NeuralMesh™ Scales Training, Feeds Inference for Generative AI in M&E

NeuralMesh runs fast S3 and file access on the same data, so generative video and image teams train, fine-tune, and serve models from a single copy, keep GPUs busy, and deliver high-quality visual content faster.

Challenges

- GPUs sit idle while training data loads
- Reads stall out during checkpoints
- Pre-training runs are lost due to storage infrastructure failures
- Video and image data is copied repeatedly into every pipeline stage
- Directories are hand-tuned to keep small-file datasets usable

Solution

NeuralMesh™ spreads data and metadata evenly across every node, so huge sets of small image and video files load without lookup stalls. Parallel striping absorbs checkpoint bursts while dataset reads keep streaming. Its S3 and POSIX run natively on the same physical data blocks, so curation, training, and inference share one copy. Distributed data protection keeps the cluster serving through a node failure, and the Kubernetes Operator deploys storage as code next to GPU jobs.

Benefits

- Keep GPUs running at more than 95% utilization on training workloads
- Sustain concurrent 1TB+ reads and writes without performance degradation
- Run pre-training without storage failures
- Serve training and inference from one copy of video and image data
- Scale small-file datasets without hand-tuning directories

Expedite Creative Timelines for Generative AI Production

Generative video and image companies compete on release cadence. A new model ships only after many experiments, and each experiment moves at the pace of the slowest system beneath it. When a run stalls, researchers lose the day and release dates slip. Creators waiting on the next model try another tool, and the team that iterates faster ships first.

NeuralMesh runs as one file system across owned bare metal, rented GPU clusters, and public cloud. Data Replication shows the destination file system before its contents arrive and pulls only what each job needs, so training can follow GPU capacity between sites. Multitenancy gives each team isolated performance and security on one shared cluster

Krea AI offers one creative platform of 64+ models, including image, video, and 3D models, serving more than **30 million users**, running on NeuralMesh which feeds hundreds of concurrent clients. **Luma AI** trains multimodal world models across video, audio, and language to enable millions of creators to generate cinematic-quality videos from text prompts, with **40x faster training startup** on NeuralMesh.

> *"NeuralMesh made it possible to train models at the scale we needed... The difference between NeuralMesh and our earlier solutions is night and day."*
>
> — Will Beddow, Head of Supercomputing and AI Infrastructure, Krea AI

KREA · Hugging Face · cohere · stability.ai · Luma AI · IIElevenLabs

© 2026 All rights reserved. WEKA and the WEKA logo are registered trademarks of WekaIO, Inc.
Other trade names used herein may be trademarks of their respective owners.