

AiW-V Local AI Workstation

Product Datasheet · AiW-V32 / AiW-V64

ALN-DS-AiW-V-EN · v1.1 · 2026-10-26 · English translation for review

Product overview

AiW-V is a budget-friendly local AI workstation using NVIDIA V100 32GB GPUs. AiW-V32 has one GPU; AiW-V64 has two. Both belong to Ai Workstation and support adapted model services, text processing and offline creative workflows.

The confirmed hardware baseline carries over, primarily using air cooling. Legacy AW-A, AW-V96 and standard liquid-cooling claims are not current product specifications. Models, knowledge applications and Agents retain distinct component roles.

Core capabilities and application value

Capability	Function and value
Local model services	Run adapted quantized Q&A, text processing and model APIs.
Document/content processing	Provide compute for local summaries, questions and drafts.
Offline video creation	V100 + ComfyUI quantized-video projects have run successfully.
Single/dual GPU options	Choose by VRAM, supported sharding or service separation.
Knowledge applications	Supply model services with separately deployed retrieval components.
Models for Agents	Provide compatible local model endpoints; Agent execution is separate.

Typical business uses

Budget-conscious individuals, studios, small teams, teaching labs and offline creators selecting for a specific model and task.

01 / Models and Technical Specifications

Model configurations and interfaces

Item	AiW-V32	AiW-V64
GPU	1 × NVIDIA V100 32GB	2 × NVIDIA V100 32GB
Total VRAM	32GB HBM2	64GB HBM2 (32GB × 2)
System RAM	32GB	64GB
Storage	512GB SSD	512GB SSD
CPU	Intel Xeon E5-2696; suffix unconfirmed	Same
Platform	X99	X99
Network	1 × Gigabit RJ45	Same
PSU rated capacity	750W	1000W
Cooling	Primarily air cooling	Same
OS	Ubuntu preinstalled; version per application	Same
Unit / package dimensions	440 × 395 × 165 mm / 590 × 500 × 270 mm	Same
Gross weight	14.9 kg	14.9 kg

Technical notes

Two GPUs do not become a single 64GB memory device; pooled use requires suitable software and does not guarantee double throughput. 750W/1000W are PSU capacities, not operating consumption. GPU form factor, interconnect and display access follow the actual configuration; dimensions preserve the confirmed order.

Unconfirmed quantitative specifications

CPU suffix/core count, GPU form factor/interconnect, USB/display interfaces, system consumption and environmental metrics require further confirmation.

02 / Model Services and Applications

Existing adaptation paths

Project	Evidence and use
Single-V100 Q&A	Quantized Qwen3.6-27B with llama.cpp CUDA has project records as a single-GPU reference.
Dual-V100 service	Qwen3.8-27B on an adapted backend has API self-check records as a dual-GPU deployment reference.
Offline video	Single-V100 text/image-to-video project records using ComfyUI and quantized video models.
Text API	Adapted backends can expose endpoints for compatible applications.
Retrieval / Agents	Retrieval, embedding, permissions and Agent tools are separate components using model compute.

These are project adaptation records, not a universal factory-installed list or fixed system performance. This edition excludes historical throughput whose GPU identity was not verified per run. A configured context limit does not establish accepted full-context concurrency.

Single GPU, dual GPUs and VRAM

V32 has one 32GB card; V64 has two 32GB cards, totaling 64GB. Sharding requires a compatible framework; VRAM does not automatically merge into one 64GB card. Separate services can be planned, but simultaneous stability requires testing; speed and production capacity do not automatically double.

Models and data paths

Once models and dependencies are prepared, adapted workflows may run locally. Cloud models, online tools and remote resources may still transfer data externally. Configure permissions, networking and dependencies for actual tasks; local hardware does not establish automatic compliance or absolute offline operation.

03 / Software Environment and Operations

Deployment specifications

Stage	Configuration requirement
OS	Ubuntu preinstalled; choose a version compatible with backend and drivers.
Compute environment	Verify NVIDIA driver, CUDA, backend and installed GPUs; retain version records.
Models	Select by VRAM, quantization, context and license; test loading.
Service endpoint	Configure bind address, port, authentication and client protocol; verify calls.
Creative workflow	Fix model, nodes, dependencies and workflow files; test inputs/outputs.
Operations	Check GPU use, disk capacity, logs and directory backups; retain configuration.

ALWork remains planned/in development and is not an already-delivered one-click model-management standard feature. An adapted V100 backend running an FP8 checkpoint does not mean native FP8 Tensor Core support. Compatibility is version-specific.

Hardware interpretation

750W / 1000W are PSU capacities, not system operating power. Air cooling is the current confirmed baseline. GPU board type, interconnect, auxiliary display access and CPU suffix require actual configuration records; legacy GPU HDMI/DP-direct-output claims do not apply.

Application verification

For model services, test loading, short/long inputs, outputs, required concurrency and recovery. For video, record resolution, frames, steps and completion time. Quantization, context and offloading affect quality/speed. Undocumented training, fine-tuning, 70B or real-time video capabilities are not general guarantees.

04 / Topology and Deployment Selection

Typical topology

Applications / Agents / API Gateway -> LAN -> AiW-V models
Creator -> adapted ComfyUI workflow -> local media output
Knowledge components -> retrieval/embedding -> model generation

Scenarios and application value

Scenario	Configuration focus	Value
Local Q&A/summaries	Adapted quantized service	Process text on site.
Offline creation	Fixed video workflow	Create short samples or animated images.
Teaching labs	Drivers, models and API setup	Learn local deployment and integration.
Knowledge projects	Retrieval application + models	Validate answers on real documents.

Model and component selection

V32 suits a defined single-GPU task; V64 suits compatible sharding or planned service separation. Validate models/backends first. Greater total VRAM does not mean every model is supported.

Deployment checks

Confirm configuration/GPU identity -> fix OS/driver/backend -> load models/dependencies -> configure endpoints/workflow -> test real tasks -> check load/cooling -> preserve environment/backups.

Website: www.alnets.net **Email:** sale@alnets.net