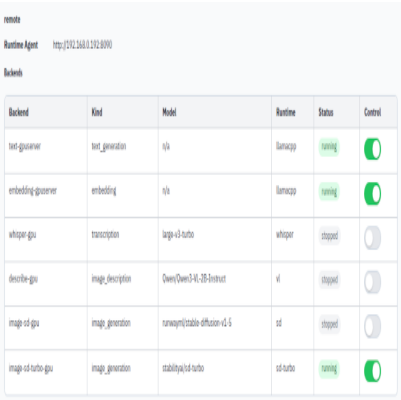


Inference Broker â?? ã?ªã?½ã?¼ã?¹ã?¹ã?±ã?¸ã?¥ã?¼ã?©ã?¼ã??è¿½å?

Description

default watermark

Inference Brokerã•¯ã?ã••ã•®ã?•ã??ã?¯ã?¨ã?³ã??ã«

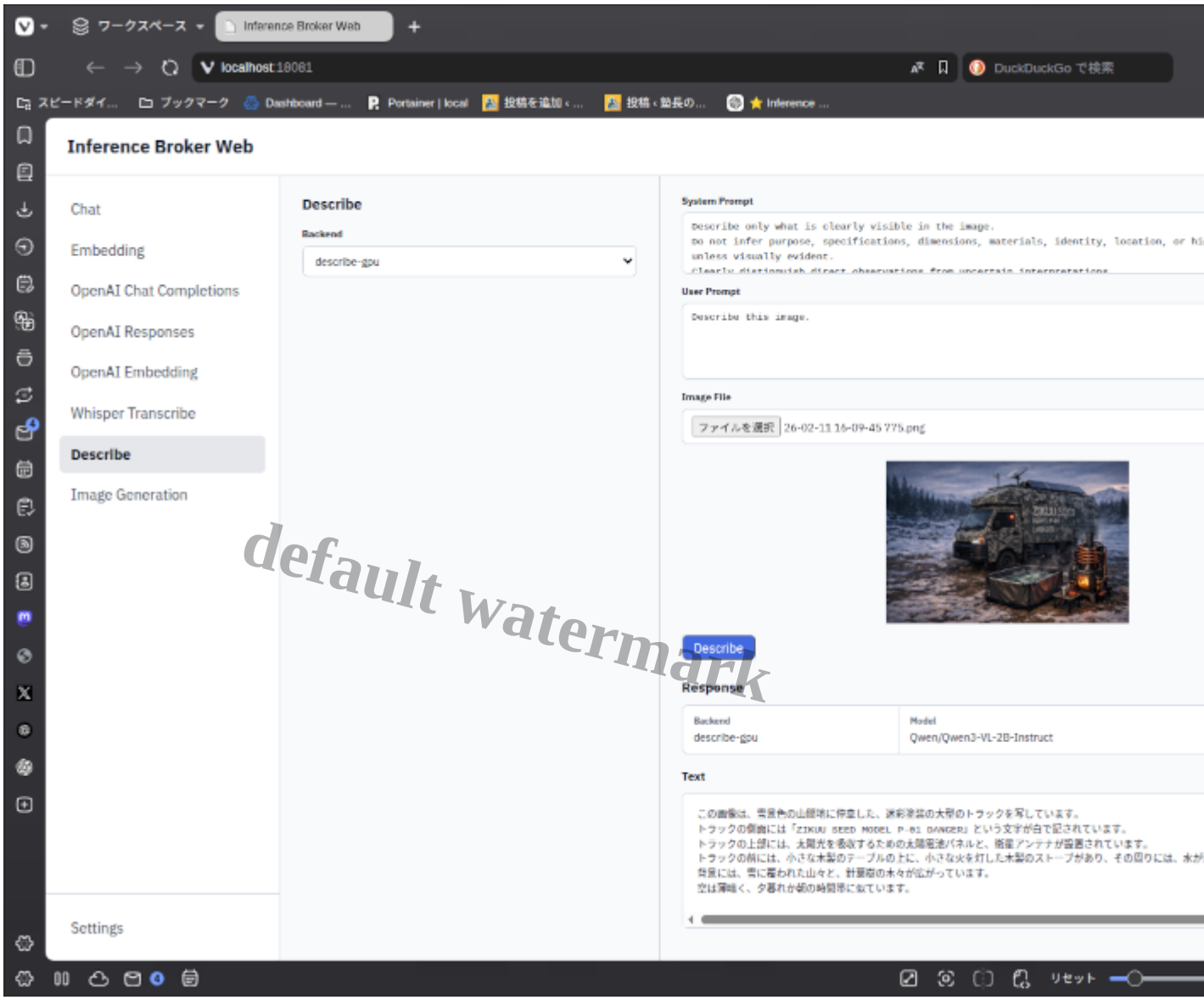
- **llama.cpp**ã?µã?¼ã?•ã?¼ï¼?æ??ç« ç??æ?•ã•¨å??ã?•è¾¼ã•¿ç??æ?•ï¼?
- **Lemonade**ã?µã?¼ã?•ã?¼ï¼?NPUæ•è¼?ã•®ZIKUU Miniã•®å´å•?ï¼?
- **SD NPU Backend**ï¼?NPUæ•è¼?ã•®ZIKUU Miniã•®å´å•?ã?•ZIKUUå? è£½ï¼?
- **GPU Inference**ï¼?NVIDIA GPUæ•è¼?ã?µã?¼ã?•ã?¼ã?•ZIKUUå? è£½ï¼?
  - **Vision**ï¼?ç?»å?•ã•®èª¬æ??ï¼?
  - **Stable Diffusion**ï¼?ç?»å?•ç??æ?•ï¼?
  - **Whisper**ï¼?é?³å£°ã•®æ??å?èµ·ã•?ã•?ï¼?

ã•?å??ã•?ã•¦ã•?ã•¦ã?•æ??ç« ç??æ?•ã?•å??ã?•è¾¼ã•¿ç??æ?•ã?•ç?»å?•ç??æ?•ã?•æ??å?èµ·ã•?ã•?ã?•ç?»å?•èª¬æ??ã??è¡?ã•?APIã??æ?•ã•£ã•¦ã•?ã•¾ã•?ã??
æ??ç« ç??æ?•ã•«ã•¤ã•?ã•¦ã•¯ã?Open AI ãº?æ•?APIã•«ã•ªã•£ã•¦ã•?ã•¦ã?•/chat/completionsã•¨/responsesã•®2ç¨®é¡?ã•?ç?¨æ?•ã•?ã??ã•¦ã•?ã??ã?•ã?µã?¼ã??ã??ã?¼ã??ã?£ã?¼ã•®AIã??ã?¼ã?«ã•¨æ?¥ç¶?ã•§ã••ã??ã??ã•?ã•«ã•ªã•£ã•¦ã•?ã•¾ã•?ã??

Inference Runtime Agentã•¨ã•?ã•?ã??ã?¼ã?¢ã?³ã•?ã?•ã•?ã??ã??ã•®ã?•ã??ã?¯ã?¨ã?³ã??ã•®ã?¹ã??ã?¼ã?¿ã?¹ç?£è¦?ã•èµ·å??ã?•å•?æ¢ã??è¡?ã•?ä»?çµ?ã•¿ã•?ã•?ã•£ã•¦ã?•Inference Brokerã•¯ã?ã••ã??ã??ä½¿ã•£ã•¦è¨?ç®?è³?æº•ã??ç®¡ç•?ã•?ã•¾ã•?ã??

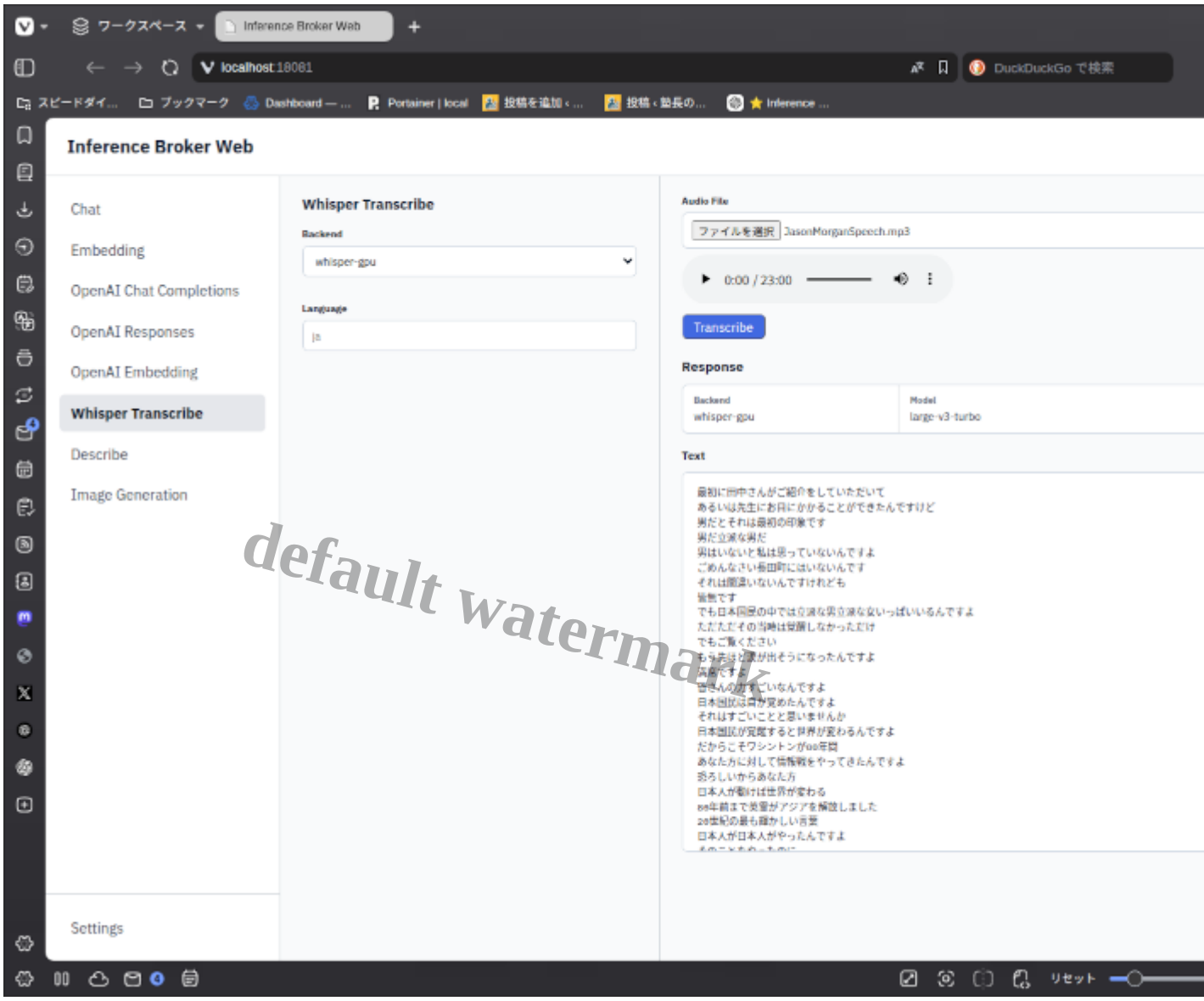
ä»?å??è¿½å? ã•?ã•?ã?¹ã?±ã?¸ã?¥ã?¼ã?©ã?¼ã•¯ã?GPU Inferenceã•?æ?•ã•£ã•¦ã•?ã??ã?•Whisper/VL/SD/SD-Turboã??ã?¢ã??ã?ªã?±ã?¼ã?·ã?§ã?³ã•®ã?ªã?¯ã?¨ã?¹ã??ã«å¿?ã•?ã•¦è?ªå??ç??ã•«å??ã??æ?¿ã•?ã??æ©?è?½ã•§ã•?ã??
ã••ã•®é??ã•«ã?•VRAMä¸•è¶³ã•«ã•ªã??ã•®ã•¯å?°ã??ã•®ã•§ã?•å¿ è¦•ã•«å¿?ã•?ã•¦ã?•VRAMã••è¼?ã•£ã•¦ã•?ã??ã?¢ã??ã?«ã??ã?¢ã?³ã?ã?¼ã??ã•?ã•¦ã?•è¦•æ±?ã•?ã??ã•?ã?¢ã??ã?«ã??ã?ã?¼ã??ã•?ã??ã?•ã•?ã•§ã•?ã??

VLã??ä½¿ã•£ã•¦ç?»å?•ã•®èª¬æ??ã??ã•?ã•?ã•?ã??
ã?•ã??ã?¯ã?¨ã?³ã??ã?èµ·å??ã•?ã•¦ã•?ã•ªã•?ã??ã•°èµ·å??ã•?ã•¦ã•?ã??æ?¨è«?ã??ã•?ã•¾ã•?ã??



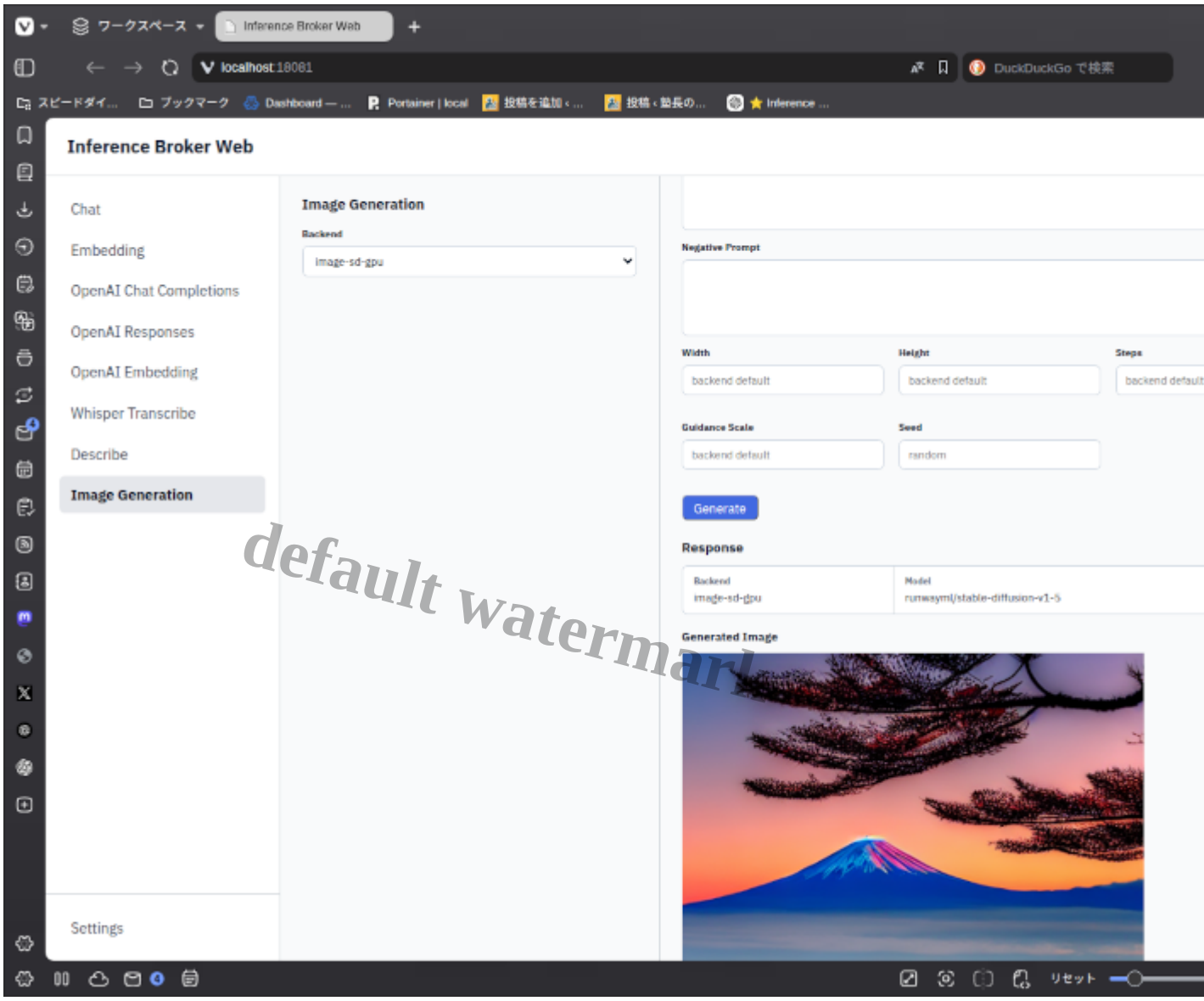
ã••ã•®ã•¾ã•¾ã?•æ??å?èµ·ã•?ã•?ã??å®?è¡?ã•?ã??ã??

VLã?¢ã??ã?«ã??VRAMã•?ã??ã?¢ã?³ã?ã?¼ã??ã•?ã•¡ã?•Whisperã?¢ã??ã?«ã??ã?ã?¼ã??ã•?ã•¦æ?¨è«?ã•?ã??ã??



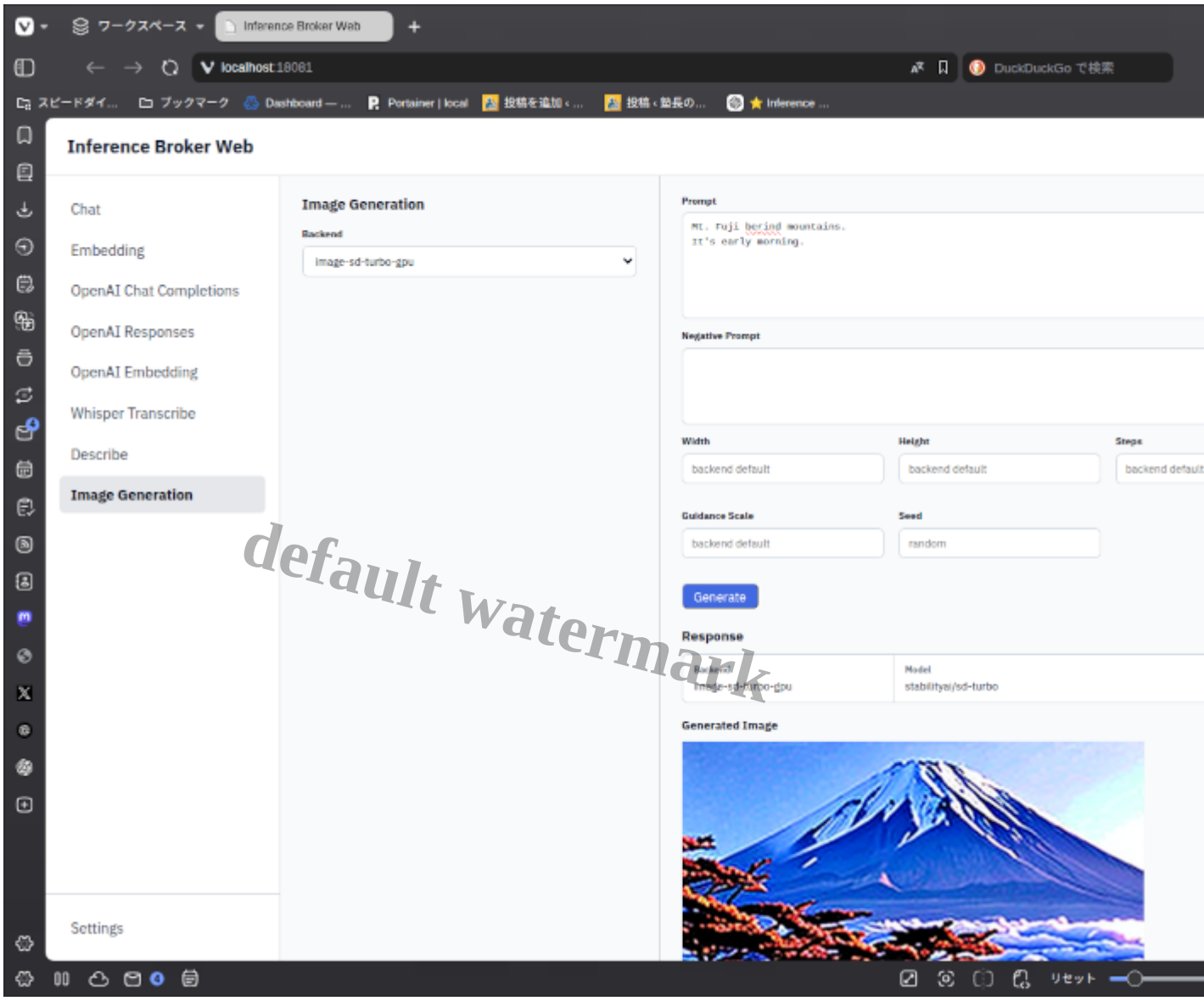
æ¬¡ã•«ã?•Stable Diffusion 1.5ã??ä½¿ã•£ã•?ç?»å?•ç??æ?•ã??ã•?ã•?ã??ã??

Whisperã?¢ã??ã?«ã??ã?¢ã?³ã?ã?¼ã??ã•?ã•¦ã?•SD 1.5ã??VRAMã•«è¼?ã•?ã•¦å?¦ç•?ã•?ã??ã??



ã•?ã??ã•«ã?•SD-Turboã??ä½¿ã•£ã•¦ç?»å?•ç??æ?•ã•?ã??ã??

SD 1.5ã??VRAMã•?ã??é?•ã?•ã•?ã•¦ã?•SD-Turboã??ã?ã?¼ã??ã•?ã•¦å?¦ç•?ã•?ã??ã??



è¤?æ?°ã•®GPUã?µã?¼ã?•ã?¼ã??ä¸¦ã•¹ã??ã•°ã?•å•³æ??å¿?ç?ã•?å¿ è¦•ã•ªã?¢ã??ã?«ã•¯VRAMã•«è¼?ã•?ã•?ã•¾ã•¾ã•«ã•?ã??ã?•ã?¢ã?¯ã?»ã?¹ã•®é »åº¦ã•?é«?ã•?ã•¨ã••ã??ã?µã?¼ã?•ã?¼é??å®³ã•?ã•?ã•£ã•?ã•¨ã••ã•¯ãº?å??ã•®GPUã?µã?¼ã?•ã?¼ã??ä½¿ã•?ã?•æ?¥ã•?ã•ªã•?å?¦ç•?ã??ã?•ã??ã?•å?¦ç•?ã•§ã•¯ã•å¿ è¦•ã•«å¿?ã•?ã•¦ã?¢ã??ã?«ã??VRAMã•«è¼?ã•?ã??ã?•ã•¨ã•?ã•?é•?ç?¨è¨è¨?ã•?ã•§ã••ã•¾ã•?ã??

æ¬¡ã•¯ã•å•?æ??ã•«ã?ªã?¯ã?¨ã?¹ã??ã•å±?ã•?ã•?ã•¨ã••ã•«ã?•ã?ã?¥ã?¼ã?¤ã?³ã?°ã•?ã??ä»?çµ?ã•¿ã??å ¥ã??ã??ã•?ã•¨æ?•ã•?ã•¾ã•?ã??

ZIKUU Miniã•®å ´å•?ã•¯ã•ã•?ã??ã??ã•®å?¦ç•?ã??è?ªå?•ã•®NPUã•¨ïGPUã??ä½¿ã•£ã•¦è¡?ã•?ã•¾ã•?ã??
ã•ªã•®ã•§ã?•GPUã?µã?¼ã?•ã?¼ã•?å¿?ã•?ã•?ã•¨ã••ã•¯ã•è?ªå?•ã•®æ?¨è«?è³?æº•ã•«å??ã??æ?¿ã•?ã??ã?•ç©ºã•?ã•¦ã•?ã??GPUã?µã?¼ã?•ã?¼ã??ä½¿ã•?ã•¨ã•?ã•?ã??ã?©ã?¼ã?«ã?•ã??ã?¯æ©?è?½ã??å ¥ã??ã•?ã•?ã•¨æ?•ã•£ã•¦ã•?ã•¾ã•?ã??

ã•?ã??ã•§ã?•ã• ã•?ã•¶æ?¨è«?ã?¤ã?³ã??ã?©ã•£ã•½ã••ã•ªã•£ã•¦ã••ã•¾ã•?ã•?ã??

ZIKUUã•¨ã•?ã•?ã•®ã•¯ã?

- ã?¢ã??ã•¥ã••ã??ã•®ã?¤ã?³ã??ã?©ã?»ã?»ã?»å·¥æ?¿ã?•æ©?æ¢°ã?•é•?å ·ã?•æ??å°?ã•§ã•• ã??æ??è¡?è?
- ç?¥è?ã•®ã?¤ã?³ã??ã?©ã?»ã?»ã?»ã??ã?¼ã?¿ã??ã?¼ã?¹ã?•Wikiã?•ã??ã?ã?°ã?•Discordã?• ã?ªã?•ã?¸ã??ã?ª
- æ?•è??ã•®ã?¤ã?³ã??ã?©ã?»ã?»ã?»Pivot Serviceã?•Context Pipeline Serverã?•*PRE*ã?•*AIå¡¾é?·*
- æ?¨è«?ã•®ã?¤ã?³ã??ã?©ã?»ã?»ã?»Inference Brokerã?•Inference Runtime Agentã?•GPU Inference

ã•?ä¸?ä½?å??ã•?ã•?ã?³ã??ã?¥ã??ã??ã?£ã?¼ã?¤ã?³ã??ã?©ã•®å§¿ã??ã•?ã•¦ã•?ã•¾ã•?ã??

ã•¤ã•¾ã??ã?•

ç•¾å®?
â??
è¨?é?²
â??
æ??è??
â??
æ?•è??
â??
æ?¨è«?
â??
å?•ã•³ç•¾å®?ã•¸

default watermark

ã•¨ã•?ã•?å¾ªç?°ã•?ã•§ã••ã??ç?¶æ ?ã•§ã•?ã??

ä»?ã?•ã•?ã•?ã•?ã•¦æ?¸ã•?ã•¦ã•?ã??æ??ç« ï¼?è¨?é?²ï¼?ã??ã?•è?ªå??ç??ã••ã•?ã•®ã?«ã?¼ã??ã??å??ã??ã•¾ã•?ã??

Category

1. æ?¥ã? ã•®æ´»å??

Tags

1. AI
2. Inference Broker
3. LLM
4. ä¸?é??å??å¸?
5. å±±æ¢¨ç??

Date Created
2026å¹´9æ??28æ?¥
Author

kazuo-tsubaki

default watermark