

• LOCAL INFERENCE • ZERO TRAINING • NO API KEYS

Ask any image. Get answers, boxes and a report.

Vision AI Agent describes, counts, inspects and finds objects in any photo, using open-source vision models that run entirely on your own computer.

 **Control Deck**
Configure the vision backend

01 MODEL

Qwen2.5-VL 7B · ✓ on this Mac

memory ~17 GB
source local folder ✓
license Apache 2.0

02 HARDWARE

 **Apple Silicon (MPS)**
GPU acceleration active gpu

Image detail Balanced

03 RUNTIME

Load / Reload Model

● model ready

Fully local. No API keys, no cloud calls, no per request cost.

Chat & Analyze

Object Detection

Report

SOURCE IMAGE



LIVE FRAME

1280x720 px

QUICK ACTIONS

Describe the scene

Count objects

Safety / PPE check

Read any text

Spot anomalies

CONVERSATION · 2 MESSAGES

YOU

Describe everything happening in this image in detail, as if explaining it to someone who cannot see it.

VISION AI AGENT

A person's hand is holding a black and yellow screwdriver. The person is holding the screwdriver in their right hand, with their left hand slightly curled around for support. The background features a white ceiling adorned with hexagonal panels, and a red curtain is partially visible in the bottom right corner.

Ask anything about this image...



02 WHY IT EXISTS

Looking at photos is still manual work.

Teams take more pictures than ever, of sites, shelves, products and paperwork. Turning those pictures into answers keeps running into the same three walls.

01

Someone has to look at every photo

Site photos, shelf photos and product shots pile up. A person opens each one, decides what matters and types notes by hand. It is slow, and the result depends on who is looking.

02

Cloud vision sends your images away

Most image AI services upload every picture to a third party and charge for each request. That is hard to accept for private sites, client work or anything confidential.

03

Custom detectors need training first

A traditional object detector only finds what it was trained on. Every new thing to look for means collecting and labelling example images before it works at all.

Vision AI Agent removes all three.



It looks for you

Describe, count, inspect and report in one click.



It stays on your computer

No uploads, no API keys, no bill per request.



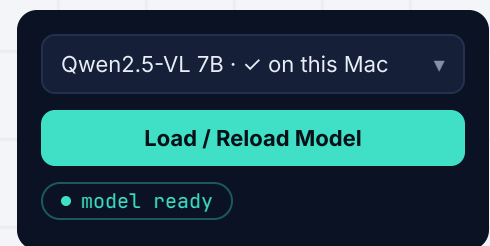
It finds what you name

No training. Type the object in plain English.

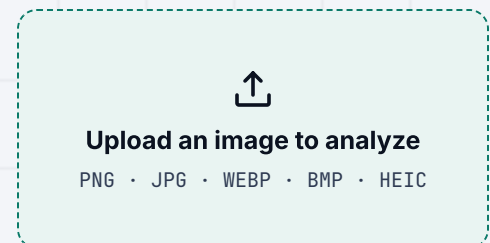
03 HOW IT WORKS

From photo to finished report in four steps.

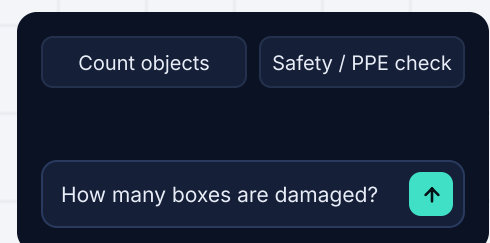
- 01 Load a model**
- Pick an engine in the Control Deck and press Load. It finds the model already on your computer, so nothing is downloaded twice.



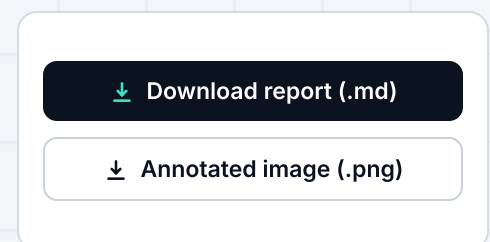
- 02 Upload an image**
- Drop in a photo from a site, a shelf, a production line or a phone. Phone photos are turned the right way up automatically.



- 03 Ask, inspect or detect**
- Chat with the image, run a one-click inspection, or type the name of an object to have every one of them boxed.



- 04 Export the report**
- Every action is logged as you go. Download the inspection report, and the image with its boxes drawn on.



04 CHAT WITH THE IMAGE

Ask anything. Then ask a follow-up.

Talk to the image the way you would to a colleague looking over your shoulder. The picture stays in context, so each question builds on the answers before it.

SOURCE IMAGE



LIVE FRAME

1280×720 px

RESOLUTION

1280×720

ORIENTATION

landscape

MEGAPIXELS

0.9 MP

DETAIL

Balanced

CONVERSATION · 3 MESSAGES

YOU

Describe everything happening in this image in detail.

VISION AI AGENT

A person's hand is holding a black and yellow screwdriver. The person is holding the screwdriver in their right hand, with their left hand slightly curled around for support. The background features a white ceiling adorned with hexagonal panels, and a red curtain is partially visible in the bottom right corner.

YOU

What kind of tip does the screwdriver have?

Analyzing image...

Ask anything about this image...

↑

Remembers the thread

Earlier questions and answers stay in context for every follow-up.

Choose the detail

Fast, Balanced or Detailed. More detail helps with small text and small parts.

Timed answers

Each reply shows how long it took, so you can pick the right model.

05 ONE-CLICK INSPECTIONS

Five inspections, one click each.

The questions people ask most are built in. Each button sends the model a carefully written instruction, and the answer goes straight into the report.



Describe the scene

A full written description, as if for someone who cannot see the image.

`"Describe everything happening in this image in detail..."`



Count objects

Every distinct object, person and item, listed by category with a count.

`"Count every distinct object, person, and item visible..."`



Safety / PPE check

Each person checked for helmet, vest, gloves and goggles, with violations flagged.

`"Act as a workplace safety inspector..."`



Read any text

Signs, labels and numbers transcribed exactly as they are written.

`"Read and transcribe any visible text, signage, labels..."`



Spot anomalies

Anything damaged, defective, out of place or unusual, pointed out.

`"Point out anything that looks damaged, defective..."`



Or ask your own

Type any question in plain words and follow up as often as you like. It all lands in the same report.

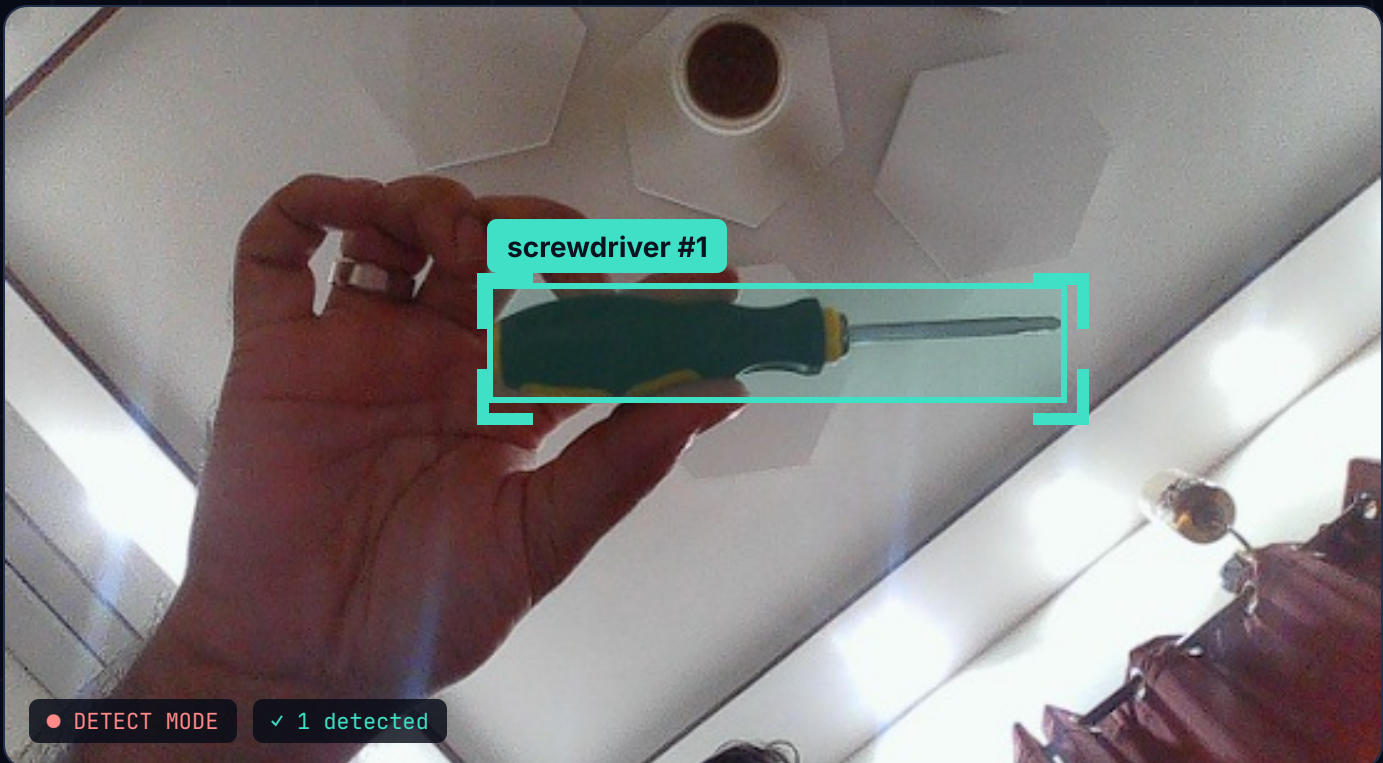
Is the fire exit blocked?



06 ZERO-SHOT DETECTION

Name it. It finds it.

Type any object in plain English, from "helmet" to "damaged box". Every match is boxed and labelled on the photo, with no training and no example images.



● DETECT MODE

✓ 1 detected

WHAT SHOULD THE AGENT LOOK FOR?

screwdriver

person

helmet

forklift

fire extinguisher

damaged box

Run Detection

INSTANCES FOUND

boxes drawn on the frame

1

■ screwdriver #1 (458, 273) → (976, 376)

[Download annotated image](#)

Every finding, written up for you.

Each quick action, detection and question is logged as you work. When you are done, the report is already written, ready to hand to a client or file with the job.

Vision AI Agent — Inspection Report

Generated 2026-10-18 14:05

Source image: tool-photo.jpg

Model used: Qwen2.5-VL 7B Instruct

Describe the scene

A person's hand is holding a black and yellow screwdriver. The person is holding the screwdriver in their right hand, with their left hand slightly curled around for support. The background features a white ceiling adorned with hexagonal panels...

Detection: screwdriver

Found 1 instance(s) of 'screwdriver'.

Detected Objects

- **screwdriver** — box: (458, 273) → (976, 376)

Full Conversation Log

Analyst: Describe everything happening in this image in detail, as if explaining it to someone who cannot see it.

Vision AI Agent: A person's hand is holding a black and yellow screwdriver...

QUICK
ACTIONS

1

DETECTIONS

1

MESSAGES

2

MODEL

Qwen2.5-VL
7B

Client deliverable

Exported as Markdown, so it opens anywhere and pastes cleanly into email, docs and tickets.

format	Markdown (.md)
image	annotated .png

↓ Download report

Sections appear in the order you ran them, followed by every detected object and the full conversation.

08 CHOOSE YOUR ENGINE

Pick the engine that fits the machine.

Switch from the sidebar at any time. Every engine is open source under the Apache 2.0 licence.

MODEL	BEST FOR	MEMORY	RUNS ON
Qwen2.5-VL 7B HIGHEST ACCURACY	Detailed scenes, small text, precise boxes. The default when it is on your computer.	~17 GB	Apple Silicon Mac, NVIDIA GPU (4-bit ~8 GB)
Qwen2.5-VL 3B BALANCED	Most of the quality at roughly half the memory. Good on 16 GB machines.	~8 GB	Apple Silicon Mac, NVIDIA GPU (4-bit ~4 GB)
Qwen2.5-VL in Ollama 4-BIT	Use a vision model you already run in Ollama. It appears in the menu by itself.	set by Ollama	Any computer running the Ollama app
Moondream 2 FAST • CPU	Quick captions and answers on laptops with no graphics card.	~4 GB	Any computer, CPU only

Then set how closely it looks

Fast

~0.4 MP

Quickest replies and least memory.

Balanced

~0.8 MP

The default for everyday photos.

Detailed

~1.0 MP

Best for small text and small parts.

09 RUNS ON YOUR HARDWARE

Built for the Mac. At home on Windows too.



Apple Silicon Mac

Uses the Apple GPU (Metal)
 Loads weights straight onto the GPU
 bfloat16 for clean, stable output
 Two-command setup



Windows or Linux + NVIDIA

Runs on the GPU with CUDA
 Optional 4-bit loading to halve memory
 One-click launcher on Windows



Any computer

No graphics card needed
 Moondream 2 runs on the CPU
 Or use a model served by Ollama

Finds your model by itself

- 1 A folder you point it to
- 2 The project's models folder
- 3 The Hugging Face cache, including one copied from a Windows PC
- 4 "Search this Mac" across your home folder and drives
- 5 Vision models in a running Ollama app

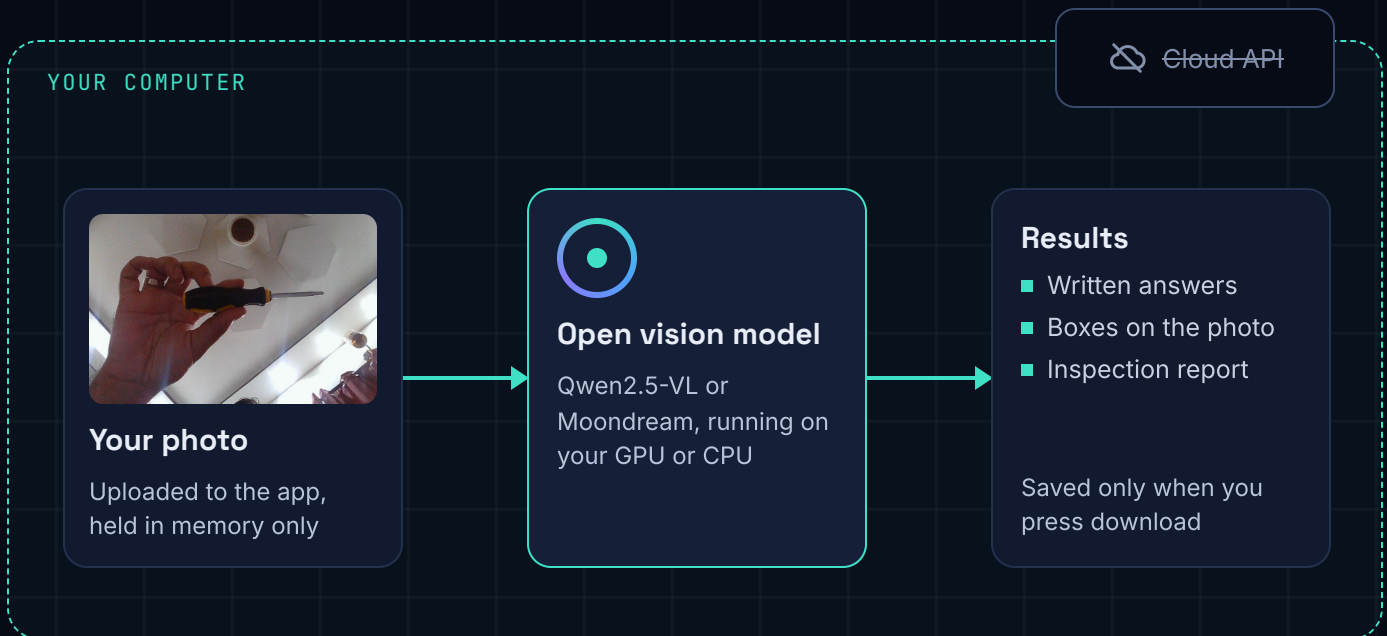
```
$ bash setup_mac.sh
✓ Apple Silicon Mac
✓ Python 3.12, virtual environment ready
Apple GPU (MPS) works ✓ precision: bfloat16
✓ All packages installed and working
➡ Looking for your Qwen2.5-VL model...
✓ READY Qwen2.5-VL 7B

$ bash run_mac.sh
Your browser will open at http://localhost:8501
```

10 PRIVATE BY DESIGN

Your images never leave your computer.

The model runs on your own processor and graphics chip. Photos, questions and reports stay on the machine in front of you.



No API keys

Nothing to sign up for or rotate.

No cloud calls

Analysis happens on this device.

No per-request cost

Run a thousand inspections for free.

Works offline

Once the model is on your computer.

Open-source models

Apache 2.0, free for commercial use.

One-click clear

Wipe the image, chat and report.

11 BUILT FOR

For teams that work from photos.



Workplace safety

Check site photos for helmets, vests, gloves and goggles, and see who is missing what.

PPE check

detect "helmet"



Warehouse and retail

Count stock on shelves, find forklifts and spot damaged boxes before they ship.

Count objects

detect "damaged box"



Quality inspection

Look over parts and finished products for damage, defects or anything out of place.

Spot anomalies

Report



Labels and signage

Read serial numbers, labels and signs exactly as written, with no retyping.

Read any text



Accessibility

Turn any picture into a clear description for someone who cannot see it.

Describe the scene



Field reports

Go from site photos to a written report the client can read, in one sitting.

Chat

Report export

Up and running in three steps.

1 Set it up once

On a Mac run `bash setup_mac.sh`. It installs everything, checks the Apple GPU and finds your model. On Windows, install the requirements and run `run.bat`.

2 Start the app

Run `bash run_mac.sh` and open `localhost:8501` in your browser.

3 Load, drop, ask

Press Load, drop in a photo and ask your first question, or run a one-click inspection.

Everything in the box

- Multi-turn chat with any image
- Detection by plain-English name
- Markdown inspection report
- Three image detail levels
- iPhone HEIC photos supported
- Five one-click inspections
- Annotated image download
- Four engines, switch any time
- Finds your model automatically
- Fully offline, no API keys

Built by

Asad Rehman

AI automation and computer vision engineer

upwork.com/freelancers/~01908fee0b3f85aa3b

Python, Streamlit and PyTorch

Hugging Face Transformers

Runs on Mac, Windows, Linux