (2023)

2. Essa imagem, então, deve ser levada ao *Stable Diffusion* no modo *txt2img* (onde o único *input* do módulo central, além das configurações, são *prompts* de texto) e inserida no módulo do **ControlNet**, com o preprocessador (um processo automático que deve ser realizado nos *inputs* do *ControlNet*, de acordo com o modo no qual ele vai ser utilizado) no modo “*lineart*”.

- 2.1. Esse passo pode ser feito também no modo *img2img*, com o *input* de imagem do modo sendo blocos de cores preenchidos a partir da *lineart* (mas sem a *lineart* em si), com o parâmetro *Denoise Strength* (o parâmetro determina qual o grau de mudança entre a imagem inicial e a imagem final - um valor de 0 gera uma imagem igual ao input, e o valor de 1 gera uma imagem completamente diferente, sem base nenhuma na imagem inicial) estando **entre 0,5 e 0,8**. Este método

²³ A cabeça do personagem, nesse caso, foi animada anteriormente pelo Processo Base. Inserir o primeiro quadro da animação gerada na imagem e a isolar com uma máscara na hora da geração para que o algoritmo pudesse “ver” o resultado já obtido e fazer a sua geração de forma mais alinhada, sem fugir muito do estilo e das cores. A máscara impede que a parte mascarada seja alterada na hora da geração.

alternativo pode ser útil quando o preenchimento automático de cores do *txt2img* se confundir sobre quais partes preencher, ou de que forma ele deveria preencher essas partes.

3. O *Stable Diffusion* pode utilizar diferentes modelos de difusão, cada um treinado de formas diferentes com datasets de imagens diferentes, o que acaba gerando estilos e aptidões diferentes para cada um. Os dois que eu utilizarei ao longo desse projeto serão o **Reliberate v2** e o **Stable Diffusion 1.5**, que por meio de testes pude determinar que tinham os estilos mais interessantes para esse projeto - os dois tem uma boa capacidade de gerar detalhes e formas coerentes com o uso simultâneo dos módulos previamente mencionados.
4. O módulo *LoRA* de nome “**add_detail**” deve ser ativado, com peso entre 0,6 e 1. Esse módulo, como o nome indica, permite um controle mais direto sobre a quantidade de detalhes gerada na imagem. Se a cena exigir um estilo mais grotesco, partes da imagem ou a imagem inteira podem ser geradas com a adição de um segundo *LoRA*, de nome “**flashlight_v15**”, treinado com fotos tiradas com flash da câmera ativados, geralmente em ambientes escuros, com o peso entre 0,5 e 0,7. Por algum motivo, a combinação desses módulos *LoRA* tendem a gerar detalhes mais grotescos, com relevos mais salientes devido à característica mais reflexiva da superfície dos detalhes.
5. O módulo principal do *Stable Diffusion* deve então ter seus vários parâmetros configurados:
 - 5.1. Seleciona-se o *sampler*: o método pelo qual a imagem será gerada a partir de ruído. Existem dezenas de tipos de *samplers*, que se dividem em linhas gerais em 2 tipos: os *estocásticos* (que tendem a variar bastante entre os *steps*) e os *convergentes* (que tendem a convergir em uma imagem específica na medida em que os *steps* passam). Através de análises externas e múltiplos testes, **os samplers “Euler” e “DPM++ 2M Karras” foram escolhidos**, por serem convergentes e obterem uma qualidade boa no resultado final.

- 5.2. Seleciona-se o número de *steps* - o número de rodadas de remoção de ruído aplicados pelo *sampler* selecionado. É importante achar um equilíbrio - uma quantidade de *steps* que gere resultados bons, mas que não acarrete em um tempo de geração muito grande. Por meio de testes, foi determinado que **uma quantidade de *steps* entre 30 e 40** é o ideal para os *samplers* selecionados.
- 5.3. O parâmetro CFG (*Classifier-Free Guidance*, ou Orientação Livre de Classificadores) determina o quão “fiel” ao *prompt* de texto a imagem final será. Através de testes, **o valor mais adequado para o CFG para o uso em questão foi o de 8**.
- 5.4. A *seed* é o valor numérico através da qual os ruídos necessários para a geração são gerados. Para experimentação, o *seed* deve ser deixado em modo randômico, mudando assim após cada geração de imagem. Caso um estilo pareça promissor, o botão de repetição da *seed* deve ser acionado, e os outros parâmetros devem ser ajustados e testados até que a imagem gerada esteja aceitável.
- 5.5. Por fim, e mais importante, deve se selecionar os *prompts* de texto. Os *prompts* servem para se descrever o conteúdo, a qualidade, o estilo e a forma da imagem. No *Stable Diffusion*, existem 2 tipos de *prompt*: o positivo e o negativo. O positivo deve tentar descrever o conteúdo, forma, estilo e qualidade da imagem que se quer gerar, e o negativo descreve o que deve ser evitado. Normalmente, o *prompt* negativo possui palavras como “*bad quality, blurry, dull*” (qualidade ruim, borrado, sem graça), entre outras coisas, mas pode ser usado também caso algum elemento indesejado continue aparecendo nas gerações sem ter sido descrito no *prompt* positivo.
6. Com os parâmetros selecionados, deve-se então gerar imagens com a *seed* em modo aleatório, até que se ache algo promissor. Então, a *seed* deve ser

travada, e o restante dos parâmetros deve ser testado até que se gere um resultado satisfatório.

7. Com a imagem formada, ela deve ser levada ao modo *img2img*:

7.1. Uma vantagem desse método é que a imagem gerada pelo *img2img* pode ser colocada como input, e assim ela pode ir acumulando detalhes e se aproximando mais dos *prompts* e dos parâmetros indicados pelos *LoRAs* utilizados ao longo das recursões.

7.2. A através de testes, foi possível determinar que **2 a 3 recursões**, ou seja, usar a imagem recém gerada como base para uma nova geração, tendem a gerar melhores resultados, que não fogem da composição original.

7.3. Através de testes, foi determinado que o valor do parâmetro *Denoise Strenghth* deve ser colocado **entre 0,3 e 0,4**, pois valores maiores fazem a imagem fugir demais da composição original, ainda mais se o módulo *ControlNet* não estiver ativado (nesse caso ele estará, pois a imagem de *input* foi importada diretamente do modo *txt2img* através do botão anteriormente mencionado).

8. Com todas as partes formadas, elas devem então ser unidas em um arquivo *.psd* do *Photoshop*, com cada uma das partes em uma camada, nos lugares corretos e com as camadas na ordem correta.

Figura 31 - Arquivo preparado, pronto para ser animado



fonte: Própria (2023)

9. O arquivo .psd seria então importado para o *After Effects*, aonde elas seriam animadas com técnicas de *Motion Design*.

4 - Conclusão

Esse projeto buscou desenvolver diversas técnicas e modos de se utilizar os novos modelos de geração de imagens disponíveis. Ao passo que consegui obter sucesso nessa tarefa, o ritmo do progresso nesse campo se provou extremamente rápido, e ao longo desses 2 semestres, pude ver o surgimento de módulos como o *AnimateDiff* e o *IP-Adapter*, que alteram completamente as possibilidades desses tipos de ferramenta. Esses módulos surgiram a tempo de eu poder aprender a usá-los e incluí-los em meu projeto, mas outros módulos, como o *Stable Video Diffusion*, surgiram tão recentemente (mais precisamente, no dia 21 de novembro de 2023) que não foi possível incluí-los nesse relatório.

Esse projeto se ateve ao *Stable Diffusion* por conta de ser um modelo de código-aberto, com diversos módulos, mas o campo das ferramentas de código fechado também avançou imensamente. O modelo *Pika 1.0* foi anunciado mais recentemente ainda (28 de novembro de 2023), e demonstrou ter capacidades muito superiores aos geradores de vídeo atualmente disponíveis. Todo avanço metodológico do uso dessas ferramentas acaba rapidamente se tornando obsoleto (ainda que possam ser usados, caso o resultado buscado seja mais estilístico e se

atenha menos à verossimilhança). Os avanços vistos, porém, não são apenas lineares, no sentido de serem apenas modelos com as mesmas funções, mas mais fidedignos. O novo modelo anunciado no mesmo dia do Pika 1.0, *AnimateAnyone*, por exemplo, toma como input uma imagem base de uma pessoa ou personagem, e uma sequência animada de movimentos corporais, e aplica o movimento à pessoa - esse novo modelo abre um meio de animação inteiramente novo, e diferente dos já apresentados nesse projeto.

Uma outra limitação desse projeto é a necessidade de ter placas de vídeo muito potentes para que seja possível usar esses algoritmos, o que limita o uso dessas ferramentas a poucas pessoas, o que vai contra a premissa democratizadora da arte que essas ferramentas têm o potencial de trazer.

5- Referências Bibliográficas

ADVADNOUN. [Twitter]. Disponível em
<<https://twitter.com/advadnoun/status/1351038053033406468>>. Acesso em: 25 jun. 2023.

ADVADNOUN. [Twitter]. Disponível em
<<https://twitter.com/advadnoun/status/1367556678896394240>>. Acesso em: 25 jun. 2023.

ADVADNOUN. [Twitter]. Disponível em
<<https://twitter.com/advadnoun/status/1382523014131064838>>. Acesso em: 25 jun. 2023.

AP NEWS. **Artificial intelligence images rights**. Disponível em
<<https://apnews.com/article/artificial-intelligence-images-rights-1c6d9e0e260e2d135a3e3bf98d5493df>>. Acesso em: 25 jun. 2023.

ARCURI, Christiane. **Nuno Ramos – narrativas visuais**. Rio de Janeiro: Ed. Multifoco, 2015 – fruto Tese de Doutorado “A arte de Nuno Ramos: um relevo na estética da contemporaneidade”, UFF, 2014

BROCK, Andrew; DONAHUE, Jeff; SIMONYAN, Karen. **Large Scale GAN Training for High Fidelity Natural Image Synthesis**, 2018. Disponível em: arXiv:1809.11096.

CARTER, Tom. Workers are worried about AI taking their jobs. Artists say it's already happening. In: **Business Insider**. [S. l.], 1 out. 2023. Disponível em:
<<https://www.businessinsider.com/ai-taking-jobs-fears-artists-say-already-happening-2023-10>>. Acesso em: 3 dez. 2023.

CIVITAI. **Detail Tweaker LoRA**. Disponível em
<<https://civitai.com/models/58390/detail-tweaker-lora-lora>>. Acesso em: 25 jun. 2023.

DAVIS, Wes. AI companies have all kinds of arguments against paying for copyrighted content. In: **The Verge**. [S. l.], 4 nov. 2023. Disponível em:
<<https://www.theverge.com/2023/11/4/23946353/generative-ai-copyright-training-data-openai-microsoft-google-meta-stabilityai>>. Acesso em: 3 dez. 2023.

GABRIEL, Ruan de Sousa. Prêmio Jabuti desclassifica edição de 'Frankenstein' ilustrada por inteligência artificial. In: **O Globo**. [S. l.], 10 nov. 2023. Disponível em:

<https://oglobo.globo.com/cultura/livros/materias-em-capitulo/2023/11/premio-jabuti-desclassifica-edicao-de-frankenstein-ilustrada-por-inteligencia-artificial.ghml>. Acesso em: 3 dez. 2023.

GITHUB. **ComfyUI**. Disponível em <<https://github.com/comfyanonymous/ComfyUI>>. Acesso em: 29 nov. 2023.

GITHUB. **Stable Diffusion WebUI**. Disponível em <<https://github.com/AUTOMATIC1111/stable-diffusion-webui>>. Acesso em: 25 jun. 2023.

GRANT, Daniel. Artists, writers, performers and their advocates call on US Congress to ban companies from copyrighting AI-generated art: The AI Day of Action, scheduled for 2 October, comes as US officials consider whether and how to regulate material generated by artificial intelligence. In: **The Art Newspaper**. [S. l.], 27 set. 2023. Disponível em: <<https://www.theartnewspaper.com/2023/09/27/artificial-intelligence-day-of-action-us-congress-copyright-regulation>>. Acesso em: 3 dez. 2023.

GUO, Yuwei et al. **AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning**, 2023. Disponível em: arXiv:2307.04725.

HO, Jonathan; SALIMANS, Tim. **Classifier-Free Diffusion Guidance**, 2022. Disponível em: arXiv:2207.12598v1.

HOME, Steward. **Assalto à cultura: utopia subversão guerrilha na (anti)arte do século XX**. São Paulo: Conrad Editora do Brasil, 2004

HU, Edward J. et al. **LoRA: Low-Rank Adaptation of Large Language Models**, 2021. Disponível em: arXiv:2106.09685.

HU, Li et al. **Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation**. 2023. Disponível em: arXiv: 2311.17117

ISHIGURO, Chiaki; TAKESHI, Okada. **"How Can Inspiration Be Encouraged in Art Learning ?"**, 2017.

LAION. **FAQ**. Disponível em <<https://laion.ai/faq/>>. Acesso em: 25 jun. 2023.

LAION. **LAION-5B**. Disponível em <<https://laion.ai/blog/laion-5b/>>. Acesso em: 25 jun. 2023.

MIDJOURNEY. **Midjourney**. Disponível em <<https://www.midjourney.com/app/>>. Acesso em: 25 jun. 2023.

OPENAI. **CLIP**. Disponível em <<https://openai.com/research/clip>>. Acesso em: 25 jun. 2023.

OPENAI. **DALL-E**. Disponível em <<https://openai.com/research/dall-e>>. Acesso em: 25 jun. 2023.

OPENAI. **DALL-E 2**. Disponível em <<https://openai.com/dall-e-2>>. Acesso em: 25 jun. 2023.

OPENAI. **DALL-E 3**. Disponível em <<https://openai.com/dall-e-3>>. Acesso em: 28 nov. 2023.

OPENAI. **Image GPT**. Disponível em <<https://openai.com/research/image-gpt>>. Acesso em: 25 jun. 2023.

OPENART AI. **The most complete guide to stable diffusion parameters**. Disponível em <<https://blog.openart.ai/2023/02/13/the-most-complete-guide-to-stable-diffusion-parameters/#:~:text=Stable%20Diffusion%20creates%20an%20image,but%20to%20a%20certain%20degree>>. Acesso em: 25 jun. 2023.

PIKA. [Blog]. Disponível em <<https://pika.art/blog>>. Acesso em: 03 dez. 2023.

QUEIROZ, Julia. Jurado do Jabuti diz que não sabia que livro foi ilustrado por IA e não teria indicado se soubesse. In: **Estadão**. [S. l.], 9 nov. 2023. Disponível em: <<https://www.estadao.com.br/cultura/literatura/premio-jabuti-livro-inteligencia-artificial-categoria-melhor-ilustracao-nprec/>>. Acesso em: 3 dez. 2023.

ROMBACH, Robin et al. **High-Resolution Image Synthesis with Latent Diffusion Models**, 2021. Disponível em: arXiv:2112.10752.

SOFTWAREKEEP. **How to use stable diffusion LoRA models**. Disponível em <<https://softwarekeep.com/help-center/how-to-use-stable-diffusion-lora-models/>>. Acesso em: 25 jun. 2023.

STABILITY AI. **Introducing Stable Video Diffusion**. Disponível em <<https://stability.ai/news/stable-video-diffusion-open-ai-video-model>>. Acesso em: 03 dez. 2023.

STABLE DIFFUSION ART. **Samplers**. Disponível em <<https://stable-diffusion-art.com/samplers/>>. Acesso em: 25 jun. 2023.

TERRA. **Foto do Papa Francisco com casaco estiloso foi gerada por IA**. Disponível em <<https://www.terra.com.br/byte/foto-do-papa-francisco-com-casaco-estiloso-foi-gerada-por-ia,f868b9bd97137b000930b7780b19c14fr2ujear6.html>>. Acesso em: 25 jun. 2023.

THE DECODER. **New CLIP model aims to make stable diffusion even better**. Disponível em <<https://the-decoder.com/new-clip-model-aims-to-make-stable-diffusion-even-better/>>. Acesso em: 25 jun. 2023.

TOWARDS DATA SCIENCE. **Stable diffusion using Hugging Face**. Disponível em <<https://towardsdatascience.com/stable-diffusion-using-hugging-face-501d8dbdd8#:~:text=Stable%20diffusion%20only%20uses%20a,are%20similar%20in%20latent%20space>>. Acesso em: 25 jun. 2023.

WILLIAMS, Richard. **Manual de animação: Manual de métodos, princípios e fórmulas para animadores clássicos, de computador, de jogos, de Stop motion e de internet**. 1ª Edição. São Paulo: Editora Senac, 2016.

YE, Hu et al. **IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models**. Disponível em: arXiv:2308.06721.

ZHANG, Lvmin; AGRAWALA, Maneesh. **Adding Conditional Control to Text-to-Image Diffusion Models**, 2022. Disponível em: arXiv:2302.05543v1.