

AI WORKSTATION SERIES

AiW-V Series

Cost-Effective Local AI Workstation

Gain accessible 32GB/64GB high-capacity VRAM. Deploy 27B-class LLMs and offline AI video generation pipelines directly to your desk.



32GB / 64GB

Physical HBM2 VRAM

Qwen 27B

Field-Tested & Proven

100% Local

Air-Gapped Data Privacy

High ROI

Budget-Friendly Entry

Take Your First Step into Local AI Without Breaking the Bank

Getting started with on-premise AI often presents a dilemma: consumer GPUs (8GB/16GB) frequently crash with out-of-memory errors on 27B models, while top-tier workstations cost a fortune. The AiW-V Series bridges this gap—delivering 32GB to 64GB of genuine enterprise VRAM at an accessible price point, so you can run practical models immediately.

01

32GB / 64GB Physical VRAM

Break through consumer memory constraints. 32GB (single GPU) or 64GB (dual GPU) HBM2 memory comfortably runs Qwen 27B models and video diffusion pipelines with zero OOM errors.

02

Zero Recurring Token Anxiety

Experimenting with AI requires iterative exploration. Stop paying per-token and per-minute on public clouds. Own your workstation and run unlimited local tasks 24/7 at fixed cost.

03

100% On-Premises Data Sovereignty

Internal SOPs, unpublished marketing briefs, and proprietary code remain securely inside your local network. Eliminate data leakage risks with true air-gapped physical privacy.

Comparison: Why AiW-V Fits Beginners & SMB Workloads			
Deployment Route	Upfront Investment	VRAM Capacity & Scale	Long-term Operating Cost
Consumer GPUs (8GB-16GB)	Moderate card price, limited expansion	Frequent OOM crashes on 27B models	Local execution, but restricted to small toy models
Cloud API SaaS	Zero upfront, pay per request	Per-token metering escalates quickly	Sensitive data leaves premises; uncontrolled expenses
ALNETS AiW-V Series	Turnkey appliance, highly accessible	Standard 32GB / 64GB HBM2 Memory	One-time purchase, unlimited free local calls, 100% data autonomy



Multi-Seat Team & Lab Collaboration

Zero server-room or cabinet dependencies. Preloaded with full AI runtime environments, standard gigabit LAN provides instant turnkey access, allowing engineers and creators across the local network to

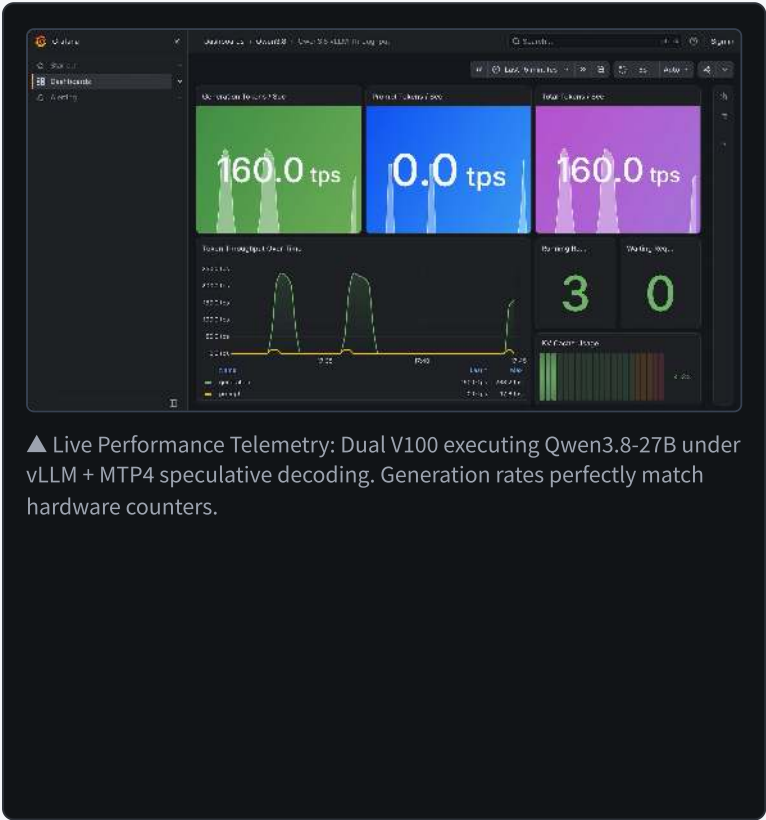


Industrial Direct Airflow • Heavy-Duty Cooling

Engineered for sustained heavy workloads across multi-step video rendering and concurrent LLM inference. All-metal chassis with high-density mesh grilles delivers high-volume

Qwen3.8-27B Verified Benchmark: Real Telemetry, Zero Guesswork

We let real production telemetry speak for itself. In our AiW-V workstation test environment paired with containerized vLLM and MTP acceleration, Qwen3.8-27B demonstrates stable, high-throughput on-premise inference.



131,072 tokens
128K Context Stress-Test Verified
Passed full 128K extreme context heavy-load stress testing. Ingest entire codebases or lengthy documentation in a single turn with rock-solid VRAM scheduling and zero OOM truncation.

425.67 tokens/s
High-Concurrency Peak Throughput
Under 8 concurrent short/medium requests, MTP4 speculative draft acceptance achieves 99.8%~100%, easily serving multi-seat local LAN offices.

OpenAI API
Standardized Local Interface
Exposes standard OpenAI endpoints. Integrate with internal web browsers, Dify, or FastGPT knowledge bases in under 5 minutes.

Test Setup & Conditions	Concurrency	Output Tokens / Request	Wall Time (s)	Measured Throughput	Scheduler Status
AiW-V32 Single 32GB (Quantized)	1 stream	256 tokens	~6.98 s	~36.7 tok/s (Smooth reading)	Flawless single-user experience
AiW-V64 Dual GPU (32K + MTP4)	8 streams	7,200 tokens (900/stream)	16.91 s	425.67 tok/s (Fast Burst)	Zero queue delay across 8 requests
AiW-V64 Dual GPU (128K Stress Test + MTP4)	3 streams	6,000 tokens (2000/stream)	25.28 s	237.26 tok/s (Steady State)	Consistent 128K long-context generation



Gigabit LAN Turnkey Integration

Standard RJ45 gigabit port connects directly to office switches, serving dedicated local model APIs to all office devices.



Field-Verified Telemetry & Runtime

Engineers verify vLLM throughput and memory scheduling directly on-device. Zero setup hurdles for team deployment.

Offline Video & Image Studio: Run Workflows at Zero Ongoing Cost

Video diffusion models are the ultimate test of GPU memory. Public platforms charge per generation and often throttle heavy usage. With 32GB+ physical VRAM, AiW-V gives creators a dedicated, subscription-free offline sandbox.

ComfyUI Offline Video & Visual Creation Pipeline

Deploy state-of-the-art open-source video diffusion and image workflows locally. Transform still product photos and concept art into 3-5s dynamic motion video clips with fluid camera movement.

Step 01

Import still image & prompt

Step 02

Configure motion camera & steps

Step 03

Render locally in offline batch



Image-to-Video Animation

Tested support for high-quality multi-step video diffusion. Convert static catalog photos into dynamic motion showcase clips for advertising and social feeds.

Dynamic Product Video

3D Orbit Camera



Batch Overnight Rendering

Tune prompts and workflow nodes during work hours, queue batch tasks before leaving, and let the workstation render offline overnight. Zero recurring bills.

Unattended Rendering

Zero Credit Metering



Digital Avatar Plates

Render natural expressions and consistent human character video plates offline, providing pristine footage for lip-sync dubbing and virtual presenters.

Virtual Avatar Plates

Consistent Facial Motion



The Essential Value of Large VRAM:

Consumer 8GB/16GB cards quickly throw memory allocation errors when handling modern video pipelines. AiW-V's 32GB/64GB capacity provides the necessary headroom for complex node-based node graphs without crashes.

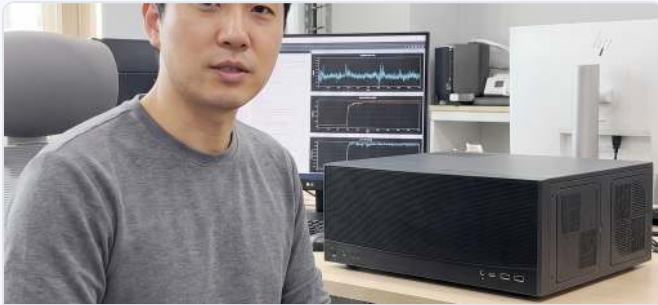


Creator's Perspective:

The greatest advantage of local generation isn't ultra-fast rendering—it's the freedom to fail without cost. On AiW-V, tweak a prompt 50 times without worrying about running out of paid credits.

Built for Practical Needs: 4 High-Value Scenarios for Beginners & SMBs

Instead of pitching unrealistic heavy industrial claims, AiW-V addresses the immediate, day-to-day productivity needs of small teams and self-starters.



Scenario 01

SMB Preferred

Internal Knowledge Q&A (Private Knowledge Base)

The Challenge:

Query employee handbooks and product specs without risking cloud exposure or purchasing expensive server racks.

AiW-V Solution:

Host Dify paired with Qwen3.8-27B. Staff query internal files via browsers in a strictly isolated local network.

✓ Low Trial Barrier | Local Collaboration | Zero Cloud Leakage



Scenario 02

Content & E-Commerce

Offline Media Studio for Creators & E-Commerce

The Challenge:

Produce daily product motion clips without burning through costly subscription credits on public platforms.

AiW-V Solution:

Run ComfyUI image-to-video batch jobs locally overnight. Turn static photos into motion clips while you sleep.

✓ Zero Per-Credit Costs | Free Iterative Trial | Batch Overnight Yield



Scenario 03

Higher Ed & Labs

Hands-On Learning & Prototyping for Students

The Challenge:

Students restricted by 8GB/16GB laptops and months-long reservation queues on university compute clusters.

AiW-V Solution:

Dedicated access to 32GB VRAM on your personal desk. Practice model deployment, prompt tuning, and agent code.



Scenario 04

Productivity Hub

Meeting Transcripts & Brief Extraction

The Challenge:

Staff face lengthy meeting logs and contract drafts while enterprise SaaS subscriptions multiply rapidly.

AiW-V Solution:

Ingest raw transcripts directly into local Qwen3.8-27B for instant summaries. Shared LAN API without limits.

AiW-V32 / AiW-V64: Shared Chassis Architecture, Dual Compute Tiers

AiW-V utilizes an industrial-grade cold-rolled steel chassis with dedicated dual airflow tunnels. Straightforward selection: choose 32GB single-GPU or 64GB dual-GPU based on team size and workload.



Reinforced Steel Chassis • Dedicated Dual-Tunnel Airflow Cooling

Built for uninterrupted heavy-load LLM inference and video rendering. High-flow top and side vents paired with speed-controlled fans keep V100 cards cool and prevent thermal throttling. Standard desktop footprint (440×395×165mm) fits beside or under standard office desks without server room racks.

AiW-V32

Entry Tier • 32GB VRAM • Solo Creator

Designed for independent creators, video editors, and university developers. Delivers single-stream Qwen3.8-27B inference and offline ComfyUI creation on an accessible budget.

1 × V100
GPU Unit

32GB
HBM2 VRAM

750W
Power Supply

AiW-V64

Advanced Tier • 64GB VRAM • Team Shared

Engineered for small studios and collaborative teams. Handles dual-GPU vLLM 128K context workloads, multi-seat concurrent office queries, or dedicated split LLM + video tasks.

2 × V100
GPU Units

64GB
HBM2 VRAM

1000W
Redundant PSU

Hardware Specifications	AiW-V32 Specification	AiW-V64 Specification
GPU Configuration	1 × NVIDIA V100 32GB	2 × NVIDIA V100 32GB
Total Physical VRAM	32GB HBM2 High-Bandwidth Memory	64GB HBM2 (32GB × 2)
Memory Bandwidth	Up to 900 GB/s	Up to 900 GB/s (Per GPU)
CPU Processor	Intel Xeon E5-2696 Enterprise Multi-Core	Intel Xeon E5-2696 Enterprise Multi-Core
System RAM & Storage	32GB DDR4 512GB High-Speed NVMe SSD	64GB DDR4 512GB High-Speed NVMe SSD
Rated Power Supply	750W Regulated Power Supply	1000W Redundant Power Supply
Cooling & Ethernet	Dual Tunnel Steel Airflow 1 × Gigabit RJ45	Dual Tunnel Steel Airflow 1 × Gigabit RJ45
Pre-installed OS	Pre-installed Ubuntu Linux (vLLM & ComfyUI ready)	Pre-installed Ubuntu Linux (vLLM & ComfyUI ready)
Chassis Dimensions & Gross Weight	440 × 395 × 165 mm Approx. 14.9 kg	440 × 395 × 165 mm Approx. 15.8 kg

* VRAM capacities reflect installed physical hardware. Actual port configuration follows order Industrial Cold-Rolled Steel | Standard Desktop Footprint | Plug-and-Play

Field AI Triad Architecture: AiR + AiA + AiW

From network infrastructure to persistent multi-agent orchestration and heavy on-premise model inference. Three products collaborate over standard LAN to deliver an autonomous, private AI loop.



Network Foundation

AiR AI Cloud Gateway

- **Dedicated AI LAN Setup:** Isolates internal model access, ensuring pristine local throughput.
- **Zero-Config Remote Tunnel:** Secure off-site employee access without complex VPN overhead.
- **High-Speed Egress Acceleration:** Accelerates HuggingFace model downloads and manages egress policies.



Persistent Agent Runtime

AiA Agent Cluster Nodes

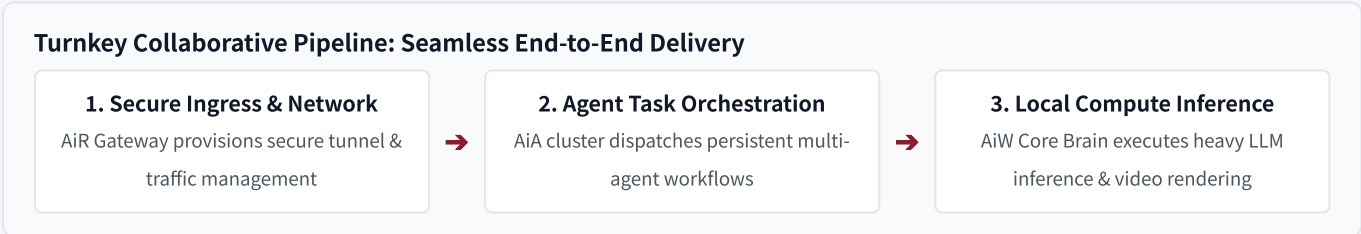
- **24/7 Always-On Autonomy:** Frees agents from laptops into persistent hardware runtimes.
- **Multi-Agent Orchestration:** Manages customer service, copy generation, and data scraping flows.
- **Compute/Logic Separation:** Handles business logic and forwards heavy tensor jobs to AiW.



Core Brain • Inference Center

AiW Workstation Brain

- **High-Throughput Local LLM:** Serves Qwen3.8-27B locally via standard OpenAI-compatible endpoints.
- **High-VRAM Video Synthesis:** 32GB/64GB HBM2 memory handles complex ComfyUI diffusion graphs.
- **100% Data Sovereignty:** Model weights, private SOPs, and vector stores remain physical on-prem.



Out-of-the-Box Readiness & Support

All units arrive preloaded with Ubuntu Linux, NVIDIA drivers, Docker container runtime, and validated deployment templates for vLLM and ComfyUI. Every workstation undergoes 72-hour burn-in stress testing, backed by standard hardware warranty and technical advisory.