

gpu01

Runtime Agent <http://localhost:8090>

Backends

| Backend            | Kind              | Model                        | Runtime  | Status  | Control |
|--------------------|-------------------|------------------------------|----------|---------|---------|
| text-gpu           | text_generation   | n/a                          | llmapp   | running |         |
| embedding-gpu      | embedding         | n/a                          | llmapp   | running |         |
| describe-gpu       | image_description | Qwen/Qwen2.5-72B-Instruct    | nl       | stopped |         |
| image-sd-gpu       | image_generation  | runway/stable-diffusion-v1.5 | sd       | stopped |         |
| image-sd-turbo-gpu | image_generation  | stabilityai/sd-turbo         | sd-turbo | running |         |

# Inference Broker for GPU-backed Stable Diffusion

## Description

GPU-backed Stable Diffusion 1.5 Turbo API for inference. This service provides a REST API for generating images using Stable Diffusion 1.5 Turbo on GPU. It supports text-to-image generation and image-to-image generation. The service is designed to be scalable and easy to integrate into existing applications.

API Endpoints:

- POST /generate: Generate an image from a text prompt.
- POST /refine: Refine an existing image using a text prompt.

Backends:

- sd-turbo: Stable Diffusion 1.5 Turbo (Running)
- sd: Stable Diffusion 1.5 (Stopped)

Backends

| Backend          | Kind              | Model                           | Runtime        | Status  | Control                             |
|------------------|-------------------|---------------------------------|----------------|---------|-------------------------------------|
| text-main        | text_generation   | n/a                             | n/a            | n/a     | n/a                                 |
| embedding-main   | embedding         | n/a                             | n/a            | n/a     | n/a                                 |
| whisper-main     | transcription     | whisper-v3-turbo-FLM            | lemonade       | running | <input checked="" type="checkbox"/> |
| image-sd-turbo   | image_generation  | SD-Turbo                        | lemonade       | running | <input checked="" type="checkbox"/> |
| image-sd15-npu   | image_generation  | amd/stable-diffusion-1.5-amdnpu | sd-npu-backend | running | <input checked="" type="checkbox"/> |
| describe-qwen3vl | image_description | qwen3vl-it-4b-FLM               | lemonade       | running | <input checked="" type="checkbox"/> |

gpu01

Runtime Agent    http://zikuu-pc-01:8090

Backends

| Backend            | Kind              | Model                          | Runtime  | Status  | Control                             |
|--------------------|-------------------|--------------------------------|----------|---------|-------------------------------------|
| text-gpu           | text_generation   | n/a                            | llamacpp | running | <input checked="" type="checkbox"/> |
| embedding-gpu      | embedding         | n/a                            | llamacpp | running | <input checked="" type="checkbox"/> |
| describe-gpu       | image_description | Qwen/Qwen3-VL-2B-Instruct      | ll       | stopped | <input type="checkbox"/>            |
| image-sd-gpu       | image_generation  | runwayml/stable-diffusion-v1-5 | sd       | stopped | <input type="checkbox"/>            |
| image-sd-turbo-gpu | image_generation  | stabilityai/sd-turbo           | sd-turbo | running | <input checked="" type="checkbox"/> |

ã•?ã??ã•šã?•ç?»å?•ç??æ?•ã•«ã½¿ã•?ã??æ?`è«?è³?æ°•ã•ã?

- NPUã??ã½¿ã•?SD1.5
- AMDã•®SD Turbo
- GPUã??ã½¿ã•?SD1.5
- GPUã??ã½¿ã•?SD Turbo

ã•®4ã•°ã•«ã•ã??ã•¾ã•?ã•?ã??

NPUç??SD1.5

Image Generation

Backend

image-sd15-npu

Prompt

a wooden workshop with hand tools on a bench

Negative Prompt

blurry, low quality

Width

backend default

Height

backend default

Steps

backend default

Guidance Scale

backend default

Seed

random

Generate

Response

backend  
image-sd15-npu

Model  
amd/stable-diffusion-1.5-amdnpu

Generated Image



AMD??SD Turbo

Page 3

Footer Tagline

Image Generation

Backend

image-sd-turbo

Prompt

a wooden workshop with hand tools on a bench

Negative Prompt

blurry, low quality

Width

backend default

Height

backend default

Steps

backend default

Guidance Scale

backend default

Seed

random

Generate

Response

|                                   |                      |
|-----------------------------------|----------------------|
| Backend <div>image-sd-turbo</div> | Model <div>n/a</div> |
|-----------------------------------|----------------------|

Generated Image



GPUç??SD1.5

Image Generation

Backend

image-sd-gpu

Prompt

a wooden workshop with hand tools on a bench

Negative Prompt

blurry, low quality

Width

backend default

Height

backend default

Steps

backend default

Guidance Scale

backend default

Seed

random

Generate

Response

backend image-sd-gpu

Model runwayml/stable-diffusion-v1-5

Generated Image



GPUç??SD Turbo

Image Generation

Backend

image-sd-turbo-gpu

Prompt

a wooden workshop with hand tools on a bench

Negative Prompt

Width

backend default

Height

backend default

Steps

backend default

Guidance Scale

backend default

Seed

random

Generate

Response

Backend

image-sd-turbo-gpu

Model

stabilityai/sd-turbo

Generated Image



ã•?ã??ã•§[Vision Broker](#)ã•®æ©?è?½ã??ã•?ã•¹ã•|å•,å•?ã•§ã•ã•¾ã•?ã•?ã??

Category

1. æ?¥ã? ã•®æ´»å??

Page 6

Footer Tagline

Tags

- 1. AI
- 2. Inference Broker
- 3. LLM
- 4. Stable Diffusion
- 5. ä,?é??å??å,?
- 6. å±±æø"ç??

Date Created  
2026å¹9æ??1æ?¥

Author  
kazuo-tsubaki

default watermark