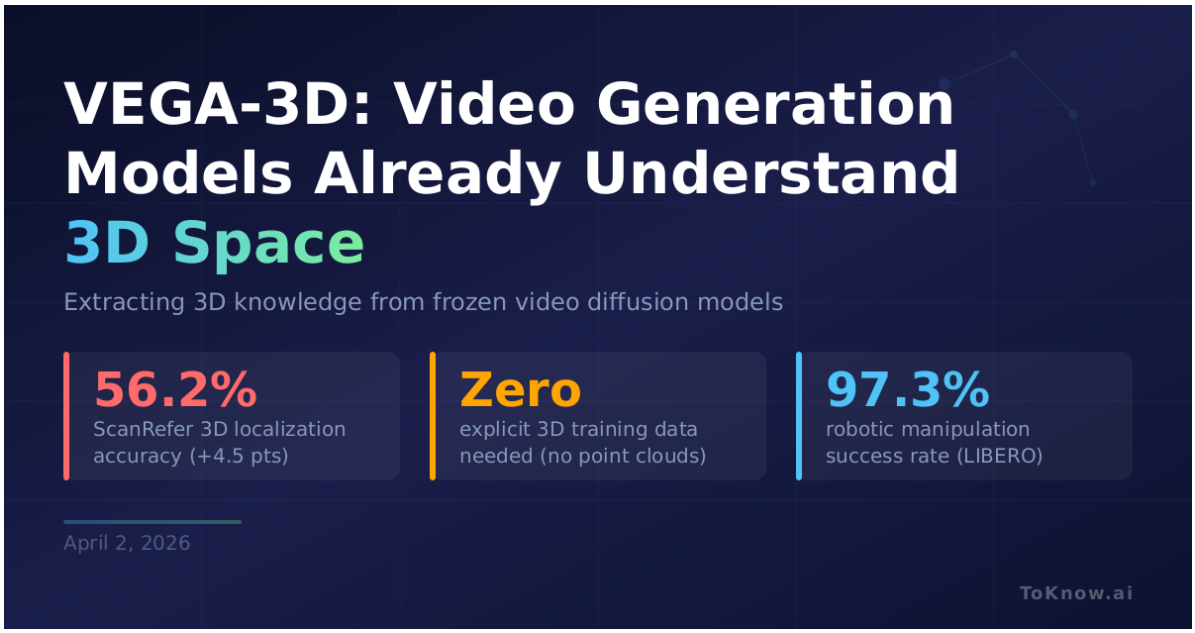
VEGA-3D: Video Generation Models Already Understand 3D Space

Kabui, Charles

2026-04-02



Researchers at Huazhong University of Science and Technology and Baidu released VEGA-3D, a plug-and-play framework that gives multimodal AI models 3D spatial awareness by tapping into something video generation models already know. The core insight: to generate temporally coherent video, a model must implicitly learn 3D geometry, depth, and object permanence. VEGA-3D exploits this by repurposing a frozen video diffusion model as a "Latent

World Simulator," extracting spatial features not from its final outputs but from intermediate denoising steps, where geometric information is richest. These features are merged with a language model's visual tokens through adaptive gated fusion, a learned mechanism that dynamically weights geometric and semantic signals per token. On ScanRefer, a 3D object localization benchmark, it scores 56.2% accuracy, a 4.5-point improvement over the baseline. On robotic manipulation tasks (LIBERO), it hits 97.3% average success rate. All without any explicit 3D training data: no point clouds, no depth maps, no 3D annotations.

Collecting and annotating 3D data is one of the biggest bottlenecks in robotics and spatial AI. VEGA-3D sidesteps the problem entirely by recycling knowledge already embedded in video models that the community has spent billions training. Because it works as a drop-in module, any existing multimodal language model can gain spatial reasoning without retraining from scratch.

Generative models may contain far more structured knowledge in their internal representations than we currently use. The valuable signal isn't the output, it's the intermediate computation. For related work on extracting 3D understanding from video, see LoGeR: DeepMind's 3D Reconstruction That Scales to 10,000 Frames.

Sources:

- VEGA-3D Paper (arXiv)
- VEGA-3D Project Page
- VEGA-3D GitHub Repository
- HuggingFace Daily Papers
