```
envs:
  MODEL_NAME: Qwen/Qwen-VL-Chat-Int4
  MODEL_ARCH: Qwen-VL-Chat-Int4
  HF_HUB_ENABLE_HF_TRANSFER: True



resources:
    # accelerators: {L4:4, A100:4, A100:8, A100-80GB:2, A100-80GB:4, A100-80GB:8} ## Large models
  accelerators: [L4, A10g, A100, A100, A100-80GB, T4, M60] ## Small models
  # cpus: 32+
  memory: 32+
  # use_spot: True
  # disk_size: 512  # Ensure model checkpoints (~246GB) can fit.
  # disk_tier: best
  ports: 8080  # Expose to internet traffic.

service:
  readiness_probe:
    path: /v1/chat/completions
    post_data:
      model: $MODEL_NAME
      messages:
        - role: user
          content: Hello! What is your name?
      max_tokens: 1
  readiness_probe: /v1/models
  replica_policy:
    min_replicas: 1
    max_replicas: 10
    target_qps_per_replica: 2.5
    upscale_delay_seconds: 300
    downscale_delay_seconds: 1200

setup: |
  pip install hf_transfer


run: |
  # Serve With Docker
  docker run -d --runtime nvidia --gpus all \
      -v ~/.cache/huggingface:/root/.cache/huggingface \
      --env "HUGGING_FACE_HUB_TOKEN=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"\
      -p 8080:8080 \
      --ipc=host \
      openmmlab/lmdeploy:latest \
    lmdeploy serve api_server $MODEL_NAME --model-name $MODEL_ARCH --server-port 8080
```