

# Do Rascunho ao Filme

Transformando Textos em Histórias Visuais  
com Inteligência Artificial

Arthur Jung Barreto



UNIVERSIDADE FEDERAL DE GOIÁS (UFG)  
INSTITUTO DE INFORMÁTICA (INF)

ARTHUR JUNG BARRETO

## **Do Rascunho ao Filme**

Transformando Textos em Histórias Visuais com Inteligência Artificial

Goiânia  
2025



UNIVERSIDADE FEDERAL DE GOIÁS  
INSTITUTO DE INFORMÁTICA

## TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO NO REPOSITÓRIO INSTITUCIONAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio do Repositório Institucional (RI/UFG), regulamentado pela Resolução CEPEC no 1240/2014, sem ressarcimento dos direitos autorais, de acordo com a Lei no 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo dos Trabalhos de Conclusão dos Cursos de Graduação disponibilizado no RI/UFG é de responsabilidade exclusiva dos autores. Ao encaminhar(em) o produto final, o(s) autor(a)(es)(as) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

### 1. Identificação do Trabalho de Conclusão de Curso de Graduação (TCCG)

Nome(s) completo(s) do(a)(s) autor(a)(es)(as): ARTHUR JUNG BARRETO

Título do trabalho: Do Rascunho ao Filme

Transformando Textos em Histórias Visuais com Inteligência Artificial

### 2. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador) Concorda com a liberação total do documento SIM NÃO<sup>1</sup>

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante: a) consulta ao(à)(s) autor(a)(es)(as) e ao(à) orientador(a); b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo do TCCG. O documento não será disponibilizado durante o período de embargo.

#### Casos de embargo:

- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro.

**Obs.: Este termo deve ser assinado no SEI pelo orientador e pelo autor.**



Documento assinado eletronicamente por **Arthur Jung Barreto, Discente**, em 10/01/2025, às 20:00, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Fernando Marques Federson, Professor do Magistério Superior**, em 15/01/2025, às 16:09, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).

---



A autenticidade deste documento pode ser conferida no site [https://sei.ufg.br/sei/controlador\\_externo.php?acao=documento\\_conferir&id\\_orgao\\_acesso\\_externo=0](https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0), informando o código verificador **5089496** e o código CRC **3C8E2710**.

---

Referência: Processo nº 23070.001545/2025-71

SEI nº 5089496

ARTHUR JUNG BARRETO

## **Do Rascunho ao Filme**

Transformando Textos em Histórias Visuais com Inteligência Artificial

Relatório final de Trabalho de Conclusão de Curso, apresentado à Universidade Federal de Goiás, como parte das exigências para a obtenção do título de Bacharel em Inteligência Artificial.

Orientador: Prof. Dr. Fernando Marques Federson

Goiânia

2025

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UFG.

BARRETO, ARTHUR JUNG

Do Rascunho ao Filme [manuscrito] : Transformando Textos em Histórias Visuais com Inteligência Artificial / ARTHUR JUNG BARRETO. - 2025.

77 f.

Orientador: Prof. Dr. Fernando Marques Federson.  
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Goiás, Instituto de Informática (INF), Inteligência Artificial, Goiânia, 2025.

1. inteligência artificial. 2. text-para-imagem. 3. texto-para-vídeo. I. Federson, Fernando Marques , orient. II. Título.

CDU 004


ARTHUR JUNG BARRETO

## **Do Rascunho ao Filme**

Transformando Textos em Histórias Visuais com Inteligência Artificial


Relatório final de Trabalho de Conclusão de Curso, apresentado à Universidade Federal de Goiás, como parte das exigências para a obtenção do título de Bacharel em Inteligência Artificial.

Data da Aprovação: 17 de dezembro de 2024.




---

Prof. Dr. Fernando Marques Federson  
Orientador (INF-UFG)



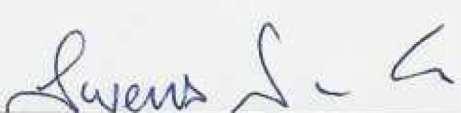
---

Prof. Dr. Aldo André Díaz Salazar  
Coordenador de TCC do BIA (INF-UFG)



---

Prof. Dr. Anderson da Silva Soares  
Coordenador do BIA (INF-UFG)



---

Prof. Dr. Iwens Gervasio Sene Junior  
(INF-UFG)

ARTHUR JUNG BARRETO

## Do Rascunho ao Filme

Transformando Textos em Histórias Visuais com Inteligência Artificial

### RESUMO

Este Relatório de Conclusão de Curso tem como objetivo reunir os resultados da minha jornada para me tornar um especialista em **Geração de Imagens (History-to-Images)**. Uma ilustração e sua narrativa descrevem os períodos de trabalho. Os Apêndices contêm os Termos de Aceite de Entrega e os resultados obtidos durante cada período de trabalho.

Palavras-chave: inteligência artificial, modelos grandes de linguagem, geração automática de datasets.

### ABSTRACT

This Course Completion Report aims to bring together the results of my journey to become an expert in **Image Generation (History-to-Images)**. An illustration and its narrative describe the work periods. The Appendices contain the Delivery Acceptance Terms and the results obtained during each work period.

Keywords: artificial intelligence, large language models, automatic dataset generation.

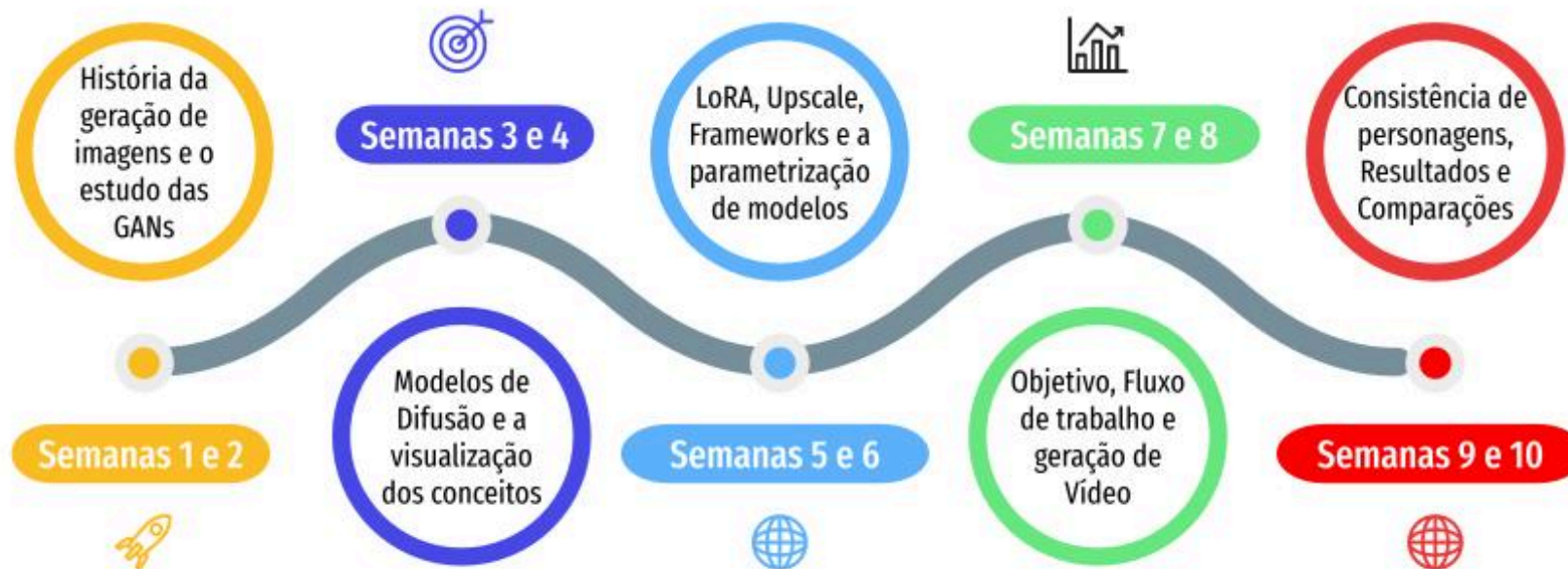
Goiânia

2025

# Minha Jornada

Arthur Jung Barreto

Especialista em: Geração de Imagens  
(History-to-Images)



## MINHA JORNADA

**Nome:** Arthur Jung Barreto

**Especialidade:** Geração de Imagens (History-to-Images)

### Objetivo deste documento

Durante o processo da disciplina Residência em IA<sup>1</sup>, foram gerados diversos resultados na construção da minha especialização. A cada semana, um conjunto de resultados foi formalizado por um Termo de Aceite de Entrega e avaliado por uma banca, considerando o planejado e o realizado para o período. Este documento tem como objetivo descrever esses resultados obtidos, fazendo referência aos Termos de Aceite de Entrega e seus documentos associados.

### Minha Jornada

Minha Jornada começou na **Semana 1** já com a área de conhecimento da minha especialização definida. Estive com dúvidas do que escolher anteriormente, tive o interesse na área de LLM, onde meu objetivo era criar prompts para injetar em modelos de geração de imagem, mas a partir da chamada dos Congressos, em especial do Congresso IPCV'24 - The 28th Int'l Conf on Image Processing, Computer Vision, & Pattern Recognition, pude perceber qual a área que eu seguiria. A formação obtida nas disciplinas Visão Computacional e Redes Neurais Profundas do Bacharelado também foram importantes para a minha escolha. Com a leitura de alguns artigos e blogs, e principalmente com a utilização de alguns modelos na prática, pude decidir, que gostaria de me aprofundar na área de **Geração de Imagens** propriamente. Portanto, na **Semana 1** me aprofundei em explorar a história da área de conhecimento, fazendo um breve levantamento de quais seriam os fundamentos e técnicas atuais por trás dessa área. Na **Semana 2** busquei entender mais sobre uma arquitetura que mudou o rumo da geração de imagens, as GANs (Generative Adversarial Networks), estudando diretamente o artigo principal, além de explorar as

---

<sup>1</sup> Dez semanas, entre setembro de 2024 e dezembro de 2024.

diferentes variações que foram sendo desenvolvidas posteriormente. As referências lidas, assim como algumas observações que considero importantes, podem ser obtidas em detalhes no material disponibilizado no **Apêndice 1**.

Na **Semanas 3**, fui estudar o que tem de mais recente em termos de arquitetura na área, modelos de difusão latente, e também o CLIP (Contrastive Language Image Pre-Training), sempre mantendo os procedimentos para uma revisão bibliográfica simples e a leitura de artigos relacionados ao tema. A **Semana 4** foi diferente em alguns aspectos, onde eu realizei também a exploração de modelos que foram lançados em cima da arquitetura de difusão latente, mas também tinha o objetivo de criar uma visualização que mostrasse em algum ponto a evolução de arquitetura e de modelos ao longo do tempo assim como demonstrar quais são alguns dos principais fundamentos necessários para entender como esses modelos de geração de imagens funcionam. No **Apêndice 2**, é possível encontrar a relação completa dos artigos lidos com algumas observações. Este material serviu de base para a continuidade dos trabalhos realizados, em especial, na organização do conhecimento e direcionamento da jornada.

Durante o período das **Semana 5 e 6**, foi realizado um estudo sobre os frameworks que possibilitaram o uso desses modelos de uma maneira prática satisfatória. Entre os frameworks encontrados vale citar um em especial o Comfy UI que foi base para a geração de todas as imagens durante o processo, nestas semanas tive também como objetivo buscar fontes confiáveis que disponibilizam esses modelos e definir os requisitos de hardware e o ambiente de teste onde eu estaria utilizando esses modelos, foi nesse período onde eu adquiri uma nova placa gráfica que possibilitou testes mais rápidos ou com modelos mais pesados, na **Semana 6** em especial comecei a realizar os primeiros testes, entendendo e documentado a diferença de parametrização dos modelos, nessa semana eu abordei outros dois tópicos relevantes na geração de imagens, o LoRA (Low-Rank Adaptation) para adaptar o modelo a um novo conjunto de imagens e upscale que permite um aumento na resolução da imagem sem uma grande perda de qualidade. A documentação e os testes realizados podem ser encontrados no **Apêndice 3**.

Minhas principais atividades nas **Semanas 7 e 8** foram em como realizar uma tarefa definida por mim como a criação de imagens que representassem um determinado texto narrativo (uma história), portanto ataquei nesse período uma área completamente diferente do que já havia estudado para montar todo um fluxo automático, desde a geração da história até a finalização dela em um formato de vídeo, sim nessas semanas eu também abordei a geração de vídeo através do Animate Diff. Portanto, nessas semanas estive envolvido com LLMs (Large Language Models) na geração de história, onde precisei me aprofundar o suficiente nas áreas de engenharia de prompt e no aprimoramento de prompt. Os testes realizados e as descobertas dessas duas semanas estão disponíveis no **Apêndice 4**.

As últimas **Semanas, 9 e 10**, foram essenciais para a conclusão de todo o processo, estive envolvido com um problema ainda não muito resolvido na área, a consistência na geração de personagens, percebi porém que o que há atualmente para resolver esse problema é a utilização de um LoRA treinado exclusivamente em cima de um personagem. Para concluir o trabalho precisei automatizar boa parte da geração através da API do Comfy UI e por fim escolher um estilo de vídeo que eu faria, a edição e geração de voz para narrar o vídeo foi feita com outra ferramenta. Além disso, tiveram muitos por menores nessas duas semanas que podem ser visualizados no **Apêndice 5**.

Em função de tudo que vivi nesta Jornada, gostaria de deixar registrado que eu aprendi bastante com o processo, o suficiente para me considerar um especialista na área, acredito que isso estará demonstrado nesse documento e de maneira prática no resultado final gerado, lembrando que isso não significa saber absolutamente tudo sobre geração de imagens, até porque a cada dia um novo modelo é lançado, mas sim saber fluir e aprender rapidamente a se posicionar diante dos desafios.

## APÊNDICE 1

### Termo de Aceite de Entrega

#### Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

**Data da Reunião (“gate”) de aprovação:** 18 de set. de 2024

**Participantes da Entrega** [matriculados em Residência em IA]:

Arthur Jung Barreto

**Entrega:** [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

- **Definição do tópico da especialização :**
  - LLMs
  - Image Generation (Geração de Imagens)
  - AI Art
  - Text-to-Image
- **História da geração de imagens :**
  - xxxx-2000
  - 2000-2014
  - 2014-2020
  - 2020-2024
  - Fundamentos e Técnicas:
    - GANs (Generative Adversarial Networks)
    - Diffusion Models
    - CLIP (Contrastive Language Image Pre-Training)

**Planejamento:** [descrever o que pretende fazer para realizar a próxima ENTREGA]

- Para a próxima entrega acredito que seja relevante entender as principais técnicas atualmente utilizadas:
  - GANs: Funcionamento, tipos, vantagens e desvantagens.
  - Diffusion Models: Comparação com GANs.
  - CLIP: Funcionamento, integração, alternativas.
- Paralelamente, entender como funciona a avaliação de modelos de imagem e entender qual a relação do tópico com a sociedade:
  - Avaliação de modelos: Métricas e desafios.
  - Ética e viés: Considerações importantes na geração de imagens.

Observação: [caso precise fazer alguma observação, de qualquer “natureza”]

---

## ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go! ▾

## Definição do tópico da especialização

Está claro que a definição do tópico reflete uma decisão pessoal mas também deve se enquadrar dentro daquilo que nós chamamos de grande área da inteligência artificial, em um primeiro momento eu estava decidido em abordar algum tópico dentro de LLMs (Large Language Models) principalmente algum tópico que fosse dedicado ao uso desses modelos na ajuda à geração de imagens, porém decidi por escolher geração de imagens como tópico, o termo que escolhi para definir o tópico então é **geração de imagens**, pesquisando em congressos e conferências sobre visão computacional (onde se enquadra esse tópico) vi que o termo que eles utilizam é realmente esse, outros termos aparecem na internet como AI Art, ele reflete um pouco da coisa mas não dá para enxergá-lo como um tópico de especialização.

O que me encanta e me motiva a escolher esse tópico é o campo de aplicação, usar as ferramentas para criar imagens, uma vez que é possível criar imagens para histórias, filmes, jogos, ainda estamos razoavelmente distantes disso, mas o que antes poderia ser só imaginação agora tem a chance de se “materializar”, mesmo que não fique 100% fiel ao que se tinha em mente. Eu sou péssimo para desenhar, mas sei descrever uma imagem para ser criada por um artista, e nesse caso o artista é o computador.

## História da geração de imagens

A geração de imagens, a capacidade de criar visuais a partir de dados ou comandos, é uma jornada longa. Iniciando com a câmera obscura, que projetava imagens invertidas, à fotografia que revolucionou a captura da realidade. A era digital trouxe softwares de edição e gráficos para computador, permitindo manipular e criar imagens complexas junto com a fotografia digital.

Nos últimos anos, a inteligência artificial tem sido a força motriz por trás de avanços sem precedentes na geração de imagens. Redes neurais artificiais e aprendizado profundo permitiram que máquinas aprendessem a gerar imagens a partir de grandes quantidades de

dados. Entre as principais técnicas e ferramentas que surgiram nessa década e na década passada estão:

- GANs (Redes Generativas Adversariais): Essa técnica revolucionária permite que duas redes neurais compitam entre si, gerando imagens cada vez mais realistas e enganando observadores humanos.
- Modelos de Difusão: Outra técnica poderosa que envolve a geração de ruído aleatório e a sua gradual transformação em uma imagem coerente.

Além dessas técnicas, o modelo CLIP (Contrastive Language-Image Pre-Training), desenvolvido pela OpenAI, desempenha um papel crucial na geração e interpretação de imagens ao vincular texto e imagem de forma eficiente. Ele permite que a IA compreenda melhor descrições textuais. Isso potencializa o processo de criação de visuais alinhados a instruções textuais.

Fontes e Materiais de Referência (Links):

<https://www.american-cse.org/csce2024/conferences-IPCV>

[https://en.wikipedia.org/wiki/Artificial\\_intelligence\\_art](https://en.wikipedia.org/wiki/Artificial_intelligence_art)

[AI Timeline - A history of text-to-image generative models — Fabian Mosele](#)

[Creating Reality: A Comprehensive History of Text-to-Image and Generative Imaging | by Sudhanva MG | Medium](#)

[A brief history of AI-powered image generation](#)

 [A History of AI Art](#) e  [The History of AI Art](#)

## Termo de Aceite de Entrega

### Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

**Data da Reunião (“gate”) de aprovação:** 26 de set. de 2024

**Participantes da Entrega** [matriculados em Residência em IA]:

Arthur Jung Barreto

**Entrega:** [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

- Entender as principais técnicas atualmente utilizadas:
  - GANs (Generative Adversarial Networks):
    - ☰ Estudo das GANs (Generative Adversarial Networks)
      - Funcionamento
      - Tipos (GANs com condição)
        - image-to-image, text-to-image, text-to-video
        - Pix2Pix, StackGAN, AttnGAN, BIGGAN
        - StyleGAN
        - BIG SLEEP - BIGGAN + CLIP
        - DVD-GAN
      - Vantagens e desvantagens.
      - Avaliação de modelos.
    - Modelos de Difusão
    - CLIP (Contrastive Language Image Pre-training)
  - Entender qual a relação do tópico com a sociedade:
    - Ética e viés: Considerações importantes na geração de imagens.

**Planejamento:** [descrever o que pretende fazer para realizar a próxima ENTREGA]

- Modelos de Difusão
- CLIP (Contrastive Language Image Pre-training)
- VAE - Variational AutoEncoder
- Ética e viés: Considerações importantes na geração de imagens. (Adiar entrega para os gates finais)

**Observação:** [caso precise fazer alguma observação, de qualquer “natureza”]

- Não consegui avançar sobre todos os tópicos do planejamento feito no último gate, uma vez que o

---

primeiro tópico se mostrou ser muito mais denso do que o imaginado.

---

## ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: [Go! ▾](#)

## Estudo das GANs (Generative Adversarial Networks).

Vantagens e desvantagens (O artigo [Generative Adversarial Nets](#) tem um capítulo exclusivo sobre isso):

The disadvantages are primarily that there is no explicit representation of  $p_g(x)$ , and that the discriminator (D) must be synchronized well with the generator (G) during training. In particular, G must not be trained too much without updating D to avoid the “Helvetica scenario,” in which G collapses too many values of  $(z)$  to the same value of  $(x)$ , reducing diversity and failing to model  $(p_{\text{data}})$  effectively. This is similar to how the negative chains of a Boltzmann machine must be kept up to date between learning steps.

On the other hand, the advantages are that Markov chains are never needed, gradients are obtained using only backpropagation, no inference is required during learning, and the model can incorporate a wide variety of functions.

Adversarial models may also gain some statistical benefits from the fact that the generator network is not updated directly with data examples but only with gradients flowing through the discriminator. This prevents components of the input from being copied directly into the generator’s parameters. Another advantage of adversarial networks is their ability to represent very sharp, even degenerate distributions, while methods based on Markov chains require the distribution to be somewhat blurry in order for the chains to mix between modes.

- cGAN - condition-to-image (condition (e.g., a label or class information)) [Conditional Generative Adversarial Nets](#)
- Pix2Pix - image-to-image - [Image-to-Image Translation w/ cGAN](#)
- StackGAN - text-to-image (2 stages) - [Stacked Generative Adversarial Networks](#)
- AttnGAN - text-to-image - [Attentional Generative Adversarial Networks](#)
- BIGGAN - [Large Scale GAN](#)
- StyleGAN - [Style-Based Generator Architecture for Generative Adversarial Networks](#)
- BIG SLEEP - CLIP + BIGGAN
- DVD-GAN - [Adversarial Video Generation](#)

## Avaliação das GANs

A avaliação de uma GAN (Generative Adversarial Network) é um desafio, pois não há uma métrica universalmente aceita para medir a qualidade das imagens geradas. Existem algumas abordagens para isso.

- Parzen window-based log-likelihood estimates: A avaliação pode ser feita utilizando janelas de Parzen, que são utilizadas para estimar a verossimilhança (log-likelihood) das amostras geradas pela GAN em relação ao conjunto de dados de teste.
- Erro padrão da média (standard error of the mean): A métrica é acompanhada pelo erro padrão da média, que é calculado ao longo de todos os exemplos testados, fornecendo uma medida da variabilidade da estimativa de verossimilhança.
- TFD (Toronto Face Dataset): No caso do TFD, o erro padrão é calculado de maneira diferente. Em vez de ser feito sobre exemplos individuais, ele é calculado entre diferentes divisões do dataset (folds) utilizadas na validação cruzada. Um parâmetro importante aqui é o  $\sigma$  (desvio padrão), que é ajustado em cada fold do conjunto de validação.

## Fontes e Materiais de Referência (Links):

- [▶ Como as GANs funcionam?](#) [▶ Introdução à GANs Condicionais](#)

## APÊNDICE 2

## Termo de Aceite de Entrega

### Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

**Data da Reunião (“gate”) de aprovação:** 2 de out. de 2024

**Participantes da Entrega** [matriculados em Residência em IA]:

Arthur Jung Barreto

**Entrega:** [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

- Entendendo o CLIP e modelos de difusão :
  - Diffusion Models:
    - [Denoising Diffusion Probabilistic Models](#)
    - [Diffusion Models Beat GANs on Image Synthesis](#)
    - [Latent Diffusion Models](#)
  - CLIP (Contrastive Language Image Pre-training)
    - [Learning Transferable Visual Models From Natural Language Supervision](#)
    - [Image Generation with CLIP Latents](#)

**Planejamento:** [descrever o que pretende fazer para realizar a próxima ENTREGA]

- Estruturar quais os principais conceitos e fundamentos necessários para um especialista em geração de imagens possuir.
- Estruturar um timeline das ferramentas (GANs, Diffusion Models, CLIP).
- Explorar os diferentes modelos de Difusão e suas diferenças.

**Observação:** [caso precise fazer alguma observação, de qualquer “natureza”]

- Compreensão em alto nível e em baixo nível.
- Existem muitas outras frentes que podem ser exploradas (open-clip, Classifier-Free Diffusion Guidance, GLIDE, VAE, text-to-video, inpainting, super resolution)

## ACEITE DA ENTREGA:

**CEDRIC LUIZ DE CARVALHO:** Go! ▾

## Entendendo o CLIP e modelos de difusão

CLIP - Contrastive Language Image Pre-training [Learning Transferable Visual Models From Natural Language Supervision](#):

O CLIP (Contrastive Language-Image Pre-training) é um modelo desenvolvido para aprender representações visuais a partir de descrições em linguagem natural, permitindo que ele generalize de forma mais eficaz para diversas tarefas visuais. Destaca-se por seu treinamento baseado em uma grande variedade de pares de imagens e textos da internet, sem depender de rótulos específicos de categorias (como em classificadores tradicionais).

O CLIP é treinado para associar imagens e descrições correspondentes por meio de uma técnica contrastiva. O modelo tenta maximizar a similaridade entre uma imagem e sua descrição correta, ao mesmo tempo que minimiza a similaridade com descrições incorretas. Isso cria representações que podem ser usadas em diversas tarefas visuais (como reconhecimento de objetos, legendas de imagens e até zero-shot classification) sem a necessidade de treino adicional específico para cada tarefa.

Nesse artigo [Hierarchical Text-Conditional Image Generation with CLIP Latents](#), o CLIP é usado na geração de imagens como uma ferramenta para guiar modelos gerativos, como redes adversariais generativas (GANs) ou modelos de difusão, na criação de imagens condicionadas por texto. A principal ideia é utilizar os embeddings do CLIP, que capturam representações ricas de imagens e textos, para alinhar a geração visual com descrições textuais de maneira mais precisa.

O processo envolve gerar uma imagem a partir de um texto usando os latents (representações internas) do CLIP. O modelo gerativo tenta criar uma imagem que corresponda à representação latente do texto, conforme aprendido pelo CLIP. Isso ocorre de forma hierárquica, onde partes da imagem são refinadas em diferentes níveis de detalhe para melhor atender à semântica do texto. O CLIP fornece um "feedback" durante a geração para garantir que a imagem se aproxime cada vez mais da descrição fornecida, permitindo uma criação visual controlada por linguagem natural.

---

Diffusion Models [Denoising Diffusion Probabilistic Models](#):

No artigo, os modelos de difusão são apresentados como uma abordagem para a geração de imagens, em que o processo de criação ocorre através de uma série de passos de refinamento de ruído para imagem. A ideia central é baseada em dois processos: difusão e desdifusão.

- Difusão: Um processo progressivo adiciona ruído gaussiano a uma imagem real, degradando-a até que ela se torne puro ruído. Esse processo é modelado de forma probabilística, garantindo que o ruído seja adicionado de maneira controlada em várias etapas.
- Desdifusão: A partir do estado de ruído puro, o modelo aprende a reverter esse processo, removendo gradualmente o ruído e reconstruindo uma imagem coerente. Esse processo de denoising é feito de forma probabilística em cada passo, até que uma imagem final seja gerada.

O modelo de difusão gera imagens "desfazendo" o processo de ruído, partindo de uma imagem completamente ruidosa e, em cada passo, aprendendo a reconstruir uma versão mais clara e detalhada da imagem, até chegar em uma imagem final. O treinamento envolve o modelo aprender a reverter com precisão cada um desses pequenos passos de ruído.

Como os modelos de difusão superaram as redes adversárias generativas (GANs) na tarefa de síntese de imagens, [Diffusion Models Beat GANs on Image Synthesis](#). A principal razão para isso é que os modelos de difusão, ao contrário das GANs, seguem um processo mais estável e controlado para gerar imagens, o que reduz problemas comuns das GANs, como mode collapse (quando o modelo gera imagens muito parecidas) e instabilidade no treinamento.

Os modelos de difusão são capazes de gerar imagens de alta qualidade porque usam um processo iterativo de remoção de ruído, em que a imagem é construída gradualmente. Isso permite um controle mais refinado e melhora a diversidade e a precisão

das imagens geradas, sem a necessidade de um treinamento adversarial complicado. Além disso, eles conseguem capturar detalhes mais finos e complexos, superando as GANs na fidelidade visual das imagens.

Os modelos de difusão superam as GANs ao oferecer um processo de geração mais estável, capaz de gerar imagens de alta qualidade com maior diversidade e menos artefatos, sem os desafios de convergência enfrentados pelas GANs.

No artigo [High-Resolution Image Synthesis with Latent Diffusion Models](#), os autores propõem uma melhoria significativa nos modelos de difusão ao introduzir a ideia de modelos de difusão latentes. A principal inovação é realizar o processo de difusão, não no espaço de pixels da imagem original, mas em um espaço latente comprimido. Esse espaço latente é uma representação mais compacta da imagem, obtida por meio de um autoencoder, o que reduz significativamente a complexidade computacional do processo.

Em vez de aplicar a difusão diretamente sobre os pixels da imagem, o modelo primeiro transforma a imagem em uma representação latente de menor dimensão. A difusão ocorre nesse espaço reduzido, o que acelera o processo de geração e permite criar imagens de alta resolução com muito menos custo computacional. Após o processo de denoising no espaço latente, a imagem é reconvertida para o espaço de pixels, gerando o resultado final.



Resumidamente, a melhoria dos modelos de difusão latentes está em mover o processo de difusão para um espaço latente comprimido, permitindo a geração de imagens de alta resolução de maneira mais eficiente e rápida, mantendo a qualidade visual.


Fontes e Materiais de Referência (Links):

<https://openai.com/index/clip/> e <https://viso.ai/deep-learning/clip-machine-learning/>

<https://impulsatek.com/clip-claramente-explicado-que-es-y-como-funciona/>

<https://theaisummer.com/diffusion-models/>

 [Stable Diffusion - What, Why, How?](#)  [Contrastive Language-Image Pre-training \(CLIP\)](#)

 [Introduction to Contrastive Language-Image Pre-training \(CLIP\)](#)

## Termo de Aceite de Entrega

### Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

**Data da Reunião (“gate”) de aprovação:** 10 de out. de 2024

**Participantes da Entrega** [matriculados em Residência em IA]:

Arthur Jung Barreto

**Entrega:** [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

- Estruturar os principais conceitos e fundamentos e um timeline dos principais modelos.
  - [Visualização Image Generation](#)
  - Árvore
  - Raiz
- Explorar os diferentes modelos de Difusão.
  - [Documentação](#)
  - Stable Diffusion (1, 1.5, 2, 3, XL)
  - Mid Journey (v1, v5, v6)
  - Imagen (1, 2, 3)
  - DALL-E (2, 3)
  - FLUX 1
- Comfy UI - Ferramenta que facilita o uso dos modelos de IA para geração de imagens.

**Planejamento:** [descrever o que pretende fazer para realizar a próxima ENTREGA]

- Identificar ferramentas que facilitem o uso dos modelos.
- Realizar testes com as ferramentas e modelos (**Stable Diffusion 1.5/FLUX 1 Schnell**).
- Listar prós e contras de cada uma.
- Listar as principais diferenças entre elas.

**Observação:** [caso precise fazer alguma observação, de qualquer “natureza”]

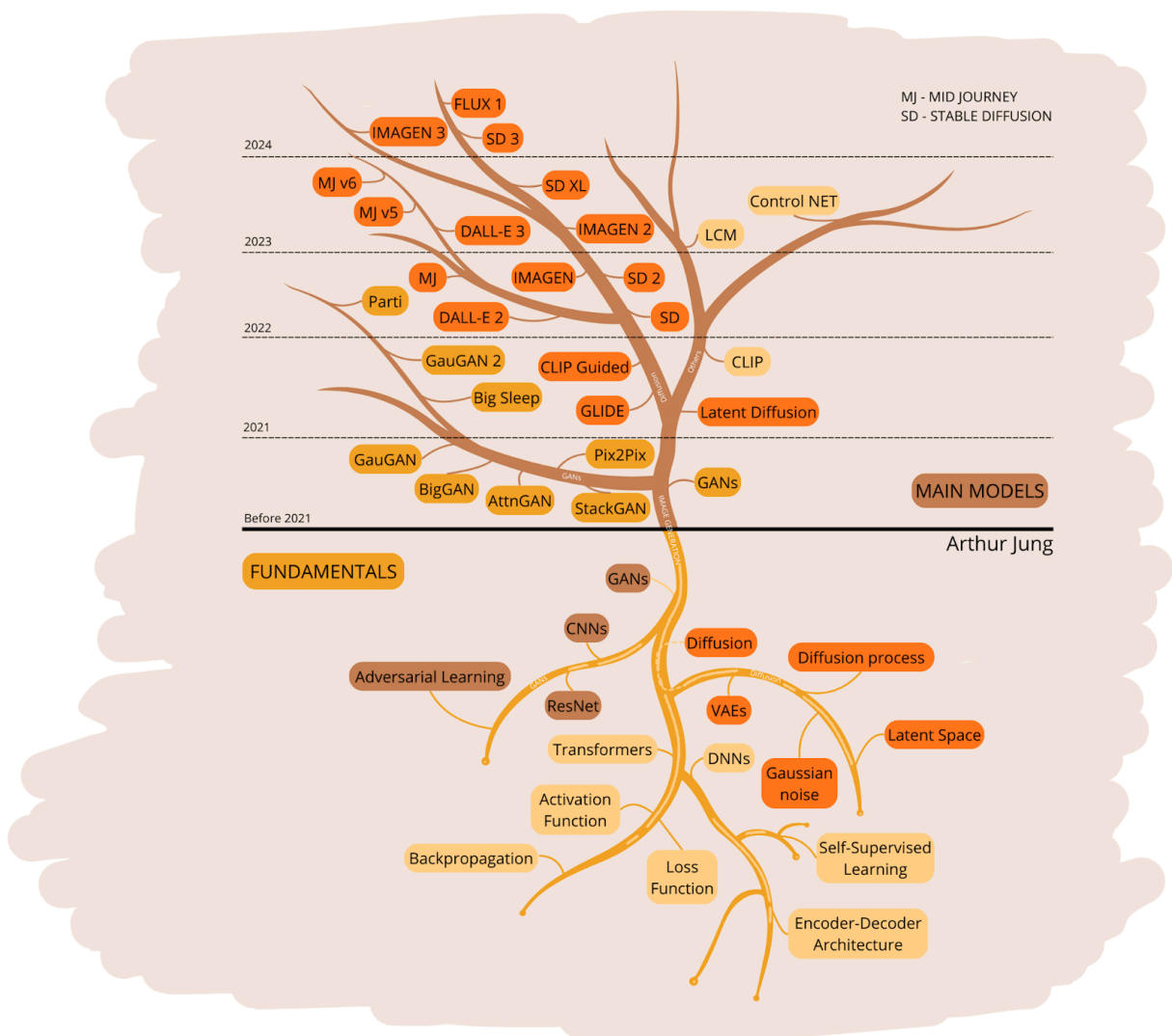
## ACEITE DA ENTREGA:

**CEDRIC LUIZ DE CARVALHO:** Go! ▾

## Estrutura de visualização

Árvore de Modelos - Uma visualização em árvore em ordem cronológica dos modelos lançados para geração de imagens a partir das GANs, utilizei como base a timeline disponibilizada no site: <https://www.fabianmosele.com/ai-timeline>

Raiz de Fundamentos - Uma visualização na forma de raiz para elencar os fundamentos necessários na compreensão das técnicas como Modelos de Difusão e GANs.



Link para a visualização: [Visualização Image Generation](#)

## Modelos Analisados

Principais Modelos de Difusão:

- Stable Diffusion -
  - Stable Diffusion 1 - O primeiro modelo de difusão da Stability AI, focado em geração de imagens de alta qualidade com pouca exigência de recursos computacionais.
    - 1.5 - Existem as versões 1.1 até a 1.5 mas a 1.5 supera as outras em qualidade, a principal diferença entre as versões é o tempo de treinamento, uma vez que a versão 1.1 é treinada em cima do checkpoint da versão 1.0 e a versão 1.2 é treinada em cima do checkpoint da versão 1.1 e as outras versões são treinadas em cima do checkpoint da versão 1.2.
  - Stable Diffusion 2 - Uma versão atualizada da versão 1 com um diferencial de ser treinada com o open-clip ao invés do CLIP da open AI.
  - Stable Diffusion XL - O modelo treinado com imagens 1024x1024 ao invés das outras versões treinadas com imagens 512x512.
  - Stable Diffusion 3 - A nova versão do stable diffusion, mas que não possui licença para uso comercial.
- Mid Journey -
  - Mid Journey - Modelo de geração de imagens com foco em criatividade e estética artística.
  - Mid Journey v5 - Melhoria na qualidade das imagens, com detalhes mais precisos, cores mais vibrantes e maior controle sobre o estilo artístico.
  - Mid Journey v6 - Aprimora a geração de texturas, iluminação e sombreamento, permitindo que as imagens pareçam ainda mais realistas e artísticas
- Imagen -
  - Imagen - Modelo de difusão da Google, altamente treinado em gerar imagens fotorrealistas a partir de descrições textuais detalhadas.

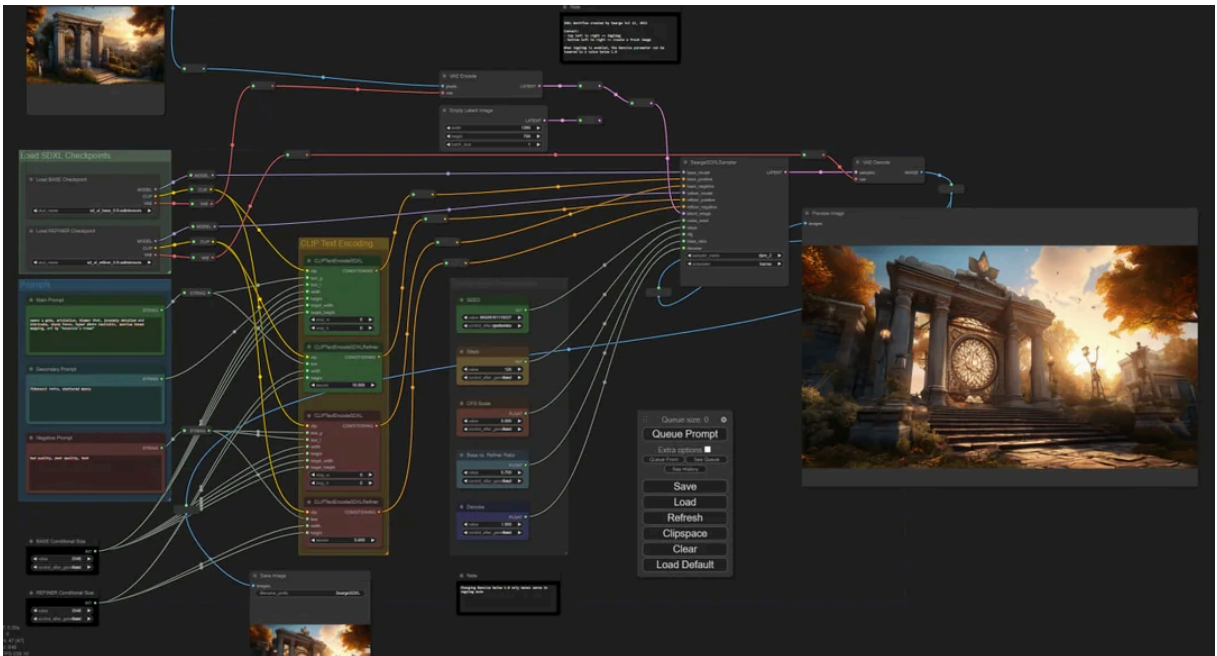
- Imagen 2 - Foco em gerar imagens mais realistas e com maior controle de estilo. Introduce uma arquitetura mais eficiente, permitindo maior diversidade de imagens geradas a partir do mesmo prompt.
- Imagen 3 - Aprimoramento significativo em termos de controle criativo e geração de imagens de alta fidelidade, com foco em contexto semântico mais preciso.
- DALL-E -
  - DALL-E 2 - Versão melhorada, com maior capacidade de gerar imagens de alta qualidade e maior coerência com os prompts textuais, essa versão utiliza difusão.
  - DALL-E 3 - Gera imagens mais coerentes, detalhadas e com precisão de interpretação textual, tornando a interação entre texto e imagem mais natural e sofisticada.
- FLUX 1 - Um dos mais recentes na categoria de difusão, focado na geração de imagens com altíssima precisão de detalhes e controle criativo, buscando superar limitações de modelos anteriores em termos de consistência visual e diversidade. Em versões Pro, Dev e Schnell.

## Frameworks

Ferramentas que facilitam o uso desses modelos e permitem a parametrização e customização como a utilização de LoRAs, upscaling e a aplicação de detalhes após a geração da imagem.

**Comfy UI** - É uma interface gráfica que facilita o uso de modelos de inteligência artificial voltados para a geração de imagens, permitindo aos usuários criar fluxos de trabalho personalizados para manipular e gerar imagens de maneira visual. Em vez de lidar com códigos complexos, os usuários podem conectar módulos em uma interface intuitiva, como em um diagrama de blocos, onde cada módulo representa uma etapa do processo, como ajuste de parâmetros do modelo, manipulação de entrada ou pós-processamento da imagem. Essa abordagem modular torna o processo de geração de imagens mais acessível, personalizável e adaptável, permitindo experimentar diferentes combinações de modelos e

parâmetros de maneira rápida e eficiente. Abaixo uma imagem de um fluxo de trabalho na Comfy UI



Link para a imagem: [Example Workflow Comfy UI](#)

Fontes e Materiais de Referência (Links):

<https://www.fabianmosele.com/ai-timeline>

<https://openai.com/index/dall-e-2/>

<https://www.midjourney.com/home>

<https://arxiv.org/pdf/2205.11487>

<https://imagen.research.google/>

<https://stability.ai/#aboutus>

<https://huggingface.co/CompVis/stable-diffusion>

<https://stability.ai/news/stable-diffusion-public-release>

<https://stability.ai/news/stable-diffusion-v2-release>

<https://docs.midjourney.com/docs/model-version-5>

<https://arxiv.org/pdf/2307.01952>

<https://cdn.openai.com/papers/dall-e-3.pdf>

<https://openai.com/index/dall-e-3/>

<https://deepmind.google/technologies/imagen-2/>

<https://arxiv.org/pdf/2403.03206>

<https://stability.ai/news/stable-diffusion-3>

<https://deepmind.google/technologies/imagen-3/>

<https://blackforestlabs.ai/>

<https://huggingface.co/black-forest-labs/FLUX.1-schnell>

## APÊNDICE 3

## Termo de Aceite de Entrega

### Objetivo deste documento


Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

**Data da Reunião (“gate”) de aprovação:** 16 de out. de 2024

**Participantes da Entrega** [matriculados em Residência em IA]:

Arthur Jung Barreto

**Entrega:** [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

- Documentação:  Frameworks (Image Generation)
- Ferramentas que facilitam o uso dos modelos:
  - **ComfyUI** - Baseada em fluxos de trabalho.
  - **A1111** - Interface Web mais amigável que o ComfyUI.
  - **Forge** - Adaptação do A1111, permite rodar o SDXL mais rápido.
  - **Foocus** - Muito mais simplificada, novos usuários, artistas.
- HuggingFace, SeaArt e CivitAI - Podem ser utilizados para buscar e baixar os modelos.
- Pytorch - Biblioteca de desenvolvimento por trás das ferramentas
- Requisitos para utilizar os modelos mais recentes:
  - “Mínimos”:
    - Placa de vídeo (GPU): 8GB de VRAM (NVIDIA RTX 4060, 4070).
  - Recomendados:
    - Placa de vídeo (GPU): 12GB+ de VRAM (NVIDIA RTX 4080, 4090).
- Realização de testes:
  - Kaggle
    - Permite rodar as ferramentas
    - 30h semanais 2x T4 GPU 16GB VRAM
    - Instável, difícil, complicado, não tem privacidade
  - Local
    - Privacidade
    - Custo
  - Modelos utilizados:
    - RealisticVisionV5.1 Hyper (VAE) Pruned
    - FLUX 1 Schnell Q4\_0

**Planejamento:** [descrever o que pretende fazer para realizar a próxima ENTREGA]

- Entender os parâmetros (Steps, CFG, Sampler, Schedule)
- Otimizar os parâmetros para os modelos selecionados (SD1.5, SDXL, FLUX 1)
- Upscaling
- LoRA

**Observação: [caso precise fazer alguma observação, de qualquer “natureza”]**

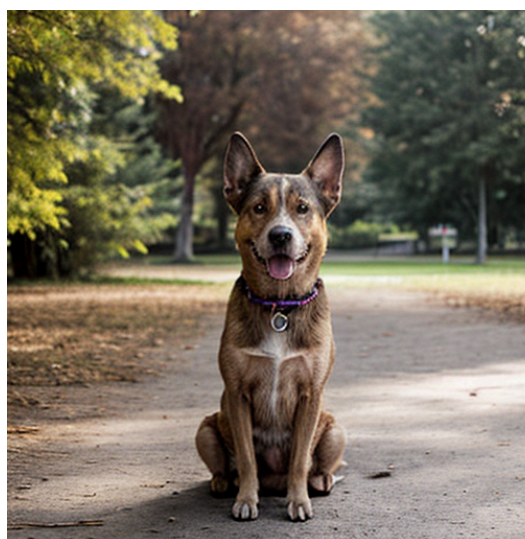
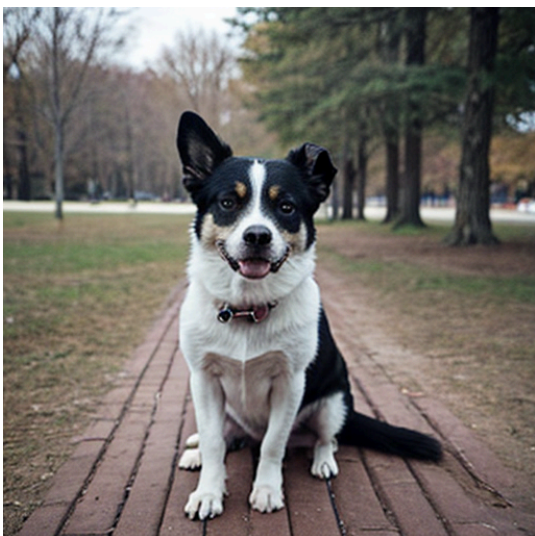
- Aquisição de uma nova placa de vídeo que vai permitir a utilização de modelos localmente

## ACEITE DA ENTREGA:

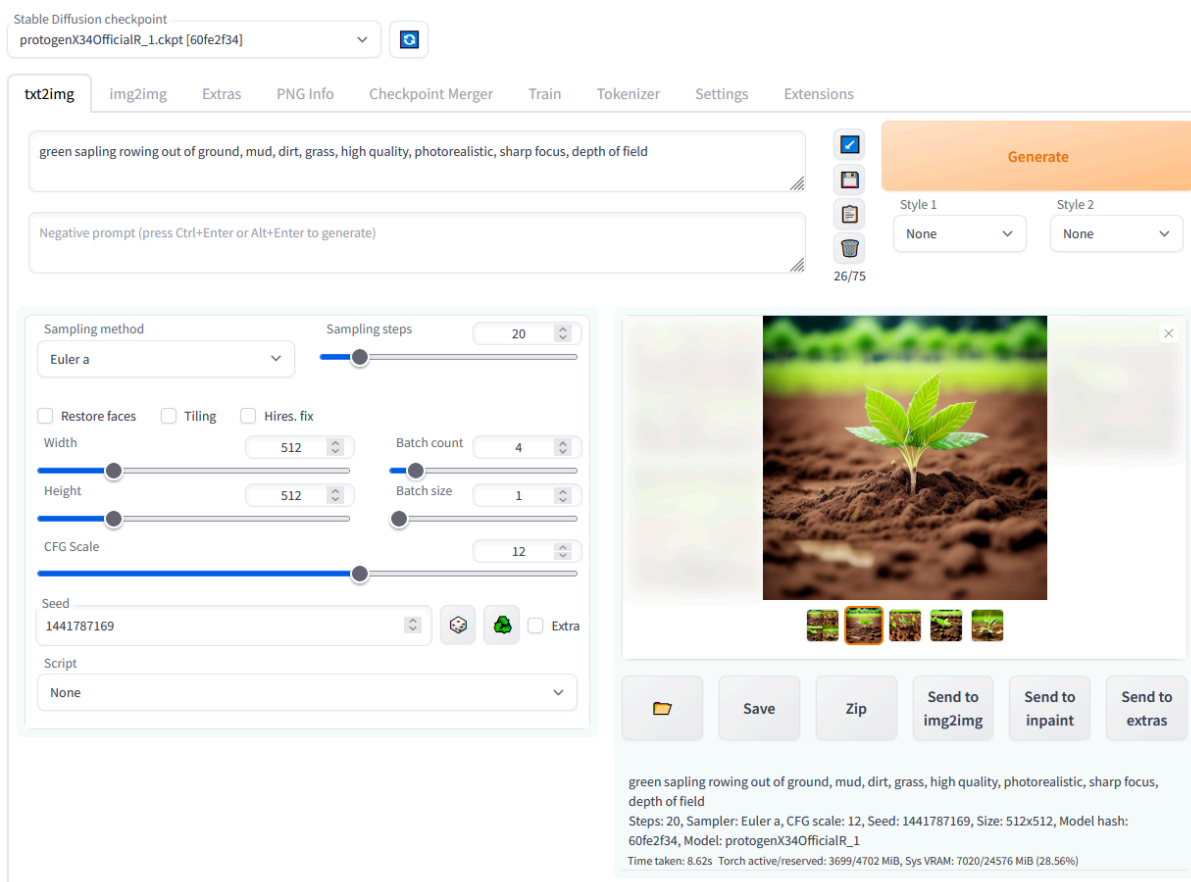
CEDRIC LUIZ DE CARVALHO: [Go! ▾](#)



Algumas imagens geradas abaixo, prompt (“dog in a park”):



**Automatic1111 (A1111)** - É uma interface web amplamente usada para Stable Diffusion, um modelo de geração de imagens por inteligência artificial. O A1111 oferece diversas funcionalidades avançadas, como controle detalhado de parâmetros de geração, suporte para diferentes samplers, ajuste fino de prompt, pós-processamento de imagem, e integração com plugins de terceiros. Ele facilita o uso de Stable Diffusion para gerar e manipular imagens sem a necessidade de lidar diretamente com código.



API • Github • Gradio • Reload UI  
python: 3.10.6 • torch: 1.13.1+cu117 • xformers: N/A • gradio: 3.15.0 • commit: 983167e6

**Forge (Stable Diffusion WebUI Forge)** - É uma interface gráfica (web UI) especificamente projetada para usuários de Stable Diffusion que desejam um ambiente de geração de imagens mais personalizável e controlado. O Forge permite que os usuários ajustem parâmetros de geração de maneira detalhada, fornecendo uma experiência rica para quem quer experimentar diferentes configurações e técnicas de geração de imagens. Ele é uma alternativa ao Automatic1111, oferecendo um conjunto de funcionalidades avançadas, como suporte para plugins, ferramentas de ajuste de prompt, controle detalhado do processo de difusão e capacidade de manipular modelos de forma personalizada.

**Foocus** - É uma ferramenta de código aberto destinada a melhorar a experiência do usuário com a Stable Diffusion, semelhante ao Automatic1111, mas com foco em tornar o fluxo de trabalho mais intuitivo e simples para artistas e criadores de conteúdo. Ela remove a

complexidade técnica e oferece uma interface gráfica limpa e direta para gerar e ajustar imagens usando modelos de difusão.

**Hugging Face** - É uma plataforma que hospeda uma ampla gama de modelos de machine learning, incluindo modelos de transformers para tarefas como processamento de linguagem natural (NLP), visão computacional, e geração de imagens. A Hugging Face oferece um repositório colaborativo onde pesquisadores e desenvolvedores podem compartilhar e explorar modelos pré-treinados, além de uma API que facilita a integração desses modelos em aplicações práticas.

**PyTorch** - É uma biblioteca de código aberto amplamente usada para desenvolvimento de modelos de deep learning. PyTorch oferece uma abordagem dinâmica para a criação de redes neurais, permitindo que os desenvolvedores ajustem, testem e implementem seus modelos de maneira mais eficiente. A biblioteca é amplamente utilizada em pesquisa e indústria, suportando desde protótipos até grandes modelos de produção, incluindo modelos de geração de imagens.

## Diferenças

**ComfyUI** é uma interface gráfica que usa uma abordagem modular e visual, onde os usuários conectam blocos que representam diferentes etapas de geração e processamento de imagens, como uma espécie de diagrama de fluxo.

### Prós:

- Interface altamente visual e intuitiva para personalização de fluxos de trabalho.
- Flexibilidade para conectar módulos e manipular todas as etapas do processo.
- Permite fácil integração de diferentes modelos e parâmetros sem código.

### Contras:

- Pode ser menos eficiente para usuários que preferem uma interface de uso rápido sem necessidade de ajustar detalhes visuais.
- Curva de aprendizado para configurar fluxos complexos.

---

**Automatic1111** é uma interface web que se destaca pela riqueza de opções e personalizações avançadas no gerenciamento de prompts e parâmetros de geração.

**Prós:**

- Altamente personalizável com diversas opções de controle de parâmetros.
- Suporte para extensões, plugins e samplers adicionais.
- Comunidade grande e ativa, com atualizações frequentes.

**Contras:**

- A interface pode ser um pouco densa e complicada para iniciantes.
- Algumas funções avançadas podem ser difíceis de entender para novos usuários.

---

**Forge** oferece uma experiência rica e personalizável para usuários avançados, focando em modularidade e personalização extrema, similar ao A1111, mas com algumas funcionalidades diferentes e uma abordagem mais voltada para usuários que querem controle absoluto. Muito utilizada para o Stable Diffusion XL por conseguir acelerar o processo de geração em comparação ao A1111 em placas de vídeo de 8GB e 6GB de VRAM.

**Prós:**

- Ferramentas avançadas de personalização, especialmente para quem deseja ajustar minúcias da difusão.
- Integração com diversas opções de plugins e suporte de modelagem.
- Interface eficiente para usuários mais técnicos.

**Contras:**

- Menos suporte da comunidade comparado ao A1111.

**Foocus** é voltado para simplificar o uso de **Stable Diffusion**, oferecendo uma interface mais amigável e limpa em comparação com A1111 e Forge, sem sacrificar as funcionalidades principais.

**Prós:**

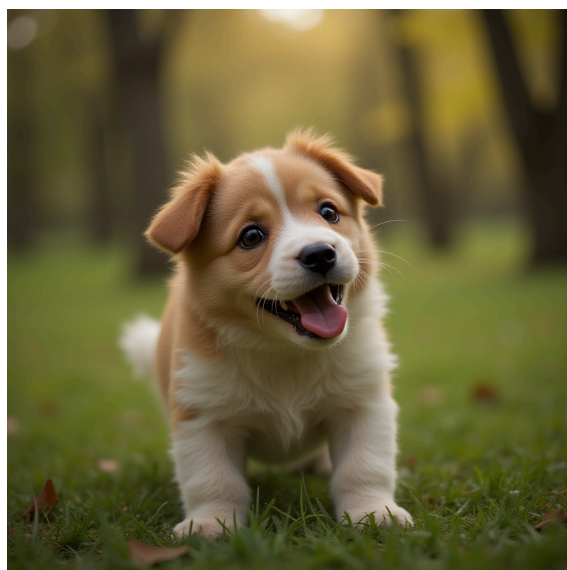
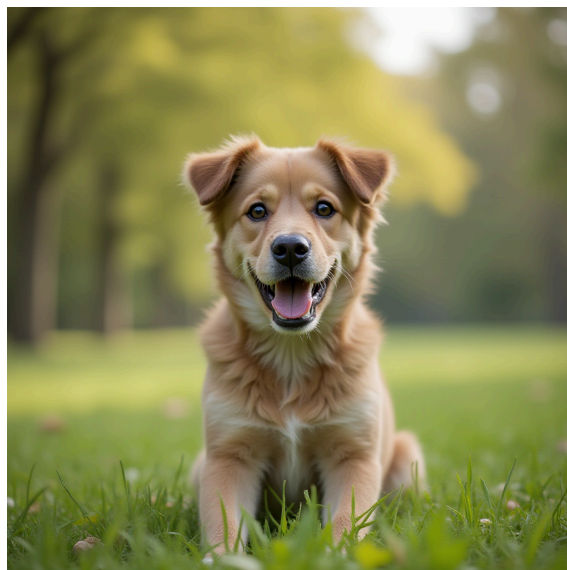
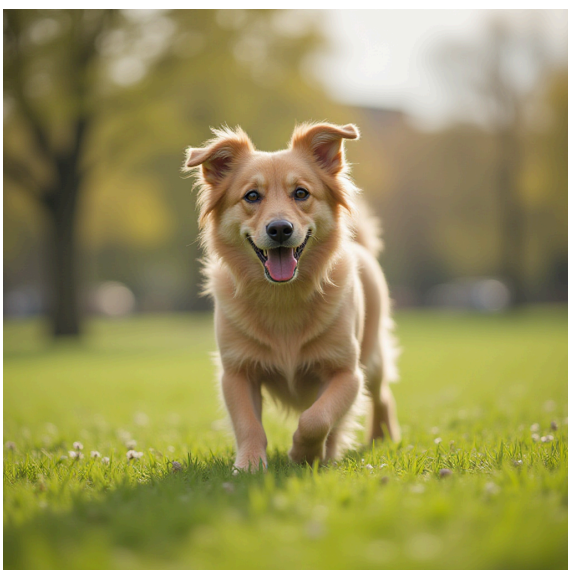
- Interface mais intuitiva e simples, com menos distrações.
- Ideal para artistas e criadores que não querem se preocupar com complexidade técnica.
- Foco em facilitar o fluxo de trabalho de geração de imagens.

**Contras:**

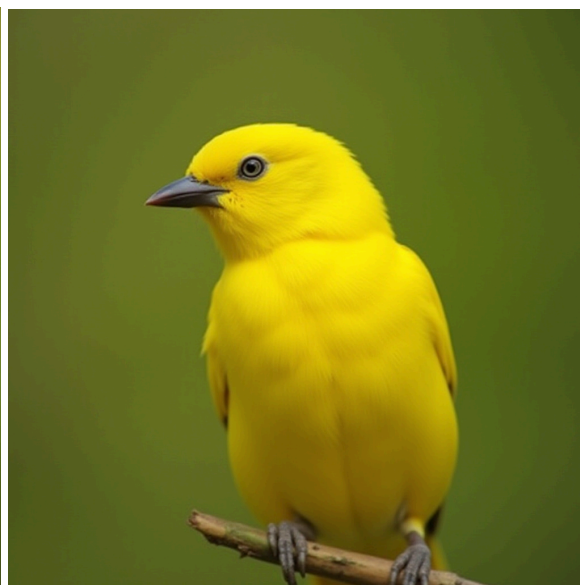
- Menos opções avançadas de personalização comparado a A1111 e Forge.
- Pode não atender usuários que procuram ajustes detalhados.

## Testes realizados

Flux 1 Schnell Q4\_0, prompt (“dog in a park”)



Flux 1 Schnell Q4\_0, prompt (“beautiful yellow bird”)



Fontes e Materiais de Referência (Links):

<https://github.com/AUTOMATIC1111/stable-diffusion-webui>

<https://github.com/comfyanonymous/ComfyUI>

<https://github.com/lllyasviel/stable-diffusion-webui-forge>

<https://github.com/lllyasviel/Fooocus>

<https://huggingface.co/CompVis/stable-diffusion-v1-4>

<https://pytorch.org/foundation>

## Termo de Aceite de Entrega

### Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

**Data da Reunião (“gate”) de aprovação:** 31 de out. de 2024

**Participantes da Entrega** [matriculados em Residência em IA]:

Arthur Jung Barreto

**Entrega:** [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

#### Documentação e Testes

- **Parâmetros:**
  - **Steps** - número de etapas de difusão que o modelo executará
  - **CFG** - técnica usada para controlar o quanto o modelo deve seguir a orientação textual fornecida.
  - **Sampler** - algoritmo usado para amostrar as imagens durante o processo de difusão.
  - **Scheduler** - sequência de ajustes feitos nos parâmetros de difusão ao longo das etapas
- **Otimizar os parâmetros**
  - Para o modelo [RealisticVision](#)
    - Steps: 20-35
    - CFG: 1.0-2.5 (1.0 para maior realismo)
    - Sampler: DPM++ 2M SDE GPU
    - Scheduler: Karras ou Beta
    - LoRA: Add Detail + More Details (0.5, 0.5)
    - Upscaling: 4x-UltraSharp (2x)
    - ImageSize: 576x768 ou 512x512
- **LoRA**
  - [Add Detail](#)
  - [More Details](#)
- **Upscaling**
  - [4x-UltraSharp](#)
  - [ESRGAN\\_4x](#)

**Planejamento:** [descrever o que pretende fazer para realizar a próxima ENTREGA]

- Tarefa - Gerar imagem de uma cena de uma história (Explorar a criação de prompts que descrevem

a cena usando LLM)

- Explorar a geração de vídeos

**Observação: [caso precise fazer alguma observação, de qualquer “natureza”]**

---

## ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: [Go! ▾](#)

## Parametrização

Entender os parâmetros é crucial para otimizar o processo e obter resultados desejáveis. No contexto da geração de imagens usando modelos de difusão existem alguns parâmetros que são importantes no momento de geração da imagem, como por exemplo: Steps, CFG, Sampler, Schedule.

**Steps** referem-se ao número de etapas de difusão que o modelo executará para transformar a imagem inicial de ruído em uma imagem final clara. Mais steps geralmente significam uma imagem de qualidade mais alta, mas também aumentam o tempo de processamento.



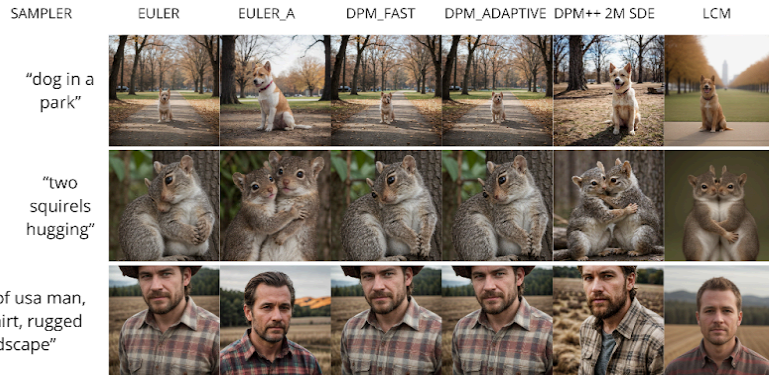
Test documentation  
**Model:** RealisticVisionV5.1 Hyper (VAE) Pruned  
**Steps:** Dynamic  
**CFG:** 1.0  
**Sampler:** DPM++ 2M SDE GPU  
**Scheduler:** Karras  
**Seed:** 0, Fixed  
**Image Size:** 512x512

**CFG (Class-Free Guidance)** é uma técnica usada para controlar o quanto o modelo deve seguir a orientação textual fornecida. Um valor mais alto de CFG faz com que a imagem gerada se alinhe mais de perto com a descrição textual, enquanto um valor mais baixo permite resultados mais criativos e diversificados.



Test documentation  
**Model:** RealisticVisionV5.1 Hyper (VAE) Pruned  
**Steps:** 30  
**CFG:** Dynamic  
**Sampler:** DPM++ 2M SDE GPU  
**Scheduler:** Karras  
**Seed:** 0, Fixed  
**Image Size:** 512x512

**Schedule** refere-se ao cronograma ou sequência de ajustes feitos nos parâmetros de difusão ao longo das etapas. Isso pode incluir ajustes na taxa de difusão para melhorar a clareza e a qualidade da imagem final.



Test documentation  
**Model:** [RealisticVisionV5.1 Hyper \(VAE\) Pruned](#)  
**Steps:** 30  
**CFG:** 1.0  
**Sampler:** Dynamic  
**Scheduler:** Karras  
**Seed:** 0, Fixed  
**Image Size:** 512x512

**Sampler** é o método ou algoritmo usado para amostrar as imagens durante o processo de difusão. Diferentes samplers podem afetar a qualidade, a velocidade e o estilo das imagens geradas, oferecendo uma variedade de abordagens para ajustar o processo de geração de imagens.



Test documentation  
**Model:** [RealisticVisionV5.1 Hyper \(VAE\) Pruned](#)  
**Steps:** 30  
**CFG:** 1.0  
**Sampler:** DPM++ 2M SDE GPU  
**Scheduler:** Dynamic  
**Seed:** 0, Fixed  
**Image Size:** 512x512

## LoRA

**LoRA**, ou "Low-Rank Adaptation", é uma técnica utilizada para adaptar modelos de inteligência artificial, especialmente em redes neurais, para tarefas específicas sem a necessidade de ajustar todos os parâmetros do modelo base. No contexto da geração de imagens, como em modelos de difusão ou GANs (Redes Adversariais Generativas), LoRA permite que o modelo seja especializado em um novo conjunto de dados ou estilo de imagem de forma mais eficiente e com menor custo computacional.



Test documentation

**Model:** [RealisticVisionV5.1 Hyper \(VAE\) Pruned](#)  
**Steps:** 30  
**CFG:** 1.0  
**Sampler:** DPM++ 2M SDE GPU  
**Scheduler:** Karras  
**Seed:** 0, Fixed  
**Image Size:** 512x512  
**Clip Skip:** 2

LoRA documentation

**Models:** [ADD\\_DETAIL](#), [MORE\\_DETAILS](#)  
**Strength Model:** 0.5  
**Strength Clip:** 0.5

Por trás dos panos, **LoRA** funciona ao inserir camadas ou matrizes de baixa dimensão (low-rank) entre as camadas do modelo original. Em vez de ajustar todos os pesos do modelo, apenas essas camadas adicionais são treinadas, o que reduz significativamente o número de parâmetros a serem atualizados. Isso permite que o modelo retenha seu conhecimento original enquanto é rapidamente adaptado para novas tarefas ou estilos de imagem.

Durante a geração de imagem, o modelo adaptado com **LoRA** utiliza essas camadas de baixa dimensão para aplicar as alterações necessárias, permitindo a criação de imagens que aderem às novas características aprendidas. Isso resulta em um processo mais rápido e eficiente, mantendo a qualidade e a diversidade das imagens geradas. Em resumo, **LoRA** é uma abordagem poderosa para especializar modelos de IA em novas tarefas de geração de imagens sem comprometer a eficiência e a performance.

## Upscaling

O upscaling no contexto de geração de imagens refere-se ao processo de aumentar a resolução de uma imagem digital, aumentando o número de pixels, de modo a melhorar sua qualidade visual em telas de alta definição. Este processo é particularmente relevante quando se trabalha com imagens de baixa resolução, que precisam ser exibidas em dispositivos com resoluções maiores, como televisores 4K ou monitores de alta definição.



Test documentation

**Model:** [RealisticVisionV5.1 Hyper \(VAE\) Pruned](#)  
**Steps:** 30  
**CFG:** 1.0  
**Sampler:** DPM++ 2M SDE GPU  
**Scheduler:** Karras  
**Seed:** 0, Fixed  
**Image Size:** 512x512  
**Clip Skip:** 2

LoRA documentation

**Models:** [ADD\\_DETAIL](#), [MORE\\_DETAILS](#)  
**Strength Model:** 0.5  
**Strength Clip:** 0.5

Upscaling documentation

**Upscale Model:** [4x-UltraSharp](#)  
**Upscale by:** 2.0  
**Model:** [RealisticVisionV5.1 Hyper \(VAE\) Pruned](#)  
**Steps:** 10  
**CFG:** 1.0  
**Sampler:** DPM++ 2M SDE GPU  
**Scheduler:** Karras  
**Seed:** 0, Fixed  
**Denoise:** 0.2

Na implementação do upscaling durante o pós-processamento, técnicas avançadas como algoritmos de interpolação bicúbica ou bilinear são frequentemente usadas para prever e adicionar pixels entre aqueles já existentes, suavizando transições e preservando detalhes. Recentemente, o uso de inteligência artificial e aprendizado de máquina também se tornou popular, com redes neurais treinadas para reconhecer padrões e texturas, preenchendo lacunas de maneira mais precisa e natural. Essas técnicas baseadas em IA, como redes generativas adversariais (GANs), conseguem resultados impressionantes, recuperando detalhes que seriam perdidos com métodos tradicionais, e garantindo que a imagem final pareça mais autêntica e detalhada.

É possível visualizar melhor as imagens em [Documentação e Testes](#).

## APÊNDICE 4

## Termo de Aceite de Entrega

### Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

**Data da Reunião (“gate”) de aprovação:** 6 de nov. de 2024

**Participantes da Entrega** [matriculados em Residência em IA]:

Arthur Jung Barreto

**Entrega:** [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

#### Documentação

- Objetivo - Gerar imagem de uma cena
  - Geração de Prompt
    - ChatGPT - Versão 4o
      - O objetivo é capturar uma cena textual e extrair dela uma sequência de prompts que, além de descreverem a cena, sigam uma ordem cronológica das ações, permitindo uma compreensão sequencial e detalhada dos eventos.
      - [Imagens Geradas](#)
  - Aprimoramento de Prompt (Prompt Enhancement)
    - Prompt Quill - Utiliza RAG com quase 5 milhões de prompts
      - O objetivo é melhorar os prompts gerados para criar imagens melhores.
- Geração de vídeos
  - AnimateDiff - [Exemplos de animações](#)

**Planejamento:** [descrever o que pretende fazer para realizar a próxima ENTREGA]

- Estudar, explorar:
  - Prompt Engineering (Específico para geração de prompts)
  - Prompt Enhancement
- Realizar Testes
- Visualizar e definir um fluxo de trabalho para a tarefa

**Observação:** [caso precise fazer alguma observação, de qualquer “natureza”]

## ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: [Go! ▾](#)

## Prompt Generation

ChatGPT 4o - Usar o ChatGPT para descrever uma cena de texto e transformá-la em uma sequência de prompts para geração de imagens é uma técnica poderosa para dar vida a narrativas escritas. Esse processo consiste em identificar os principais elementos de uma cena textual — como personagens, cenários, ações e atmosferas — e convertê-los em prompts concisos que instruem um modelo de geração de imagens.

Essa técnica permite transformar uma cena escrita em uma série de prompts que, juntos, representam a narrativa visualmente, ideal para aplicações como storyboards ou ilustrações digitais.

## Imagens Geradas

Texto da Cena:

Fui ontem ao Pireu com Glauco, filho de Ariston, para orar à deusa, e também para me certificar de como seria a festividade, que eles promoviam pela primeira vez. A procissão dos atenienses foi bastante agradável, embora não me parecesse superior à realizada pelos trácios. Após termos orado e admirado a cerimônia, estávamos regressando à cidade quando, no caminho, fomos vistos à distância por Polemarco, filho de Céfalo. Ele mandou seu jovem escravo correr até nós, para nos pedir que o esperássemos.

Prompts Gerados:

Prompt 1:

"Ancient Greek city port of Piraeus, bustling with activity, with Greek temples and marble columns in the background under a clear sky. Two men in traditional Greek tunics, one named Glauco, son of Ariston, are standing in a reverent pose as they prepare to pray to the goddess. Lively crowd and scenic port visible in the background."

Prompt 2:

"A solemn Greek religious festival procession with Athenians in white and colored tunics, carrying offerings and banners as they walk towards a grand Greek temple. The scene is marked by reverent faces, some with garlands, others holding baskets of fruit and flowers. Traditional music instruments are played by participants, adding to the ceremony's vibrant atmosphere."

Prompt 3:

"Close-up of Glauco and a friend, watching the Athenian festival procession with expressions of admiration. They wear simple tunics, observing as the Athenians participate in their first-time festival for the goddess. The crowd is arranged in orderly lines, showing respect and deep cultural pride."

Prompt 4:

"Side-by-side comparison of two processions: Athenians with white and pastel tunics moving in a calm, ordered line, contrasted against Thracians with bright, colorful clothing, lively movements, and diverse musical instruments. The Athenians' solemn mood contrasts with the festive, energetic Thracian procession."

Prompt 5:

"Two men, Glauco and his friend, walking away from the Piraeus temple, discussing the ceremony as they start their journey back to the city. The sun begins to set, casting a warm glow over the ancient Greek landscape with distant hills and cityscape visible. They look content and thoughtful, reflecting on the event they just witnessed."

Prompt 6:

"A distant view of two figures walking along a dirt path back to the city of Athens. They are seen from afar by Polemarco, son of Cephalus, who stands on a higher path, dressed in traditional Greek attire. He raises his hand, noticing his friends with a look of recognition and eagerness to catch up."

Prompt 7:

"Young Greek servant or slave running towards Glauco and his friend with a sense of urgency, sent by Polemarco. The servant is dressed in simple attire, dashing down a dusty path under the golden light of the evening. Glauco and his friend look back, surprised and pausing to wait."

Imagens geradas:



### Prompt Enhancement

Prompt Quill - É uma ferramenta projetada para ajudar os usuários a criar prompts melhores para a geração de imagens com IA, como modelos do tipo Stable Diffusion. Essa solução, alimentada por recuperação aumentada por geração (RAG), utiliza uma vasta base de quase 5 milhões de prompts armazenados em um banco de vetores, facilitando a inspiração e a criação de instruções detalhadas para obter imagens de alta qualidade.

Além de facilitar o desenvolvimento de prompts para iniciantes e especialistas, Prompt Quill oferece uma interface Gradio que simplifica a criação de textos mais sofisticados para imagem. Também inclui uma opção para teste automatizado de modelos, útil para aqueles que treinam modelos de geração de imagens.

## Vídeo Generation

AnimateDiff - É uma tecnologia de geração de animações que utiliza modelos de difusão (como o Stable Diffusion) para criar vídeos animados a partir de prompts de texto. Ele adapta o processo de geração de imagens para gerar frames sequenciais, garantindo que cada frame mantenha continuidade com o anterior, criando assim uma animação fluida e visualmente coesa.

- **Consistência entre Frames:** O modelo é otimizado para gerar frames que mantêm consistência visual entre si — elementos como iluminação, cor e posição são controlados para evitar "saltos" entre os frames, garantindo uma transição suave.
- **Aplicações e Usos:** AnimateDiff é útil para criar animações curtas, loops visuais, vídeos artísticos e até mesmo protótipos de cenas para animações mais complexas.

Com o AnimateDiff, é possível gerar vídeos curtos e animados a partir de uma descrição textual, oferecendo uma nova maneira de produzir conteúdo visual. [Exemplos de animações](#)

Fontes e Materiais de Referência (Links):

[Animate Diff](#)

[Prompt Quill](#)

## Termo de Aceite de Entrega

### Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

**Data da Reunião (“gate”) de aprovação:** 14 de nov. de 2024

**Participantes da Entrega** [matriculados em Residência em IA]:

Arthur Jung Barreto

**Entrega:** [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

#### Documentação

- Prompt Engineering
  - 2 ou 3 vezes
  - Criação do texto narrativo (Opcional)
  - Criação de descrições
  - Criação e melhoramento de prompts
- Prompt Enhancement
  - Prompt Quill vs Homemade
- Fluxo de Trabalho
  - Visualização na documentação
- Testes
  - [Prompt Generation](#) - [Resultado dos testes](#)
  - [Modelo](#) utilizado
- Artigos
  - Qualidade - Modelo
  - Consistência - Arquitetura e Modelo
  - Versatilidade - Arquitetura

**Planejamento:** [descrever o que pretende fazer para realizar a próxima ENTREGA]

- Concluir a tarefa focando na Consistência pela Arquitetura

**Observação:** [caso precise fazer alguma observação, de qualquer “natureza”]

- Napkin AI - <https://www.napkin.ai/>

## ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: [Go!](#)

## Visualização do Fluxo de Trabalho

Essa é uma visualização do fluxo de trabalho detalhado para a criação de imagens ou vídeos a partir de texto narrativo, utilizando uma abordagem de transformação de texto em prompts através de uma LLM. O processo inclui a melhoria dos prompts gerados, visando otimizar a entrada para um modelo de difusão estável, que é responsável pela geração final das imagens ou vídeos. A visualização abaixo:

### Transformando Histórias em Visuais

#### Gerar Visuais

Usar um modelo de difusão para produzir as imagens ou vídeos finais.

#### Refinar Prompts

Aprimorar e otimizar os prompts para melhor qualidade de entrada.

#### Gerar Prompts

Criar prompts iniciais a partir do texto da história usando um modelo de linguagem.



Essa imagem foi gerada através da plataforma Napkin AI

## Prompt Engineering

Prompt Engineering é uma técnica que visa explorar o aperfeiçoamento do prompt a fim de se obter uma resposta melhor do modelo. Em função do fluxo que trabalho apresentado o prompt engineering vai aparecer 2 ou 3 vezes dependendo se o texto narrativo já estiver pronto. A primeira vez é sobre a criação do texto narrativo, uma vez

tendo uma breve descrição do que o usuário gostaria a LLM deve gerar um texto narrativo, caso o texto já esteja pronto ou tenha sido escrito pela LLM temos a segunda aplicação do prompt engineering onde a LLM deve pegar o texto narrativo e gerar um conjunto de descrições seguidos de trilhas sonoras que podem ser utilizadas para cada trecho que ela identificar no texto narrativo, por fim a terceira aplicação é na conversão de descrições para prompt onde também será aplicada uma técnica de prompt enhancement.

### **Prompt Enhancement**

Prompt Enhancement é uma técnica que visa aprimorar o prompt já existente em um prompt que gere melhores resultados para um modelo específico, nesse caso para a utilização do modelo Stable Diffusion V1.5. Existem formas de realizar o prompt enhancement, uma delas citada anteriormente é o Prompt Quill que pega um conjunto enorme de prompt lançados em um banco de dados vetorial e através de RAG seleciona os “n” mais parecidos com o prompt inicial para usar de exemplo em uma LLM para que ela converta o prompt inicial em um prompt semelhante aos selecionados, basicamente uma técnica de prompt engineering.

O Prompt Quill não funcionou corretamente para mim, consegui ter acesso a base de dados porém o restante apresentou muitos erros, então decide realizar um “Prompt Quill caseiro” onde ao invés de possuir quase 5 milhões de prompts e então escolher alguns semelhantes, apenas selecionei prompts de imagens geradas no site Civitai que eu considero boas e coloquei como exemplo para uma LLM.

## Abordagens para aprimoramento de prompt em fluxos de trabalho de narrativa.



### Prompt Quill

Aproveita um grande banco de dados para correspondência de similaridade



### Homemade Prompt Quill

Usa exemplos selecionados para aprimoramento direcionado

Essa imagem foi gerada através da plataforma Napkin AI

### Testes realizados

A ideia geral para os testes é:

- Demonstrar a capacidade de geração de textos narrativos melhores pela LLM usando Prompt Engineering
- Demonstrar a capacidade de geração de descrições de textos narrativos junto de uma trilha sonora adequada ao trecho melhores pela LLM usando Prompt Engineering
- Demonstrar a capacidade de geração de prompts com as descrições melhores pela LLM usando Prompt Engineering
- Demonstrar a melhoria das imagens e vídeos gerados pela utilização da técnica Homemade Prompt Quill
- Demonstrar todo o funcionamento do conjunto
- [PromptGeneration](#) - [Resultado dos testes](#)

---

## Artigos

Encontrei um termo que pode ser usado para se designar a essa tarefa que estou realizando, “Story-to-Image”, “Story-to-Video”, “Story-to-Visual”, então procurei por artigos que tinham propostas semelhantes. Artigos encontrados:

Make-A-Story: Visual Memory Conditioned Consistent Story Generation

TaleCrafter: Interactive Story Visualization with Multiple Characters

AutoStory: Generating Diverse Storytelling Images with Minimal Human Effort

Esses artigos apresentam pontos importantes para a realização da tarefa:

- (1) high quality: the generated images must be visually appealing and have a reasonable layout
- (2) consistency: not only the generated images should be consistent to the text descriptions, but also the identities of the characters and scenes in different images should be consistent; and
- (3) versatility: to satisfy a wide range of users’ needs, it needs to be able to be easily applied to different styles, characters, and scenes.


## Sites

Encontrei ainda alguns sites que tentam realizar a tarefa:

<https://perchance.org/ai-story-with-pictures>

# AI Story Generator


(with pictures!)

 delete story and restart

Um grego antigo ao lado de seu cachorro explorando sua região

 delete  stop  regen

Em meio à exuberante paisagem da Grécia Antiga, um velho e sábio grego de barba cana e olhos azuis penetrantes ponderava as maravilhas da natureza com a companhia inseparável de seu leal cachorro, Argos. As pedras arenosas do caminho sob os pés descalços do homem e os latidos distantes de Argos ecoando pelas colinas, eram testemunhas silenciosas da inabalável amizade que os unia. Os dias de sol ardente e os verdes vales repletos de oliveiras e vinhedos se estendiam à sua volta, e a cada passo que dava, o grego sentia o peso da história antiga que jazia em cada pedaço da terra que calçava. Juntos, exploravam a beleza da sua região com o entusiasmo de um criança e a sabedoria de alguém que sabia que cada pedregulho e cada vento que acariciava o rosto eram pequenos fragmentos de um universo imensurável e misterioso que aguardava por ser descoberto.

 insert block here

 delete  stop  regen




delete stop regen

Enquanto caminhavam, o grego antigo contemplou os monumentos antigos que punham a paisagem: restos de templos deuses abandonados, estátuas erodidas pelo tempo e vestígios de antigas estradas que outrora testemunharam o passar dos heróis e filósofos que tinham dado vida à terra que agora andavam. Argos, o cachorro, corria alegremente à procura de tesouros escondidos e aromas desconhecidos, mas parava de repente, erguendo as orelhas, cada vez que detectava o som de passos invisíveis ou o cheiro do vento trazendo histórias de terras distantes. O homem e a besta andavam em harmonia, a sabedoria do passado fluindo no ar à medida que os diálogos sussurravam em cada som da natureza que os envolvia. A cada encontro com um viajante ou pastor que cruzava o caminho, o grego compartilhava histórias antigas e conselhos sábios, e em troca, recebia notícias do vento e do sol que brilhava nas terras vizinhas. A vida era simples, mas cheia de maravilhas e sabedoria para aqueles que sabiam ouvir e observar. E assim, com a paciência do tempo e o olhar penetrante do saber, o grego e Argos continuaram a explorar a beleza da Grécia Antiga, desvendando os segredos que a terra lhes oferecia e fortalecendo a eterna amizade que os mantinha unidos.

+ insert block here

delete stop regen



<https://flowgpt.com/p/situational-ai-story-and-image-generator>

Não consegui acessar com uma conta para gerar as imagens mas ele gera o texto narrativo.

Fontes e Materiais de Referência (Links):

<https://civitai.com/models/4468/counterfeit-v30>

<https://huggingface.co/g sdf/Counterfeit-V3.0>

[Make-A-Story: Visual Memory Conditioned Consistent Story Generation](#)

[TaleCrafter: Interactive Story Visualization with Multiple Characters](#)

[AutoStory: Generating Diverse Storytelling Images with Minimal Human Effort](#)

[https://www.canva.com/design/DAGWdav0qG4/5IDQj2vD9SM1p1Nh\\_HarWw/edit?utm\\_content=DAGWdav0qG4&utm\\_campaign=designshare&utm\\_medium=link2&utm\\_source=sharebutton](https://www.canva.com/design/DAGWdav0qG4/5IDQj2vD9SM1p1Nh_HarWw/edit?utm_content=DAGWdav0qG4&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton)

<https://colab.research.google.com/drive/1EsAjUoqLFNLZRpFYiu34DT7bX-5OnYR?usp=sharing>

## APÊNDICE 5

## Termo de Aceite de Entrega

### Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

**Data da Reunião (“gate”) de aprovação:** 28 de nov. de 2024

**Participantes da Entrega** [matriculados em Residência em IA]:

Arthur Jung Barreto

**Entrega:** [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

#### Documentação

- Consistência na criação de personagens
  - Criar personagem
    - Workflow pronto para (Flux 1, SD 3.0)
      - ControlNet - OpenPose
  - LoRA
    - Treinar LoRA (Flux 1, SD 1.5)
      - FluxGym
      - Kohya ss
      - One Trainer
    - Baixar LoRA (SD 1.5)
- Geração de vídeo
  - Animated Diff
  - Cog Video
  - Pyramid Flow
- Alternativas
  - Chat GPT 4o + DALLE 2 - Consistência de personagens
  - Kling AI - Geração de vídeo
- Adicionais
  - Face Detailer
  - RIFE VFI
  - Comfy UI API (permite integração do criador de textos narrativos Chat GPT com o modelo de geração de imagens)

**Planejamento:** [descrever o que pretende fazer para realizar a próxima ENTREGA]

- Conclusão
  - Comparar a solução com as alternativas de story-to-video
  - Entregar vídeos completos
  - Visão da sociedade sobre geração de imagens/vídeos
    - Ética
    - Usos
  - Geração de imagens/vídeos hoje e quais os possíveis rumos dessa tecnologia

**Observação: [caso precise fazer alguma observação, de qualquer “natureza”]**

---

## ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: [Go! ▾](#)

## Consistência na criação de personagens

Criação de personagens, nessa etapa temos que definir como será o nosso personagem e descrevê-lo no prompt para ser gerado, existe uma técnica que permite a criação de uma imagem onde o personagem estará em diferentes posições e em diferentes ângulos e para isso nós utilizamos o Control Net.

Control Net - Open Pose, serve para muitas outras coisas, inclusive é muito bom para gerar imagens ou vídeos que o personagem tenha que estar em uma determinada posição, mas para este caso utilizaremos para criar uma imagem onde o personagem estará em diferentes posições.

Flux vs SD 3.0 vs SD 1.5, encontrei um workflow para comfy UI que já realiza todo o esse processo automaticamente, porém esse workflow só funciona para modelos como o Flux e o Stable Diffusion 3.0 que são modelos grandes, para os modelos utilizados por mim como o Stable Diffusion 1.5 eu precisei fazer algumas adaptações e mesmo assim não consegui gerar adequadamente todas as imagens do personagem devido a dimensão das imagens geradas.

LoRA, eu encontrei um método para criar personagens consistentes, que consiste em treinar um LoRA do seu personagem e depois aplicá-lo no momento da geração. Para essa tarefa então será necessário treinar um LoRa.

Treinar LoRA, encontrei dificuldades no treinamento, primeiramente porque comecei tentando treinar um LoRA de modelos maiores como o Flux já que havia conseguido criar boas imagens de um personagem com ele, e então busquei algumas ferramentas para esse treinamento.

Flux Gym vs Kohya ss vs One Trainer, essas foram as três ferramentas encontradas para realizar o treinamento do LoRA, as duas primeiras para modelos grandes como o Flux onde não obtive êxito, os motivos são erros sem solução no momento do treinamento, toda a preparação de dados foi realizada corretamente, foi aí que optei por treinar o LoRA de

modelos menores SD 1.5 utilizando a ferramenta do One Trainer, finalmente consegui treinar um LoRA, mas não funcionou, a segunda etapa era utilizar o mesmo para gerar imagens consistentes do meu personagem o que não aconteceu.

Baixar LoRA, já que as coisas não saíram como o planejado decide então fazer o download de u, LoRA de algum personagem que alguém disponibilizou na internet para testar se isso realmente funcionava, e sim funcionou, porém sem conseguir treinar o LoRA eu mesmo isso quebra o fluxo de trabalho planejado, podendo apenas utilizar personagens que outras pessoas criaram.

### **Vídeo Generation**

Animated Diff é uma técnica avançada para animar modelos de difusão texto-para-imagem (T2I), como o Stable Diffusion, sem a necessidade de ajustes específicos no modelo base.

Cog Video é um modelo de geração de vídeo a partir de texto. Ele possui 9,4 bilhões de parâmetros e foi treinado em 5,4 milhões de pares de texto e vídeo. Uma característica do CogVideo é a capacidade de produzir vídeos de 4 segundos com uma resolução de 768 x 1360 pixels a 16 quadros por segundo, mantendo movimentos fluidos e coerentes com o texto fornecido.

Pyramid Flow é uma ferramenta open-source para geração de vídeos, desenvolvida por pesquisadores de universidades chinesas em colaboração com a Kuaishou Technology. Ela utiliza uma técnica inovadora chamada pyramidal flow matching, que combina eficiência computacional e qualidade visual, gerando vídeos de até 10 segundos com resoluções de 384p a 768p e 24 FPS. Essa abordagem reduz custos computacionais ao construir vídeos progressivamente, começando com resoluções mais baixas e refinando até atingir a alta resolução final.

### **Alternativas**

[Chat GPT 4o + DALLE 2](#) - Consistência de personagens

[Tensor Art](#) - Consistência de personagens

[Kling AI](#) - Geração de vídeo

## Adicionais

[Face Detailer](#) - é um método de pós-processamento de imagens, muito útil quando o modelo não consegue gerar os rostos corretamente, uma yolo identifica o rosto na imagem e faz com que o modelo gere novamente somente a parte do rosto.

[RIFE VFI](#) e [RIFE VFI Comfy UI](#)- Real-Time Intermediate Flow Estimation for Video Frame Interpolation, é um algoritmo que faz a interpolação entre frames de vídeos possibilitando aumentar a quantidade de frames que por sua vez gera uma sensação de suavidade para o mesmo.

[Comfy UI API](#) - API que faz uma requisição de geração de imagem sem passar pela interface o que permite uma integração do criador de textos narrativos Chat GPT com o modelo de geração de imagens via código python.

## Resultados

Muitos vídeos e imagens geradas, mas não consegui compilar tudo isso de uma forma que fique fácil a compreensão, segue só alguns exemplos para visualização:

[Resultados-Canva](#)

Fontes e Materiais de Referência (Links):

<https://github.com/kijai/ComfyUI-PyramidFlowWrapper>

<https://pyramid-flow.github.io/>

<https://arxiv.org/abs/2410.05954>

<https://huggingface.co/rain1011/pyramid-flow-miniflux>

<https://huggingface.co/THUDM/CogVideoX-2b>

<https://github.com/Kosinkadink/ComfyUI-AnimateDiff-Evolved?tab=readme-ov-file>

<https://www.youtube.com/watch?v=FOtJ1Chxy6Y>

---

<https://www.runcomfy.com/pt/comfyui-workflows/consistent-characters-with-flux-comfyui-workflow>

<https://www.youtube.com/watch?v=MbQv8zoNEfY>

<https://www.patreon.com/posts/free-workflows-113743435>

<https://github.com/city96/ComfyUI-GGUF>

<https://www.youtube.com/watch?v=va8Jkc7o9d4>

[https://github.com/Nuked88/DreamingAI/blob/main/dreaminAI\\_websockets\\_api\\_example.py](https://github.com/Nuked88/DreamingAI/blob/main/dreaminAI_websockets_api_example.py)

## Termo de Aceite de Entrega

### Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

**Data da Reunião (“gate”) de aprovação:** 4 de dez. de 2024

**Participantes da Entrega** [matriculados em Residência em IA]:

Arthur Jung Barreto

**Entrega:** [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

#### Documentação

- Comparações:
  - Inv Video AI
  - Fliki AI
  - Pictory AI
- Entrega Final (2 vídeos):
  - [Resultado](#)
- Casos:
  - Coreia do Sul
  - Taylor Swift
- Atualidades e próximos passos da tecnologia

**Planejamento:** [descrever o que pretende fazer para realizar a próxima ENTREGA]

**Observação:** [caso precise fazer alguma observação, de qualquer “natureza”]

## ACEITE DA ENTREGA:

**CEDRIC LUIZ DE CARVALHO:** Go!

## Comparações

Inv Video AI: é uma plataforma projetada para criar vídeos automaticamente a partir de ideias ou prompts textuais. Ele oferece funcionalidades como:

- Geração de scripts personalizados para diversos tópicos.
- Uso de mais de 16 milhões de imagens e vídeos de banco de dados para compor os vídeos.
- Criação de vídeos "faceless" (sem rosto humano), adequados para nichos como educação, motivação e narrações.
- Geração de narrações com vozes humanas em vários idiomas e sotaques.
- Integração de música de fundo e transições automáticas. É uma ferramenta ideal para criar vídeos para redes sociais ou automação de canais no YouTube, sem a necessidade de habilidades avançadas em edição

Fliki AI: transforma textos em vídeos curtos e envolventes, voltados para marketing e redes sociais. A ferramenta se destaca por:

- Geração automática de vídeos a partir de textos simples.
- Uso de bibliotecas de mídia integrada para vídeos e imagens.
- Narrações com vozes realistas e opções de personalização. Ela é útil para criar conteúdo rápido e visualmente atrativo para branding e campanhas de marketing.

Pictory AI: permite transformar conteúdos longos, como artigos e transcrições, em vídeos curtos e dinâmicos. As principais características incluem:

- Resumo automatizado de textos longos em cliques curtos.
- Edição baseada em texto, permitindo ajustar cenas e conteúdo rapidamente.
- Geração de vídeos para storytelling, marketing e tutoriais. Ideal para criadores de conteúdo que desejam reaproveitar material escrito para formatos de vídeo.

## Atualidades e próximos passos da tecnologia


O que existe hoje em questão de geração de imagens e vídeos já é espetacular para diversas atividades, porém ainda existem pontos que precisam ser resolvidos. Destaco também o que ocorre com as melhores ferramentas para esse fim, existem diversas ferramentas gratuitas e que podem ser utilizadas localmente, porém as novas ferramentas que se destacam, ou não são abertas ao público (não é possível utilizar) ou não são disponibilizadas para serem utilizadas localmente uma vez que só através dos sites ou via API você tem acesso e muitas vezes o acesso é pago. Sobre a questão dos recursos computacionais necessários, para a inferência (geração da imagem ou vídeo) a grande maioria dos modelos podem ser executados em GPUs de desktop é claro as melhores delas com no mínimo 8GB de VRAM, entretanto, para o treinamento desses modelos é preciso de muito poder computacional, o que usuários comuns não têm acesso.

O futuro da geração de imagens depende do futuro dos computadores, o poder computacional vai ditar a acessibilidade a essas ferramentas e a qualidade dessas ferramentas, além dessa dependência existe a possibilidade do surgimento de novas arquiteturas. Bom, essa tecnologia tem diversos usos na indústria de entretenimento, jogos, filmes, vídeos em redes sociais, tem usos na educação, gerar visualizações e usos na criação de dados sintéticos para outros fins.

Vídeos gerados completos (2 vídeos e descrições):

[Vídeos gerados \(Resultado Final\)](#)

Pasta com coleção de imagens e vídeos gerados sem muita organização:

 Vídeos e Imagens Geradas

Fontes e Materiais de Referência (Links):

<https://invideo.io/make/ai-faceless-video-generator/>

<https://fliki.ai/>

<https://pictory.ai/>