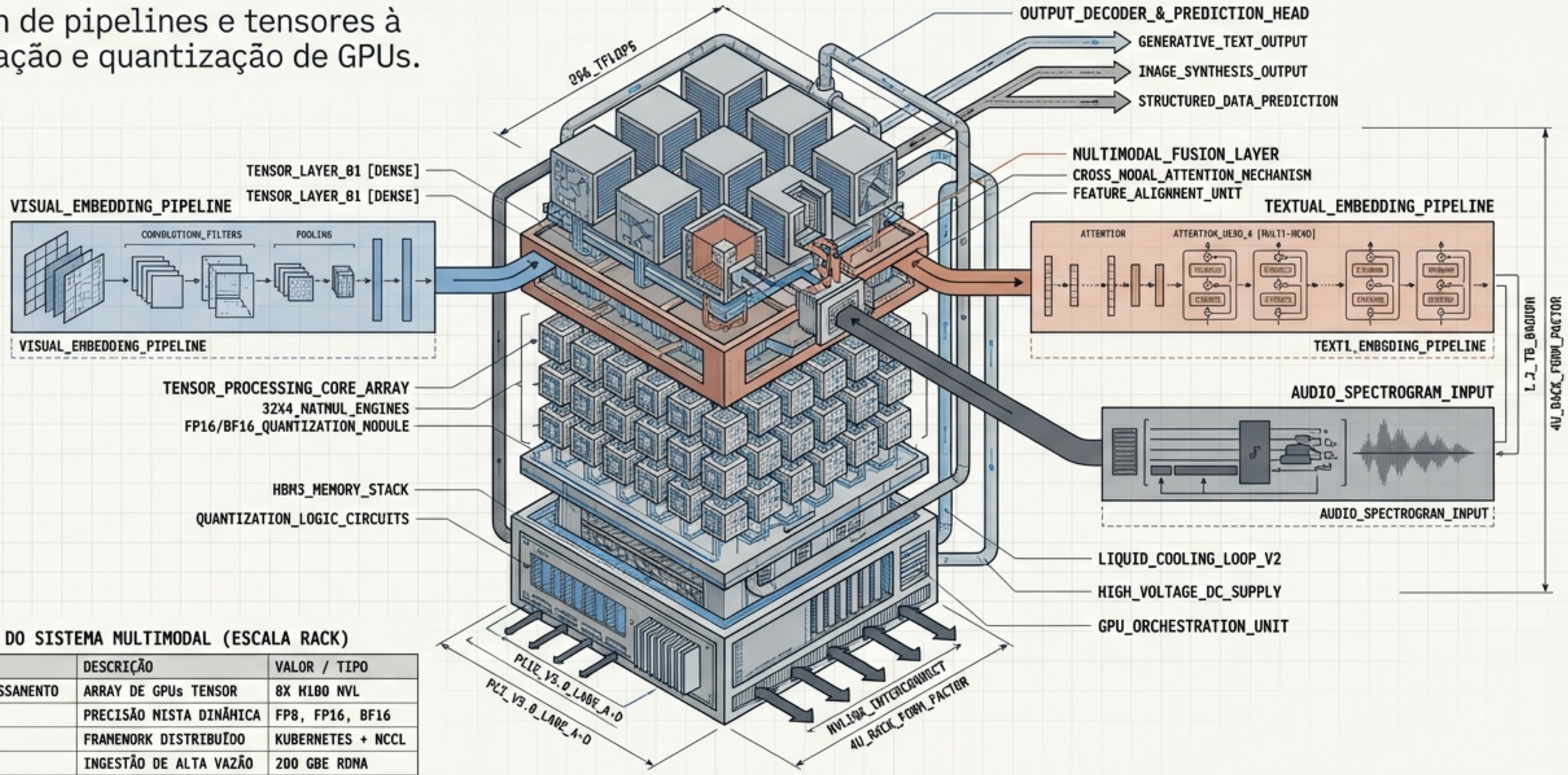
BLUEPRINT DE SISTEMAS: ARQUITETURA DE IA MULTIMODAL

[CONFIDENTIAL / ENGINEERING_DRAFT_V1]

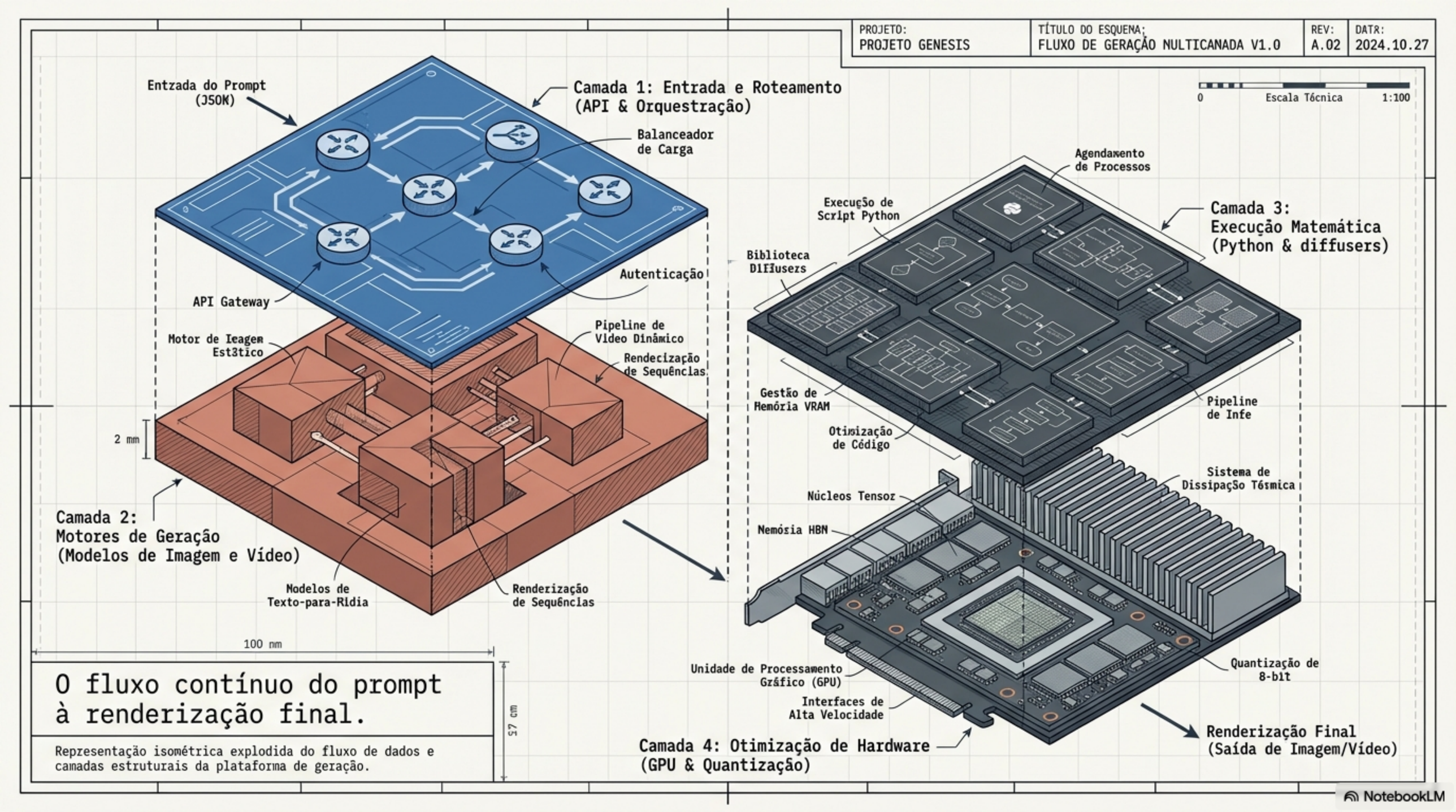
Do design de pipelines e tensores à orquestração e quantização de GPUs.



ESPECIFICAÇÕES DO SISTEMA MULTIMODAL (ESCALA RACK)

PARÂMETRO	DESCRIÇÃO	VALOR / TIPO
UNIDADES DE PROCESSAMENTO	ARRAY DE GPUs TENSOR	8X H100 NVL
QUANTIZAÇÃO	PRECISÃO MISTA DINÂMICA	FP8, FP16, BF16
ORQUESTRAÇÃO	FRAMEWORK DISTRIBUÍDO	KUBERNETES + NCCL
PIPELINE DE DADOS	INGESTÃO DE ALTA VAZÃO	200 GBE RDMA
ARMAZENAMENTO	BUFFER DE MODELO	512 GB HBM3 / GPU

PROJETO:	TÍTULO DO ESQUEMA:	REV:	DATA:
PROJETO GENESIS	FLUXO DE GERAÇÃO NULTICANADA V1.0	A.02	2024.10.27



O fluxo contínuo do prompt à renderização final.

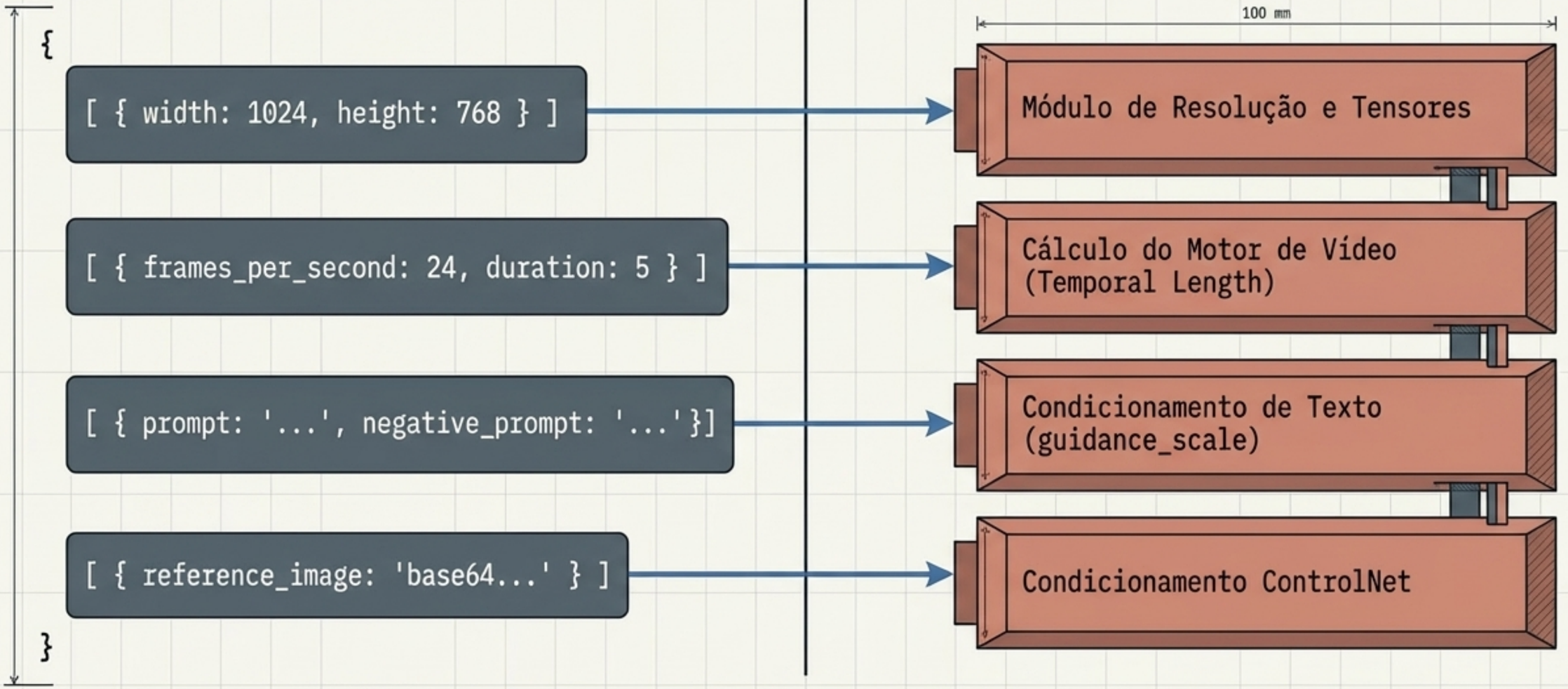
Representação isométrica explodida do fluxo de dados e camadas estruturais da plataforma de geração.

PROJETO:		REV:	DATA:
PROJETO GENESIS		A.03	2024.10.27

A Ponte API-Backend

Dados estruturados roteando a intenção criativa diretamente para os pesos da rede neural.

0 Escala Técnica 1:100



PROJETO:	TÍTULO DO ESQUEMA:	REV:	DATA:
PROJETO GENESIS	A Mecânica dos Prompts e Guidance Scale	A.04	2024.10.27

A Mecânica dos Prompts e Guidance Scale

Escala Técnica 0 — 1:100

Entrada de Texto

Vetorização

Prompt Positivo

Direciona o vetor da IA em direção ao conceito desejado (ex: estilo cyberpunk, hiper-realista, iluminação dramática).

Prompt Negativo

Instrui o modelo sobre quais subespaços latentes evitar (ex: desfocado, baixa resolução, anatomia deformada).

Ajusta a força matemática com que a IA segue as diretrizes do prompt contra sua própria distribuição latente.

Escala de Intensidade

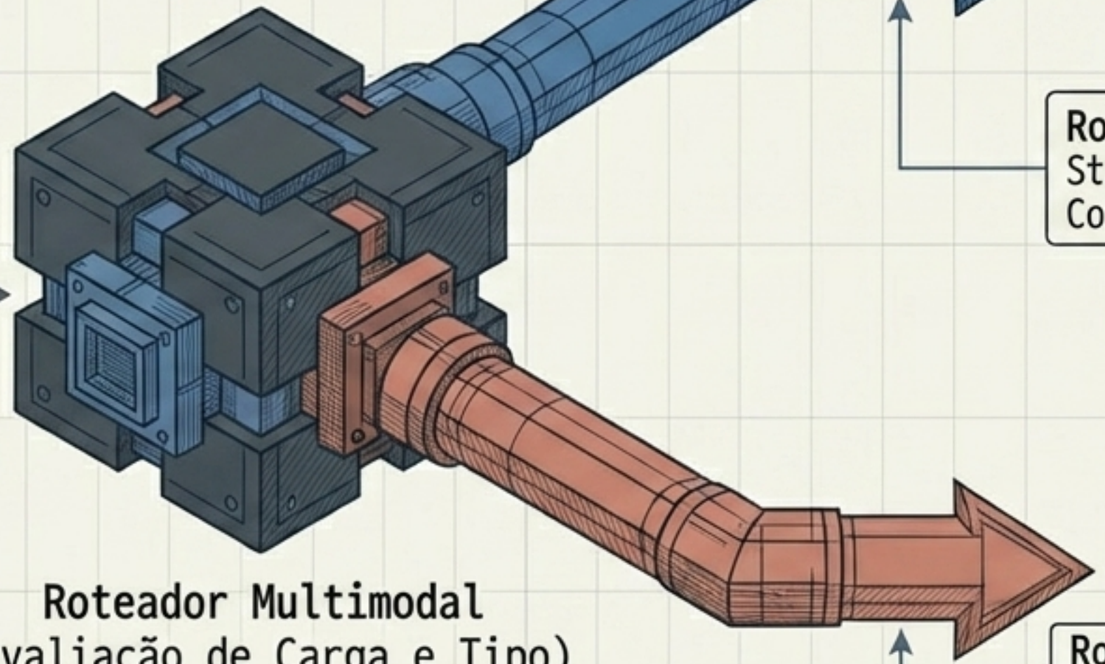
0 — 1:100

PROJETO:	TÍTULO DO ESQUEMA:	REV:	DATA:
PROJETO GENESIS	O Ecossistema da API: Roteamento	A.05	2024.10.28

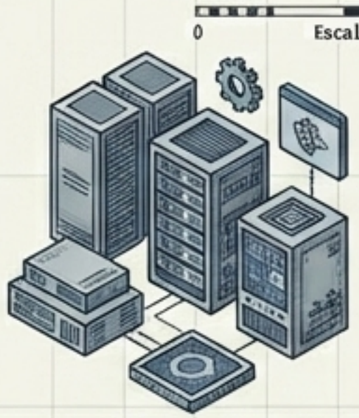
O Ecossistema da API: Roteamento Inteligente

Balanceando custo computacional e controle criativo.

Usuário
(Mobile/Web)

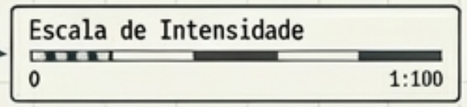


Roteador Multimodal
(Avaliação de Carga e Tipo)



Rota A: Geração de Imagem
Stable Diffusion + ControlNet
Controle Granular & Open-Source

Rota B: Geração de Vídeo
APIs Proprietárias
RunwayML, Pika Labs, Google Veo



O Fluxo de Pipeline Python

A biblioteca diffusers da Hugging Face atua como a camada de abstração, empilhando modelos de base e condicionadores de forma modular.

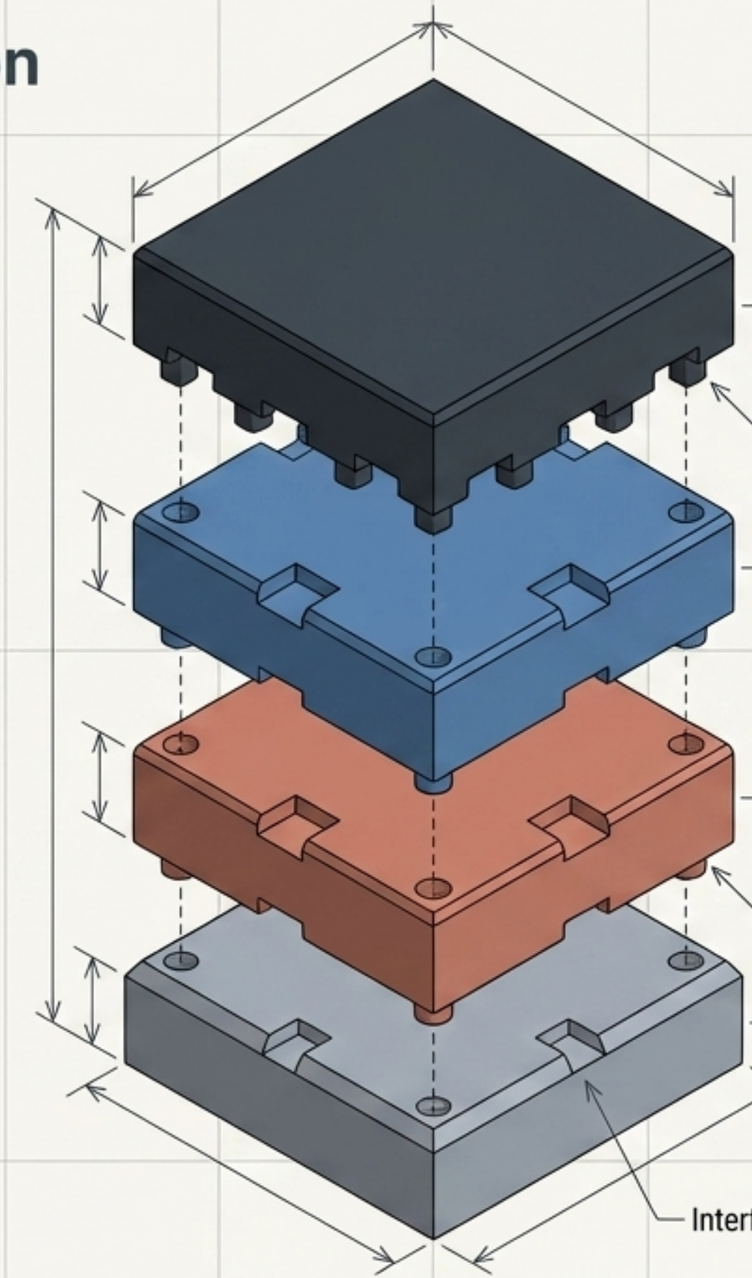
PROJETO:
PROTETO GENESIS

TÍTULO DO ESQUEMA:
O Fluxo de Pipeline Python

REV:
A.86

DATA:
2024.10.29

Escala Técnica



1. StableDiffusionControlNetImg2ImgPipeline
Camada de Orquestração

Ponto de Encaixe da API

2. Pesos do Stable Diffusion Base
O Motor de Geração

3. Pesos do ControlNet
O Condicionamento Visual

Ponto de Encaixe da API

4. PyTorch / CUDA
A Base de Execução

Interface de Execução de Tensor

0 2 6 10

A Equação ControlNet

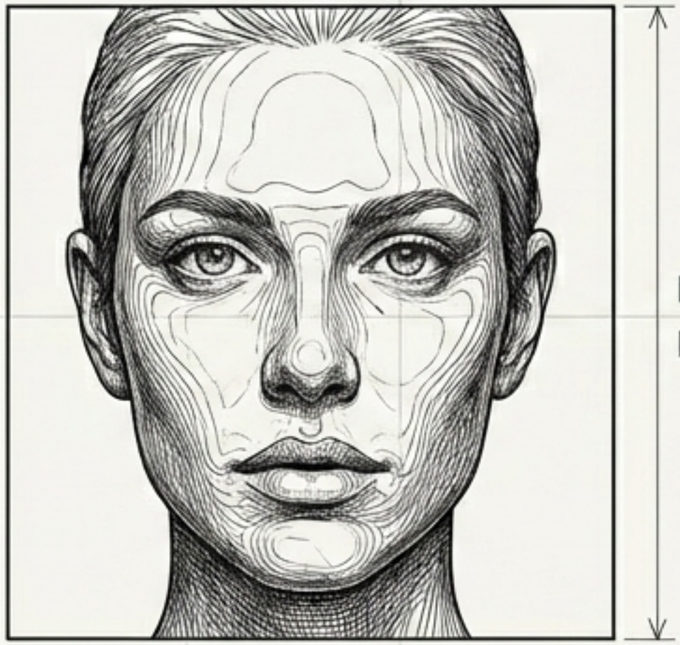
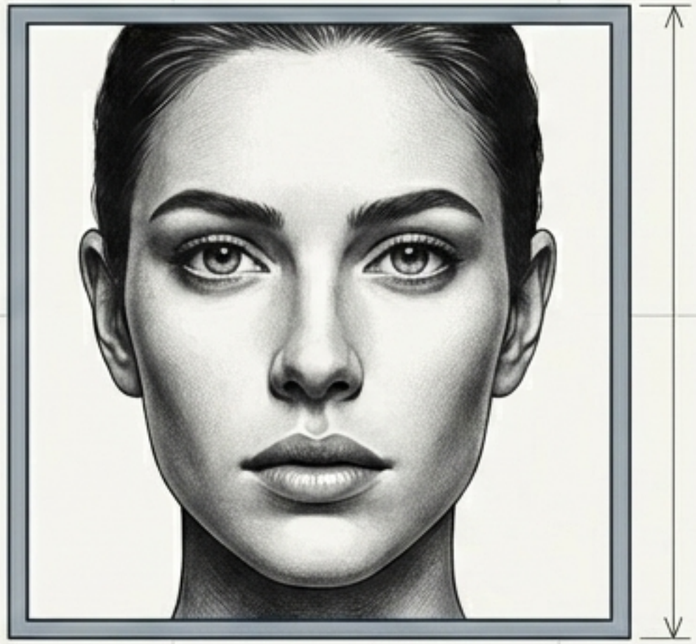
Forçando o modelo principal a respeitar estritamente a estrutura, composição ou pose de uma entrada base.

PROJETO: PROTETO GENESIS

TÍTULO DO ESBOÇO: A Equação ControlNet

REV: A.07

DATA: 2024.10.30

[Prompt de Texto] +  = 

[Imagem de Referência]

[Saída Altamente Fiel]

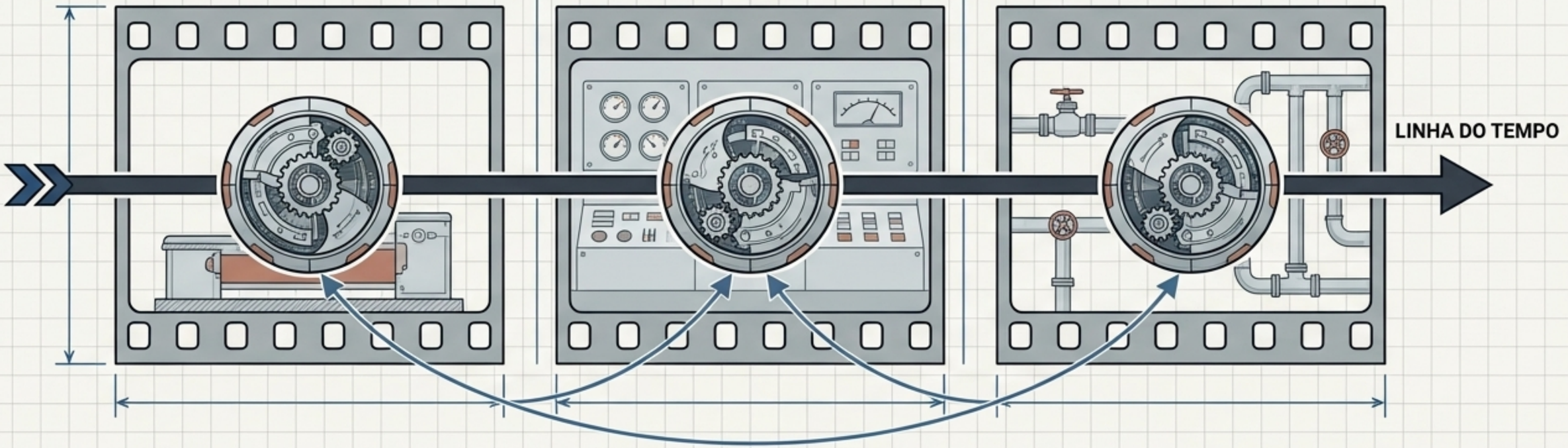
ESCALA TÉCNICA 0 1:100

REV: A.07

DATA: 2024.10.30

A Anatomia da Coerência Temporal

O verdadeiro desafio do vídeo: distinguir 10 quadros contínuos de 10 imagens isoladas.



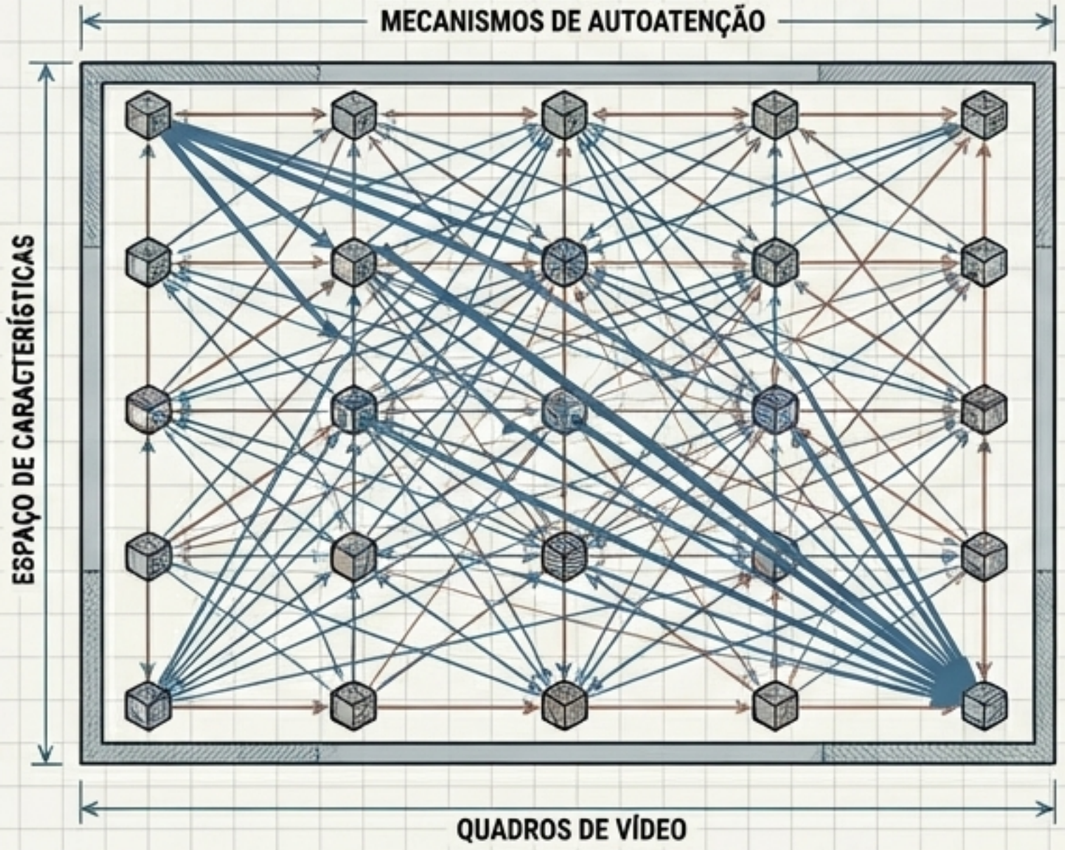
PROJETO: GERAÇÃO DE VÍDEO AVANÇADA		
TÍTULO DO ESQUEMA: A Anatomia da Coerência Temporal		
REV: A.08	DATA: 2024.11.01	ESCALA TÉCNICA 0 2 4 6 10

Arquiteturas de Geração de Vídeo

O debate estrutural no processamento de imagens em movimento.

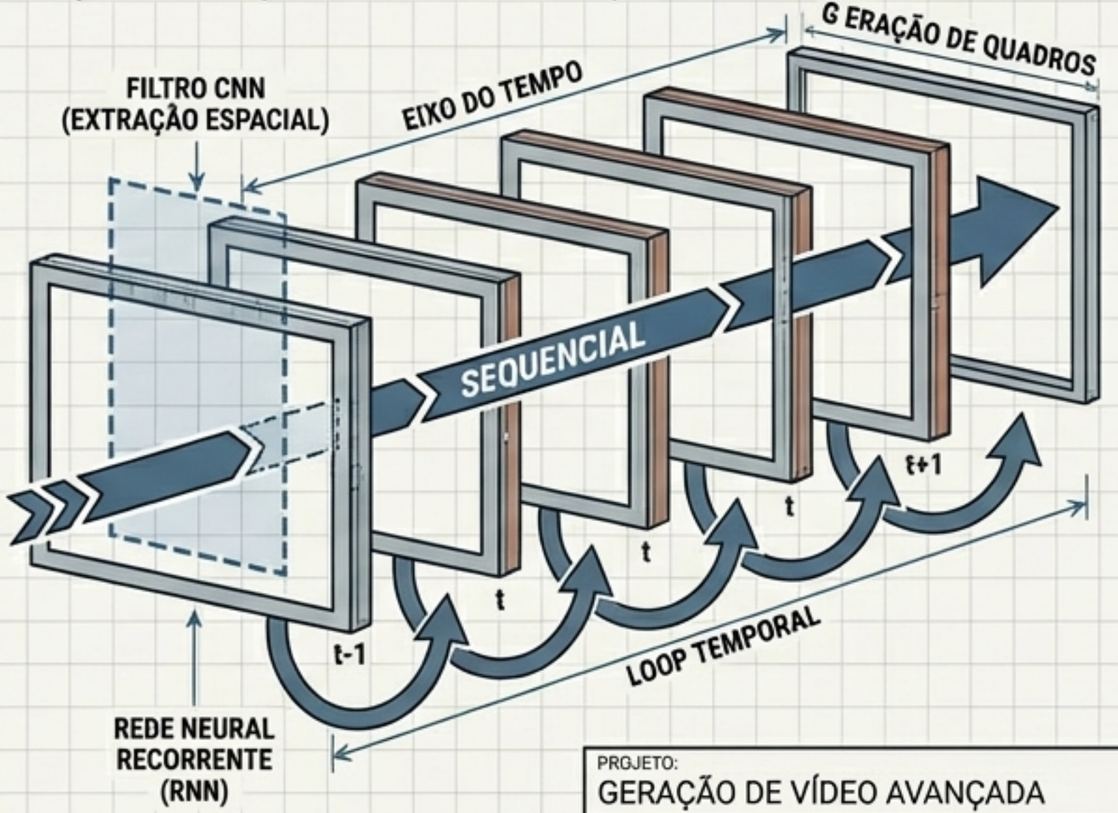
Modelos Baseados em Transformers

Compreensão holística e de longo alcance temporal. Processamento via mecanismos de autoatenção em todos os quadros simultaneamente.



Abordagem Híbrida (CNN + RNN)

Extração de características espaciais combinada com geração sequencial de quadros no eixo do tempo.



PROJETO: GERAÇÃO DE VÍDEO AVANÇADA		
TÍTULO DO ESQUEMA: Arquiteturas de Geração de Vídeo		
REV: A.09	DATA: 2024.11.02	ESCALA TÉCNICA 0 2

Matriz de Geração: Imagem vs. Vídeo

Dimensão	Motor de Imagem	Motor de Vídeo
Complexidade Principal	**Coerência Espacial**	**Coerência Temporal & Física**
Arquitetura Dominante	**Difusão Padrão (Stable Diffusion)**	**Transformers de Vídeo / CNN+RNN**
Custo Computacional	**Baixo/Moderado (Frame único)**	Exponencial (Atenção cross-frame)
Controle de Saída	**Alto (ControlNet)**	**Moderado** (Recursos limitados de movimento)

PRGJETO:
GERAÇÃO DE VÍDEO AVANÇADA

TITULD DO ESQUEMA:
Matriz de Geração: Imagem vs. Vídeo

REV:	DATA:	ESCALA TÉCNICA
A.10	2024.11.03	0 2 4 6 10

Matriz de Decisão Arquitetônica

Open-Source vs. APIs Proprietárias

Stack Customizada (Open-Source)

- **Tecnologia:** Hugging Face, Stable Diffusion
- **Controle Granular:** Máximo (Pesos customizados, ControlNet)
- **Custo de Manutenção:** Alto (Infraestrutura GPU, pipeline)
- **Curva de Implementação:** Íngreme

APIs Hospedadas (Proprietárias)

- **Tecnologia:** RunwayML, Pika Labs, Google Veo
- **Controle Granular:** Limitado pelos endpoints
- **Custo de Manutenção:** Baixo (Cobrança via API)
- **Curva de Implementação:** Rápida

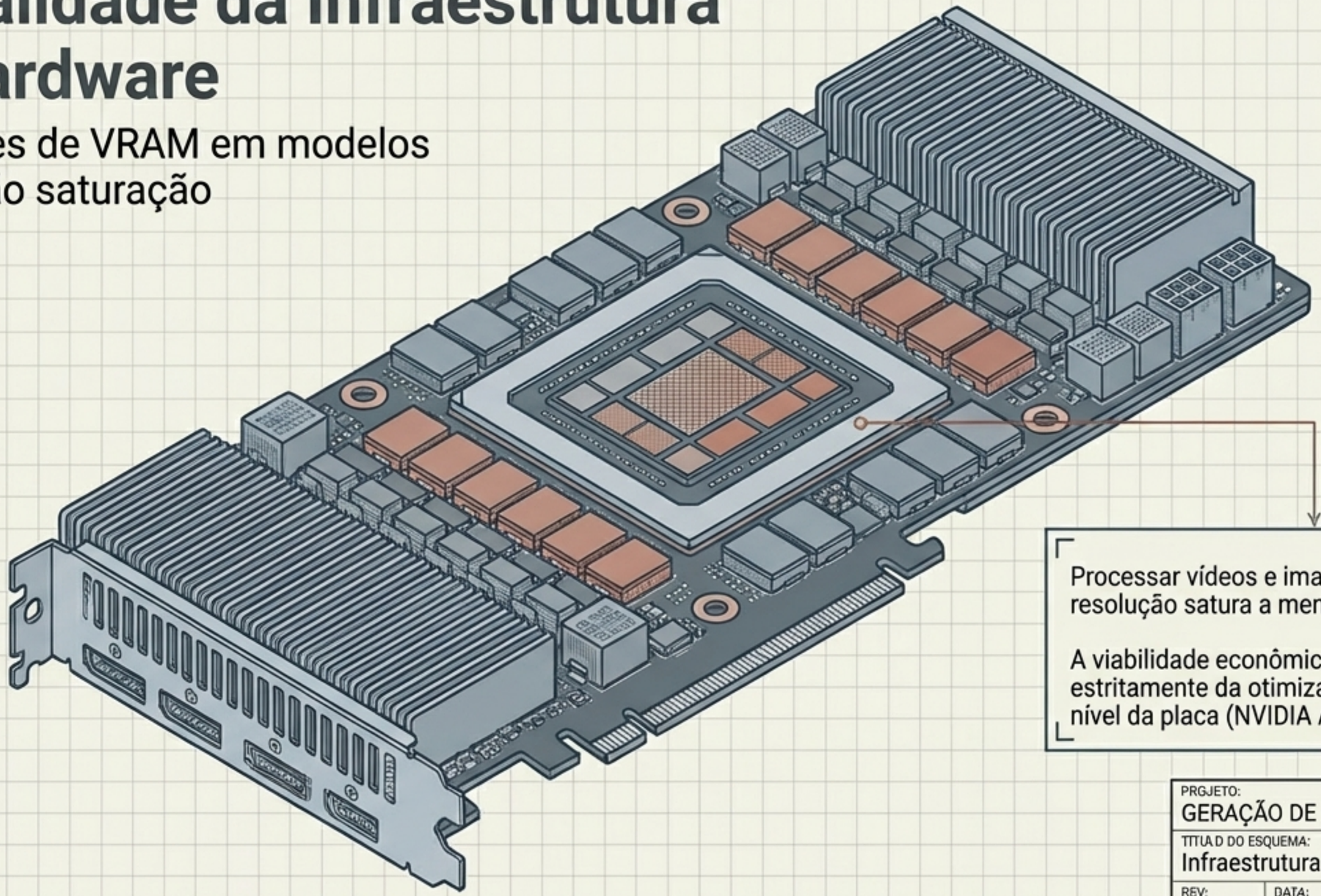
PRGJETO:
GERAÇÃO DE VÍDEO AVANÇADA

TITULO DO ESQUEMA:
Matriz de Decisão Arquitetônica

REV:	DATA:	ESCALA TÉCNICA
A.11	2024.11.04	0 2 4 6 10

A Realidade da Infraestrutura de Hardware

Restrições de VRAM em modelos de difusão saturação



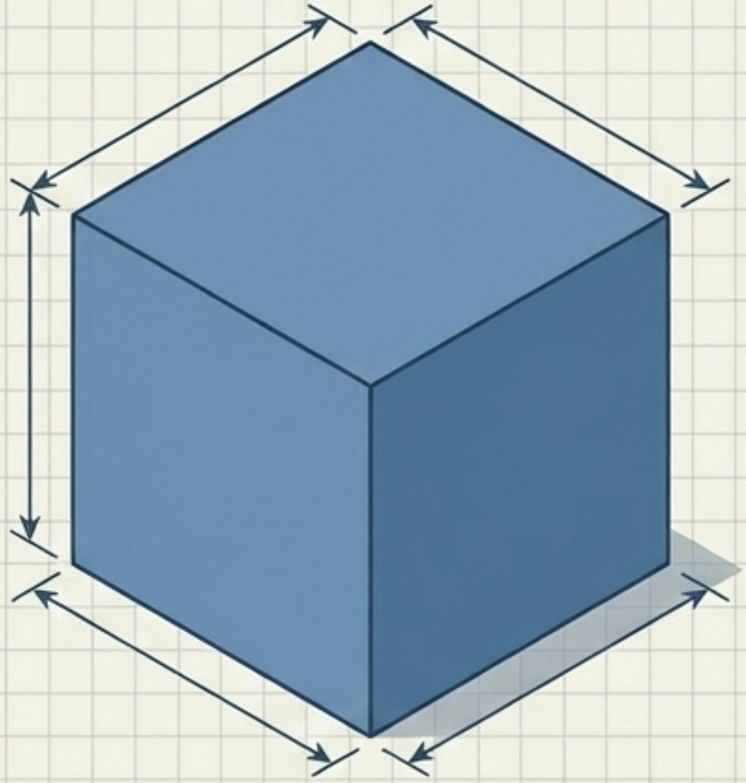
Processar vídeos e imagens em alta resolução satura a memória rapidamente.

A viabilidade econômica do software depende estritamente da otimização matemática no nível da placa (NVIDIA A100 / H100).

PRGJETO: GERAÇÃO DE VÍDEO AVANÇADA		
TITULO DO ESQUEMA: Infraestrutura e VRAM		
REV: A.12	DATA: 2024.11.05	ESCALA TÉCNICA 0 2 4 6 10

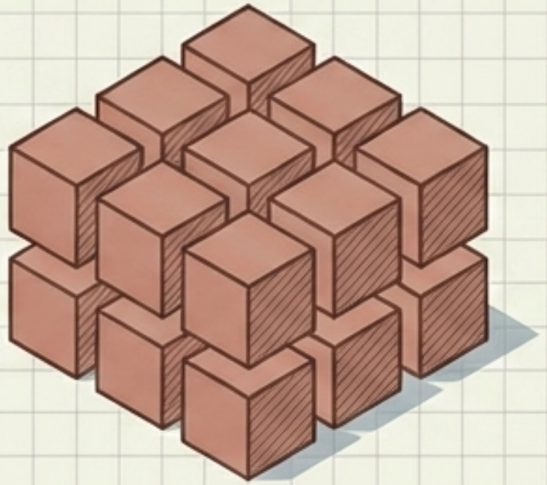
Otimização via Quantização

Implementação via biblioteca BitsAndBytes: comprimindo parâmetros para reduzir o gargalo de memória sem sacrificar a coerência visual.



Precisão 32-bit
(Antes)

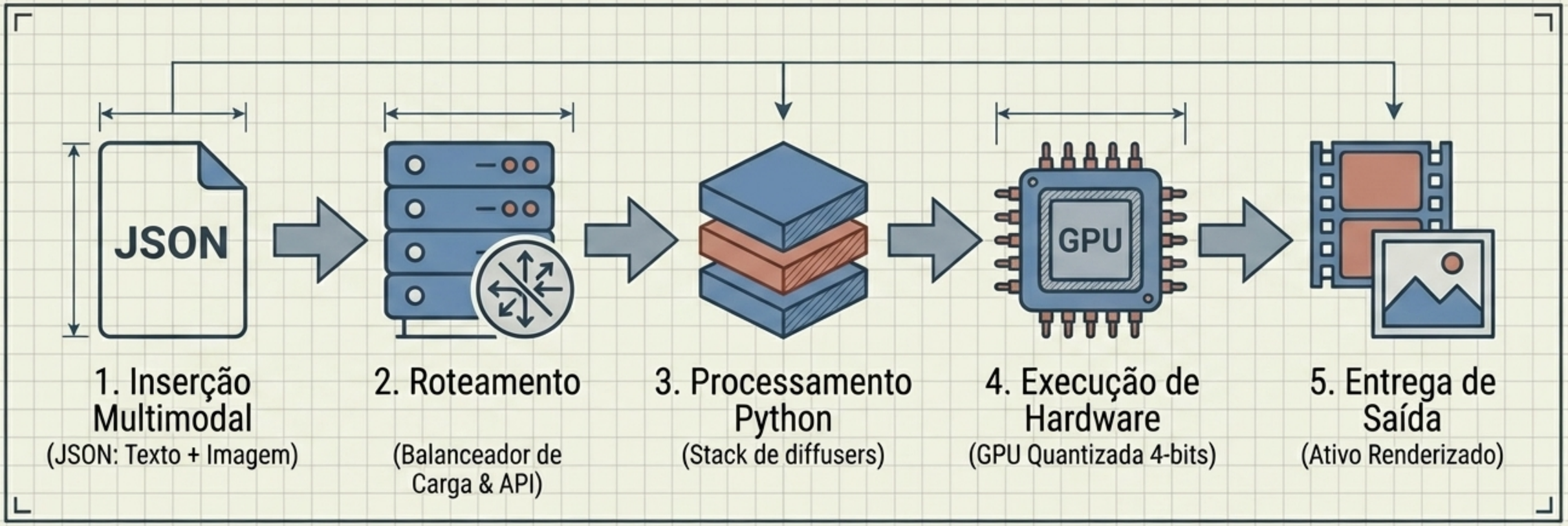
`load_in_4bit=True`



Quantização 4-bit
(Depois)

PROJETO: GERAÇÃO DE VÍDEO AVANÇADA		
TÍTULO DO ESQUEMA: Otimização via Quantização		
REV: A.13	DATA: 2024.11.06	ESCALA TÉCNICA 0 2 4 6 10

Arquitetura Multimodal Unificada (End-to-End)



PRGJETO: GERAÇÃO DE VÍDEO AVANÇADA		
TITLA D DO ESQUEMA: Arquitetura Multimodal		
REV: A.14	DATA: 2024.11.07	ESCALA TÉCNICA 0 2 4 6 10

O Futuro em Código e Tensores

A verdadeira vantagem competitiva em IA multimodal não reside apenas nos modelos que você escolhe, mas na precisão com que você orquestra a coerência, otimiza o silício e unifica o pipeline.

> [END_OF_TRANSMISSION]