



From Fast Pixels to Interactive Worlds

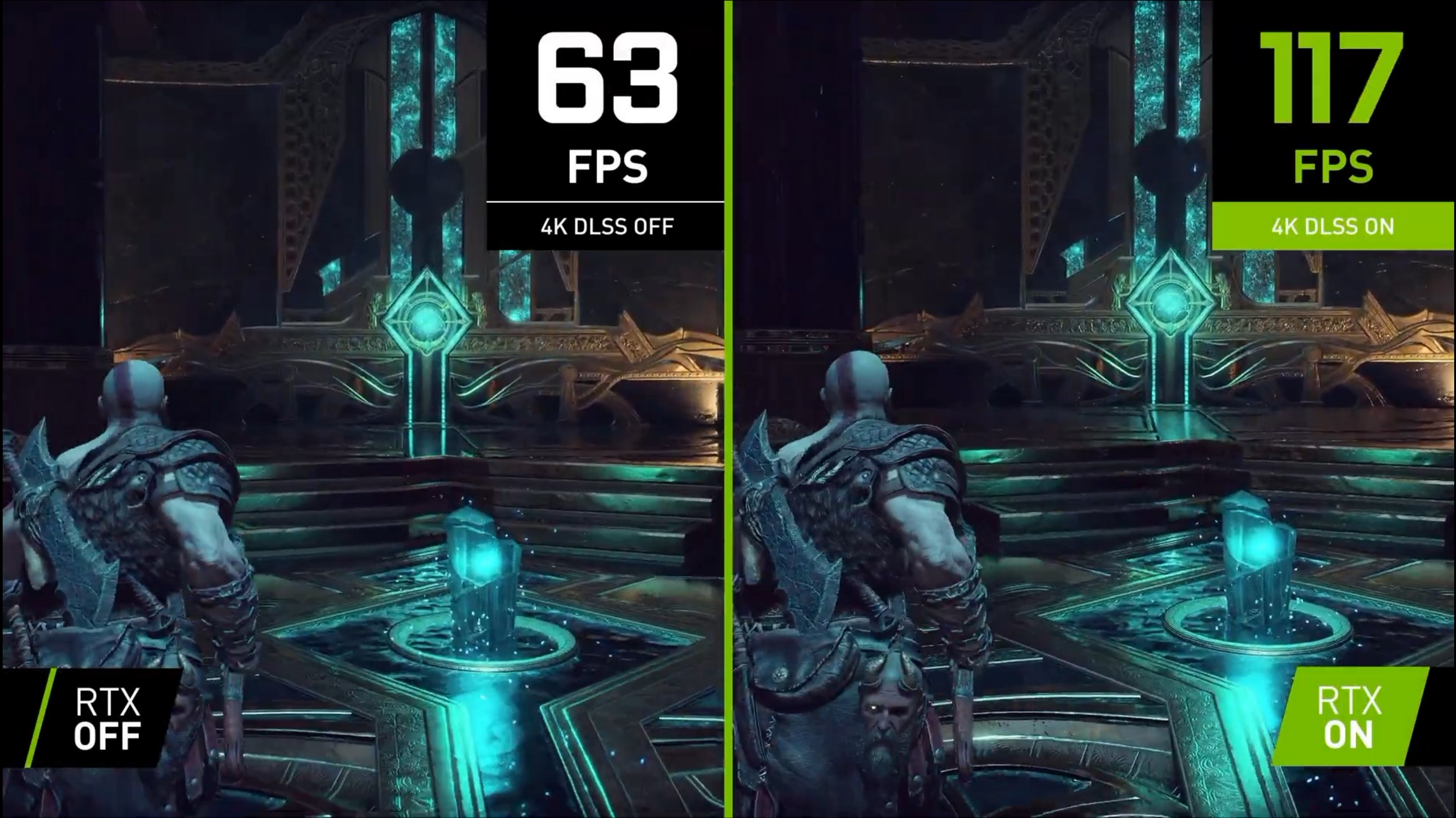
Challenges in Efficient and Responsive World Models

Anjul Patney, AI for Experiences at NVIDIA

SIGGRAPH 2026



State-of-the-art in AI Graphics: Super Resolution



DLSS SR

State-of-the-art in AI Graphics: Frame Generation



DLSS FG

State-of-the-art in AI Graphics: Ray Reconstruction

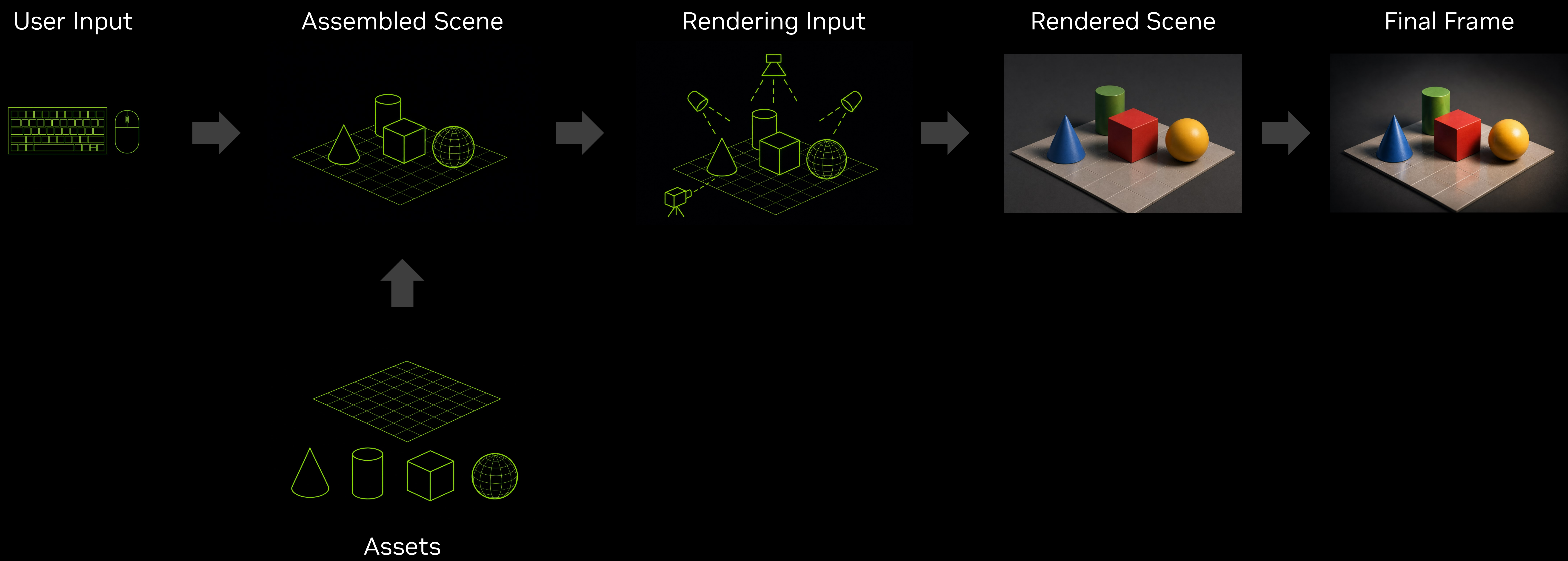


Ray Reconstruction Off

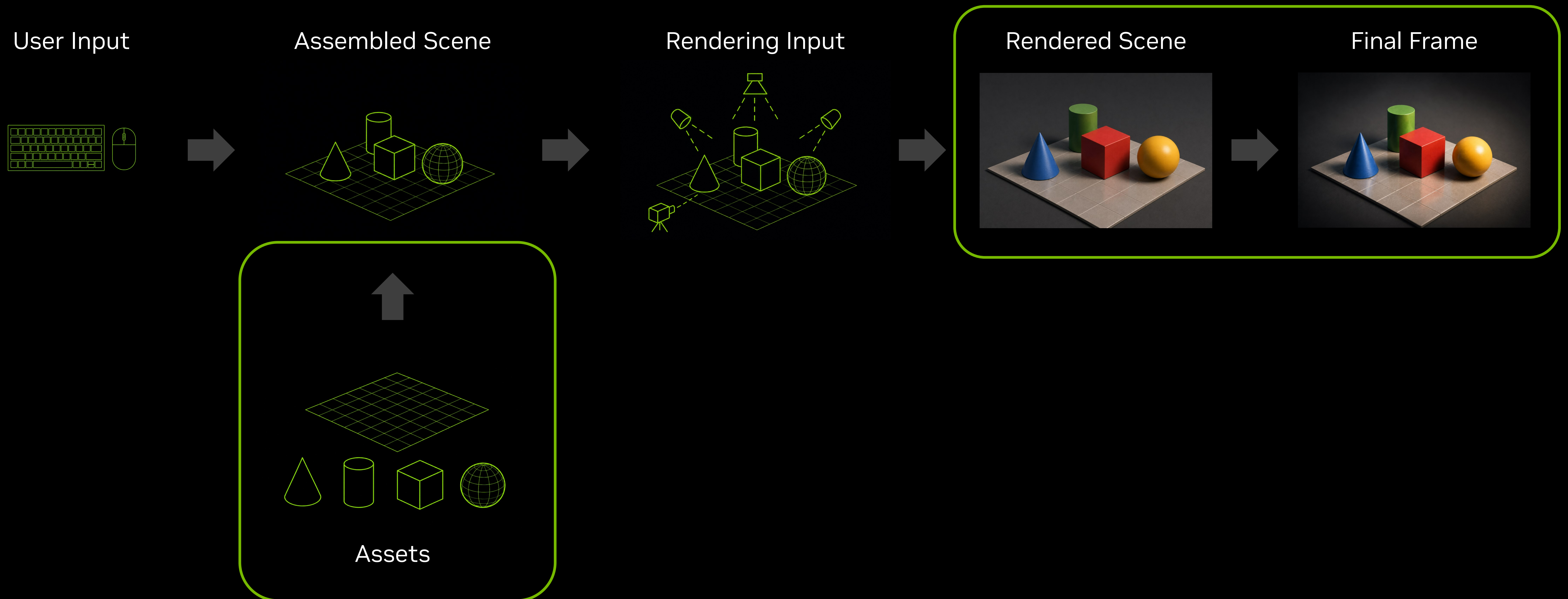
Ray Reconstruction On

DLSS RR

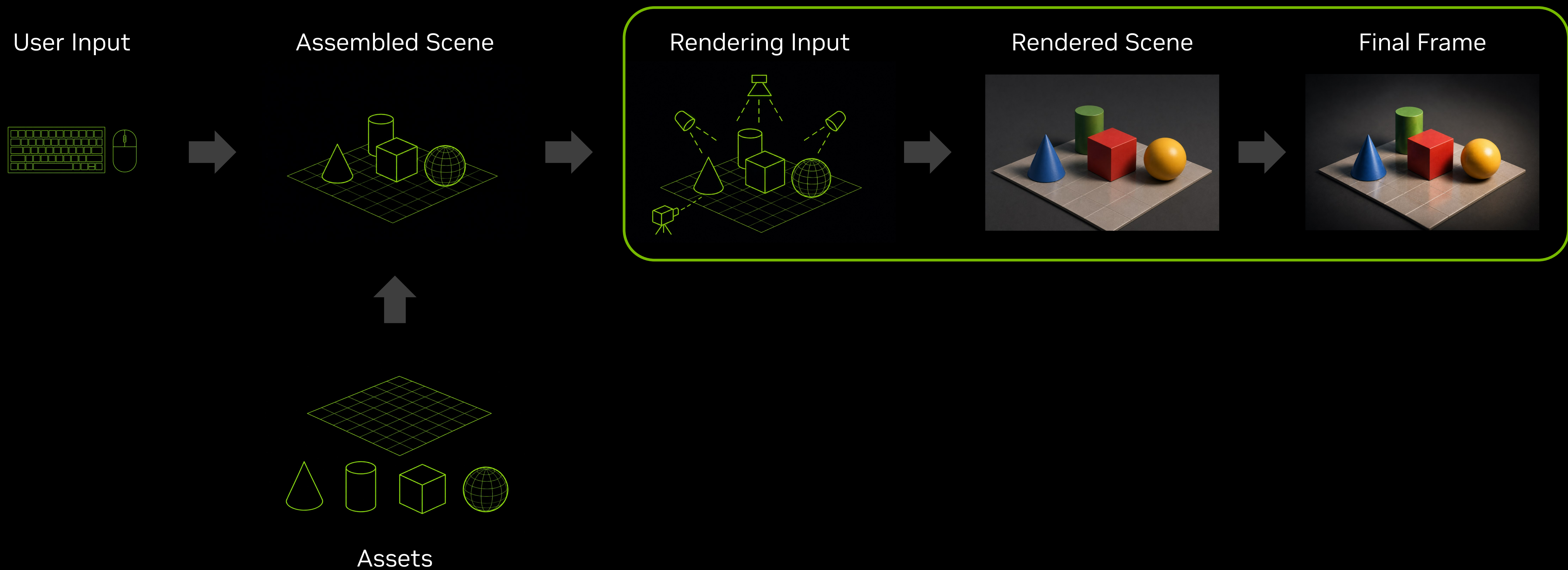
So far, the rendering pipeline is largely unmodified



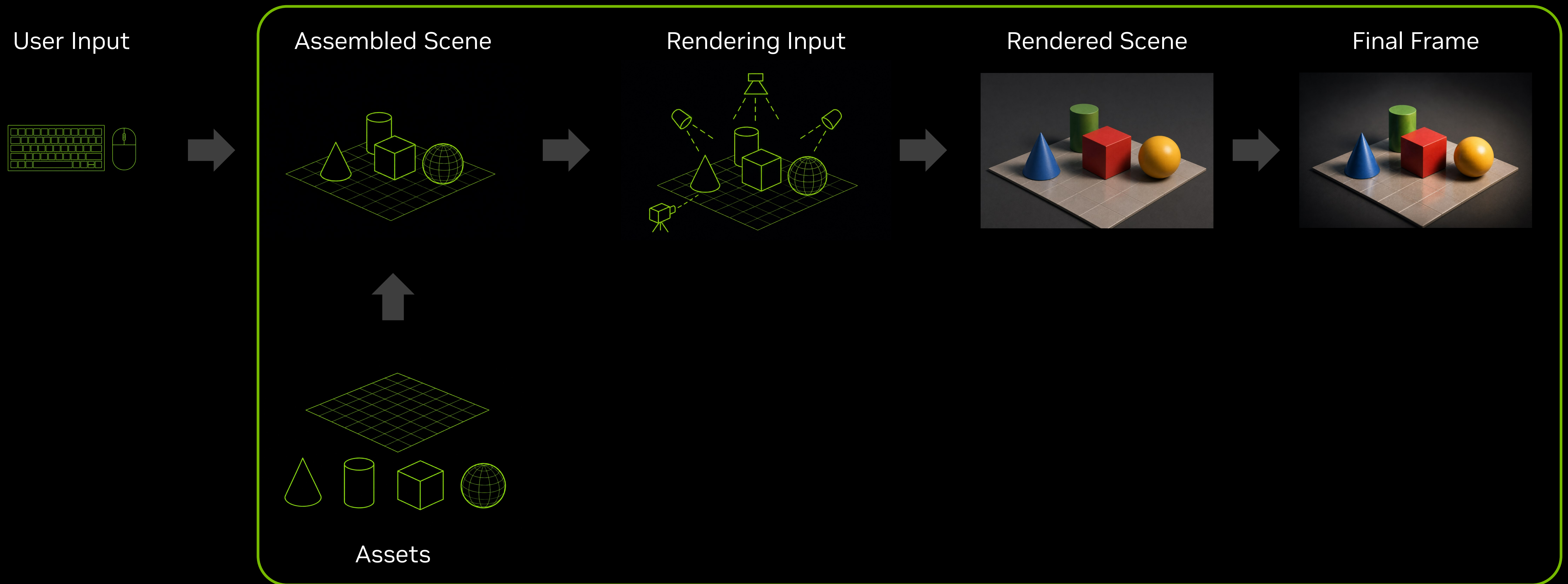
Real-time AI works at the edges of the graphics pipeline



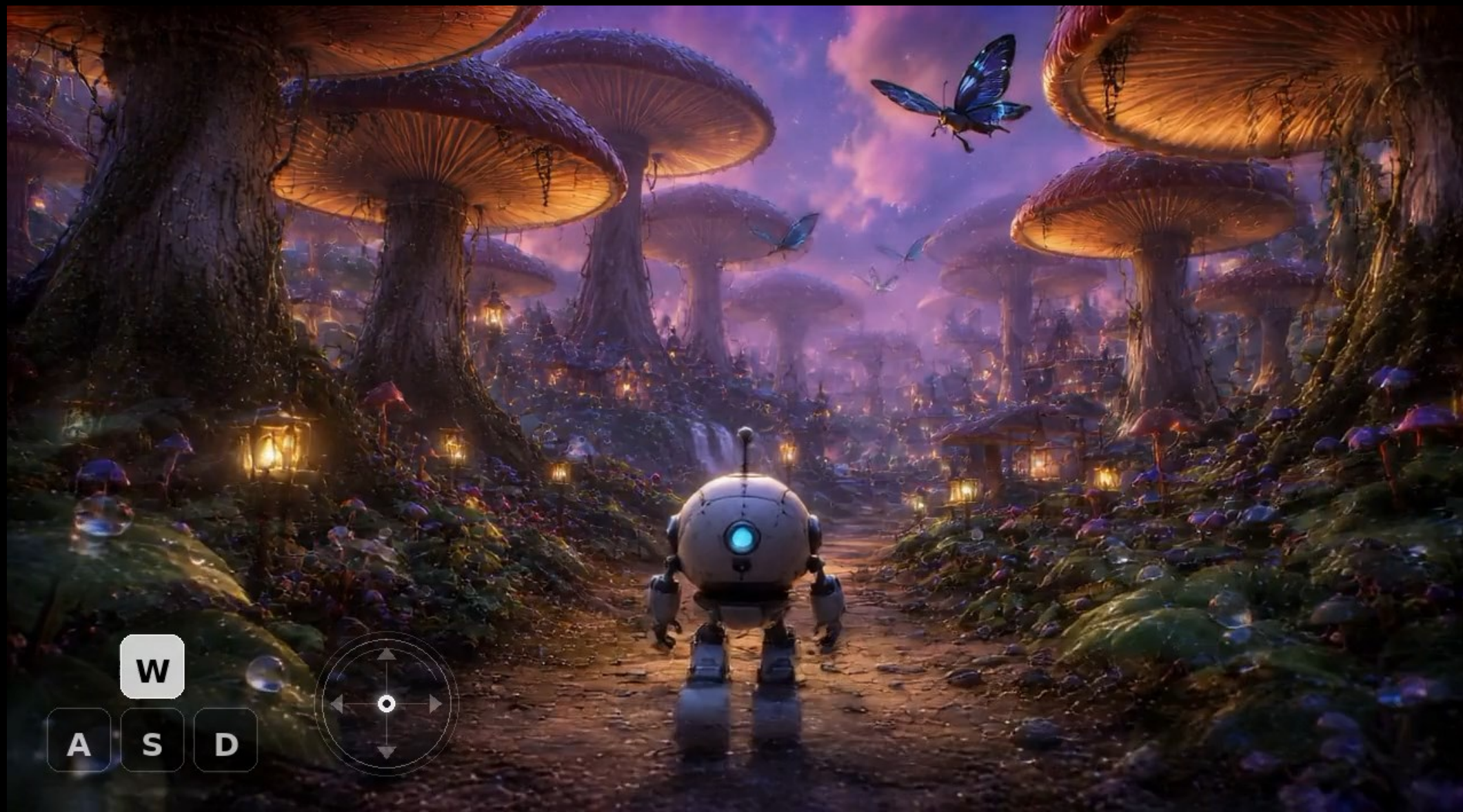
Generative graphics enables increased AI responsibility



What if we could replace the entire pipeline?



Interactive world models could redefine video game graphics



SANA-WM



Genie 3

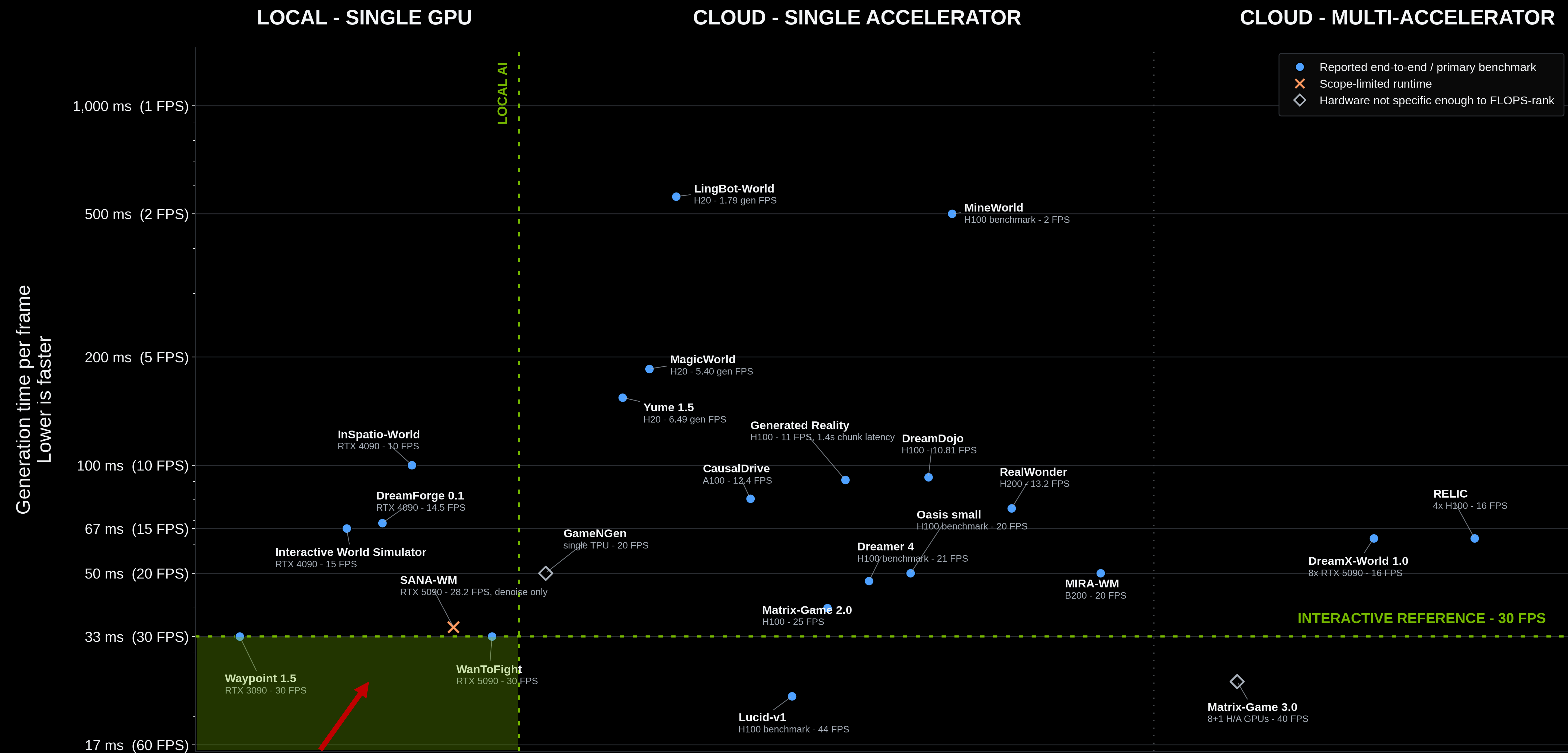


LingBot-World

“World Models Are AI’s Next Frontier”

—TIME Ideas (July 2026)

Interactive world models depend on cloud AI



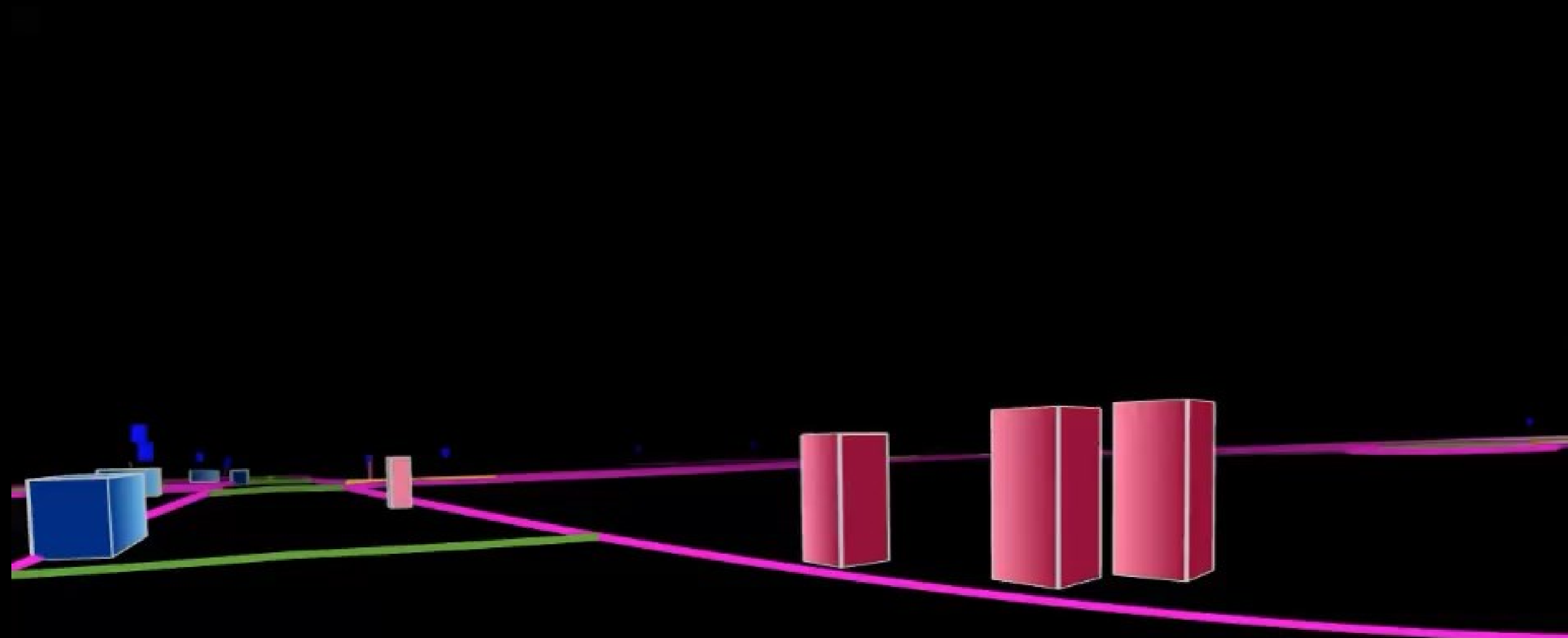
Interactive Local
World models



Achieving interactive
framerate



OmniDreams: A Cosmos 2.5-based driving world model

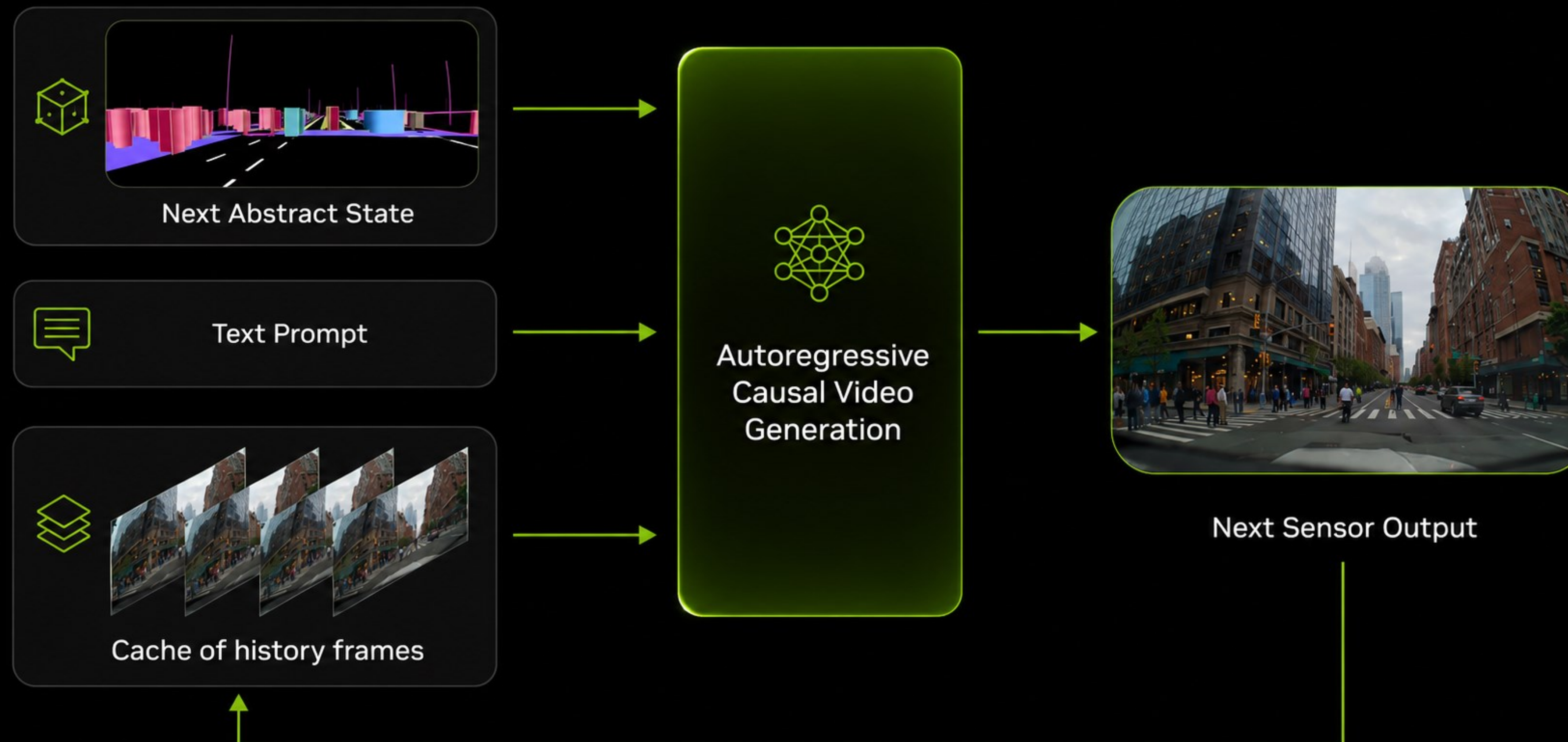


Abstract Scene State



Output

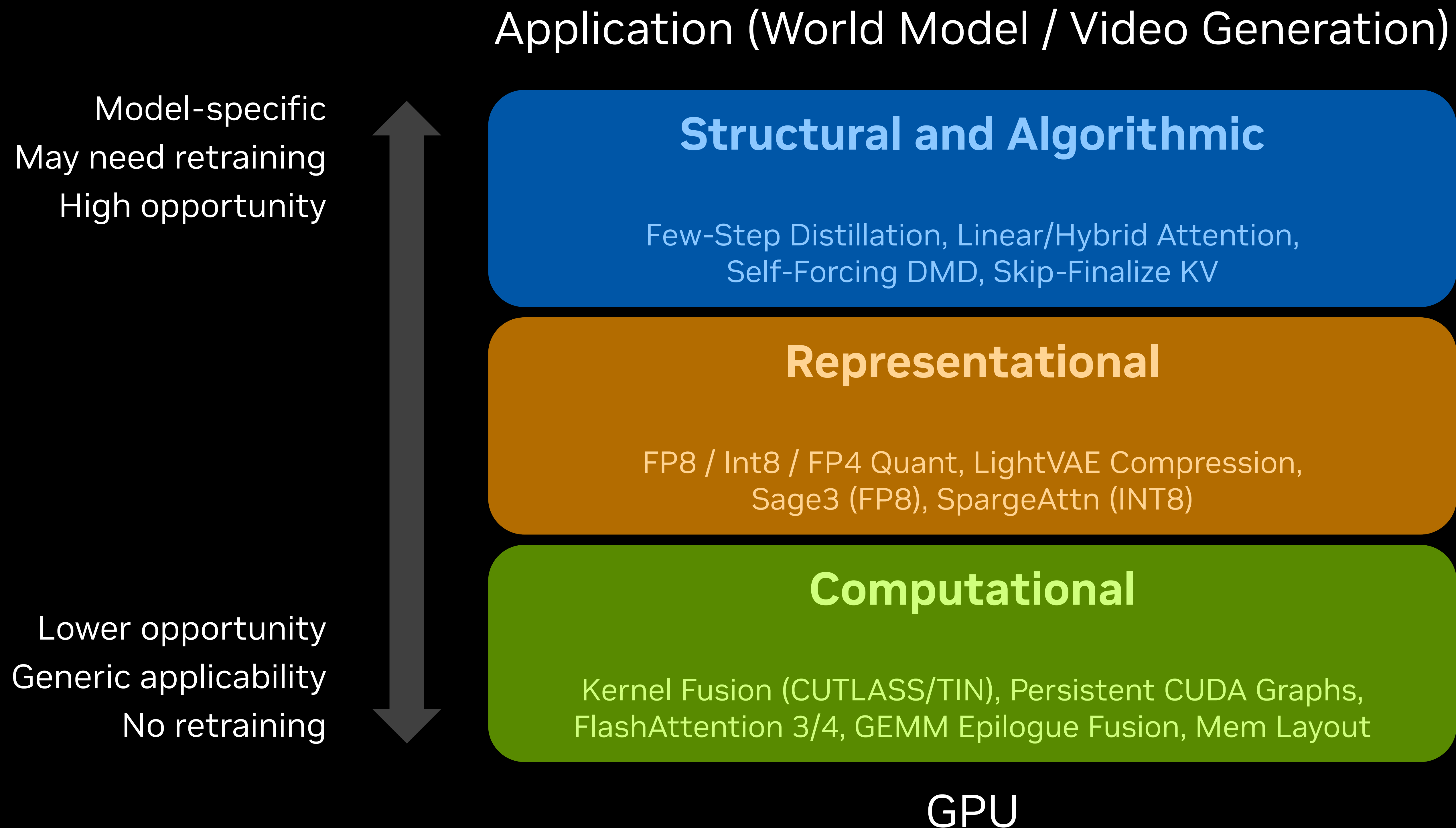
OmniDreams builds on autoregressive video generation



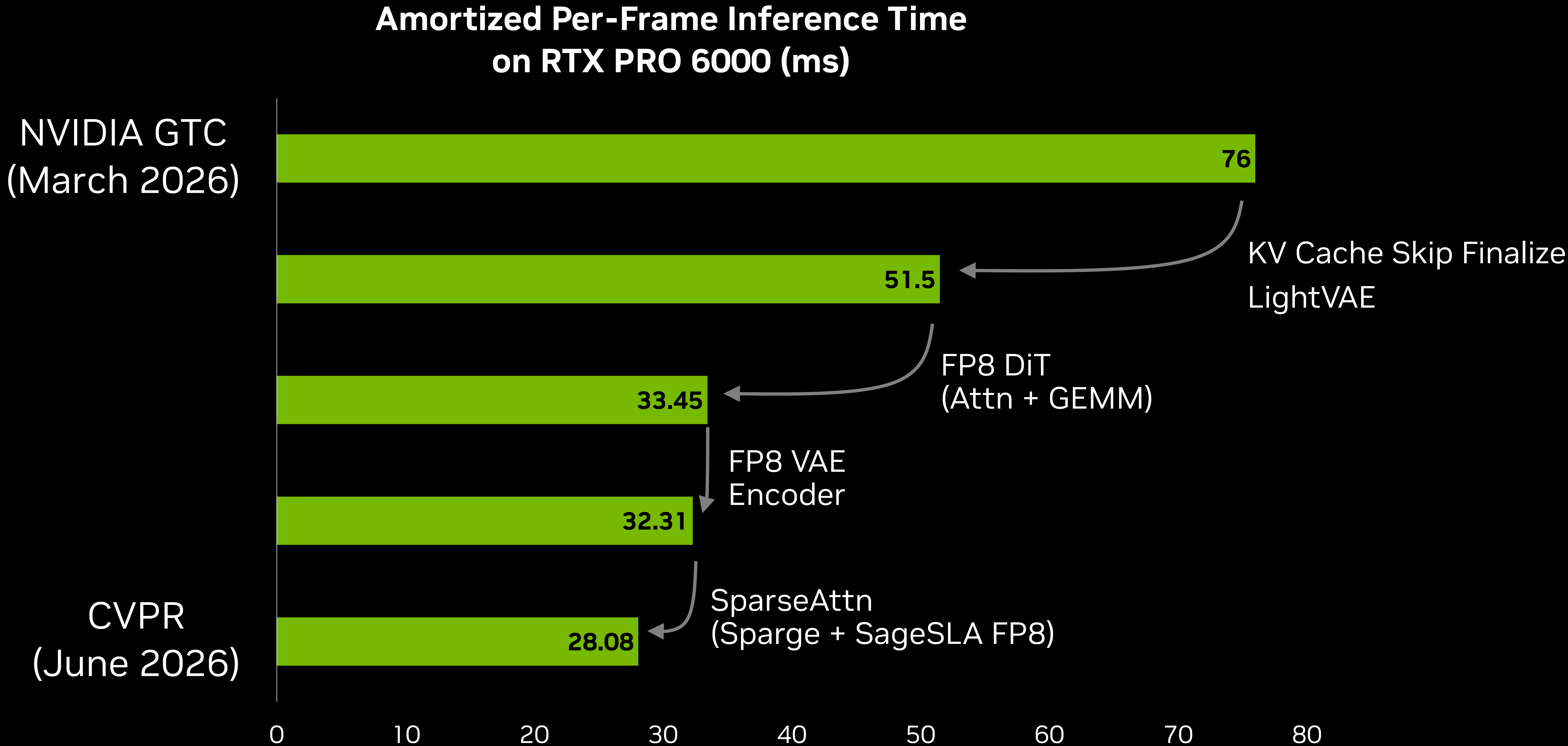
OmniDreams at GDC 2026: 52 FPS on GB300



Optimization opportunities span the implementation stack



From datacenter (GB300) to a backpack PC (RTX PRO 6000)



OmniDreams at CVPR 2026: 30 FPS on RTX PRO 6000)

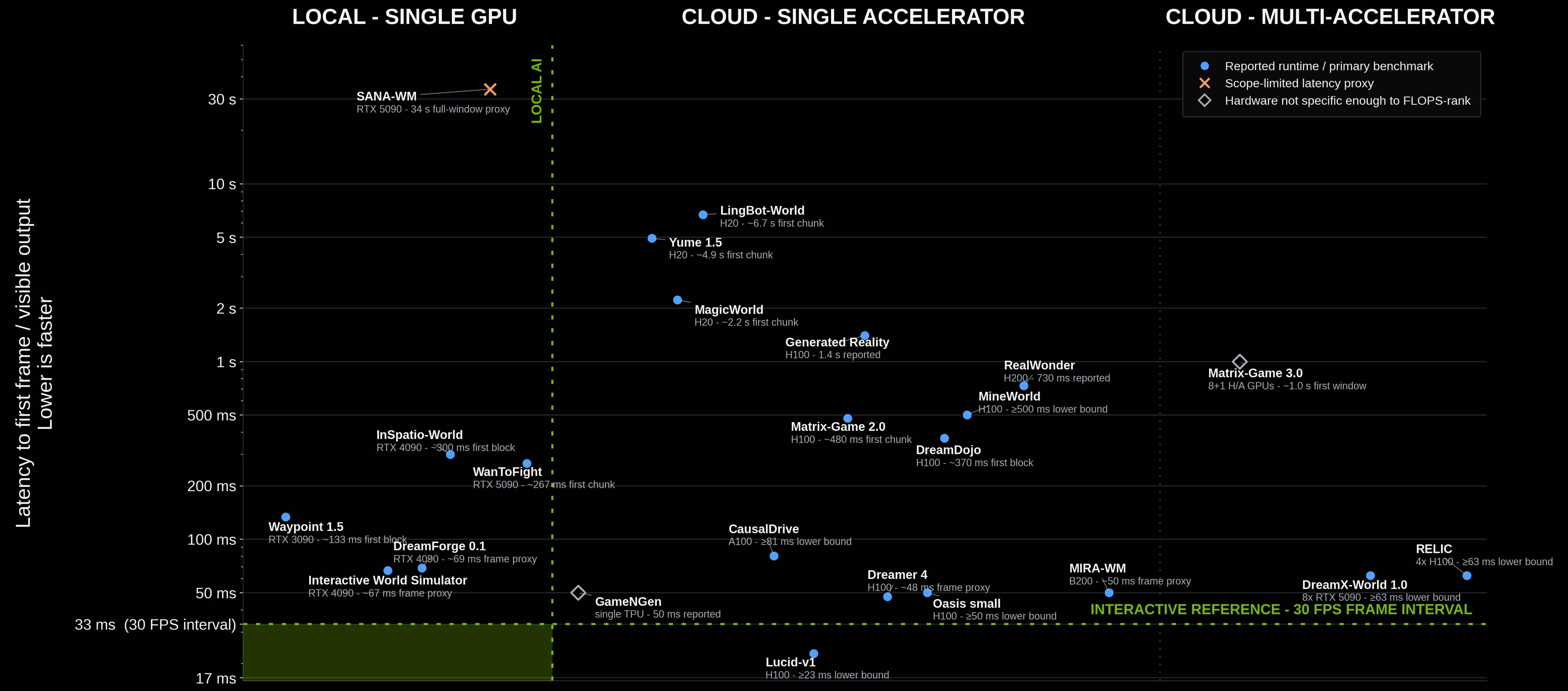




Reducing action-to-
response latency

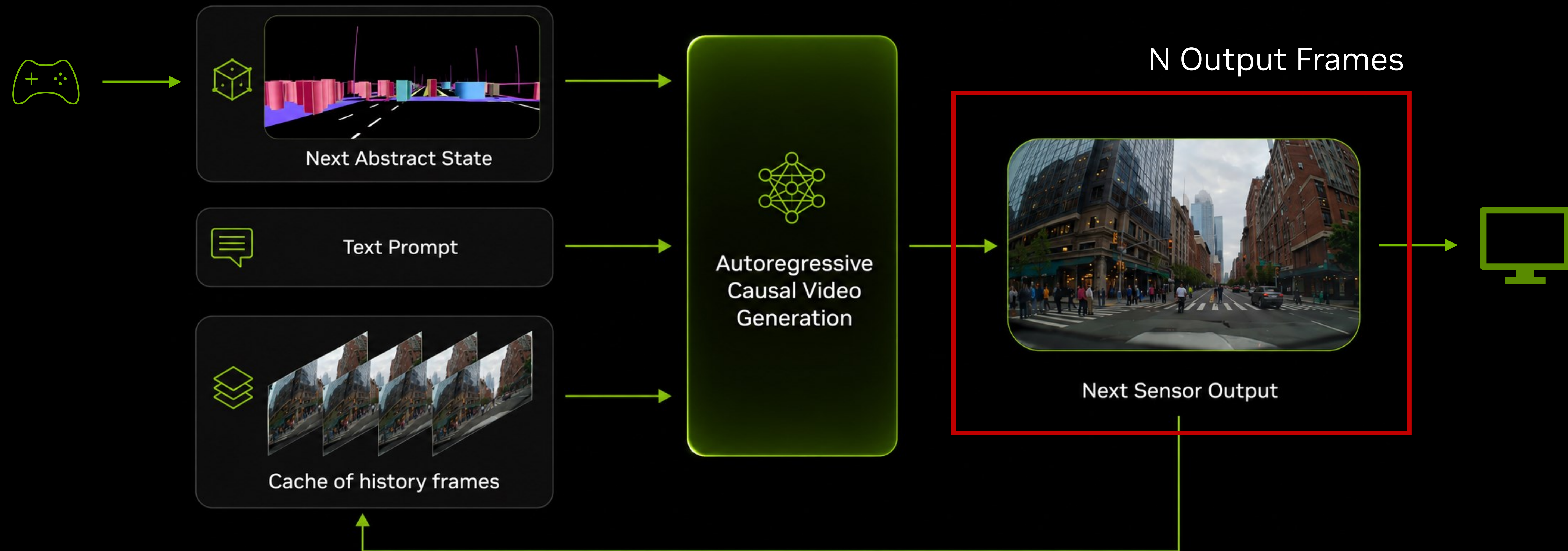


The latency situation is much more dire

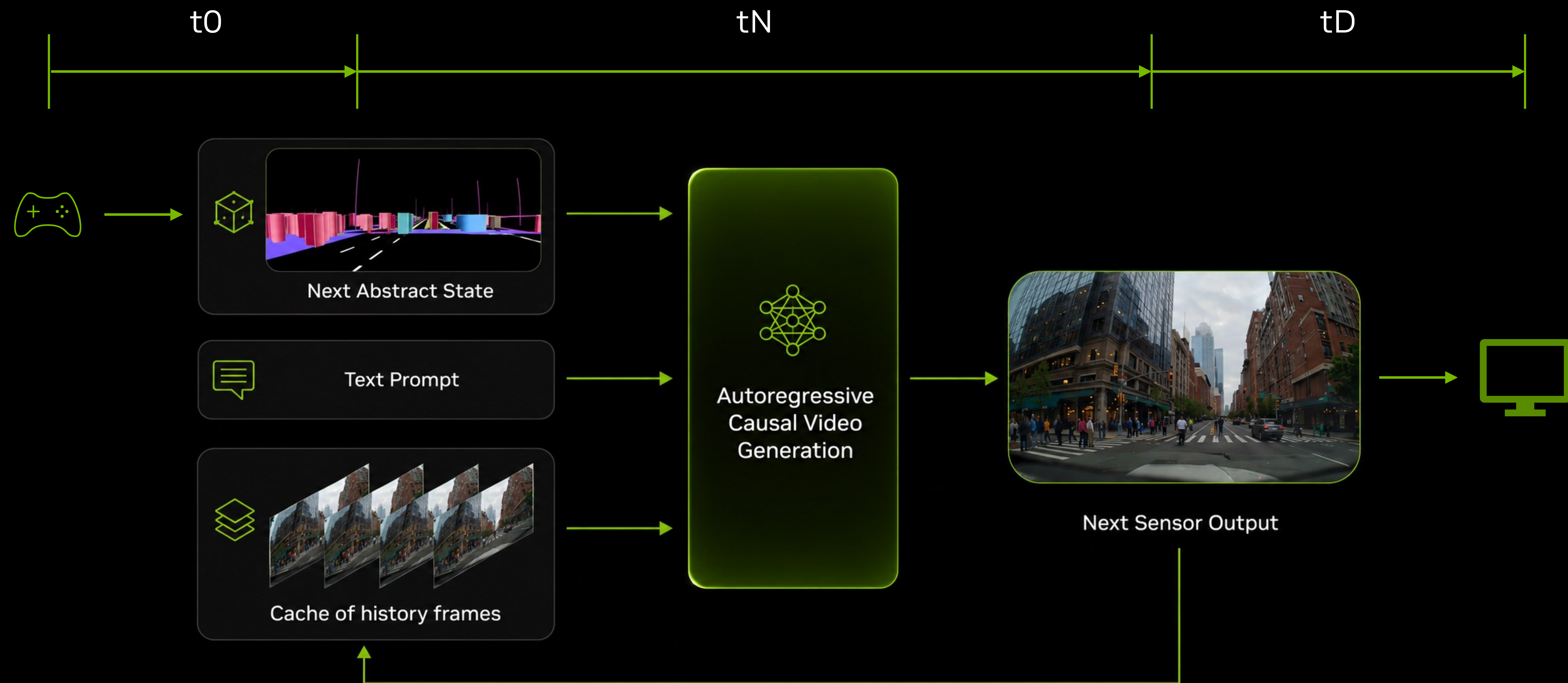


Reported action-to-output latency where available; otherwise a documented first frame/block/chunk proxy or one-frame lower bound. Cold start, display, and network overhead are excluded unless the source includes them.

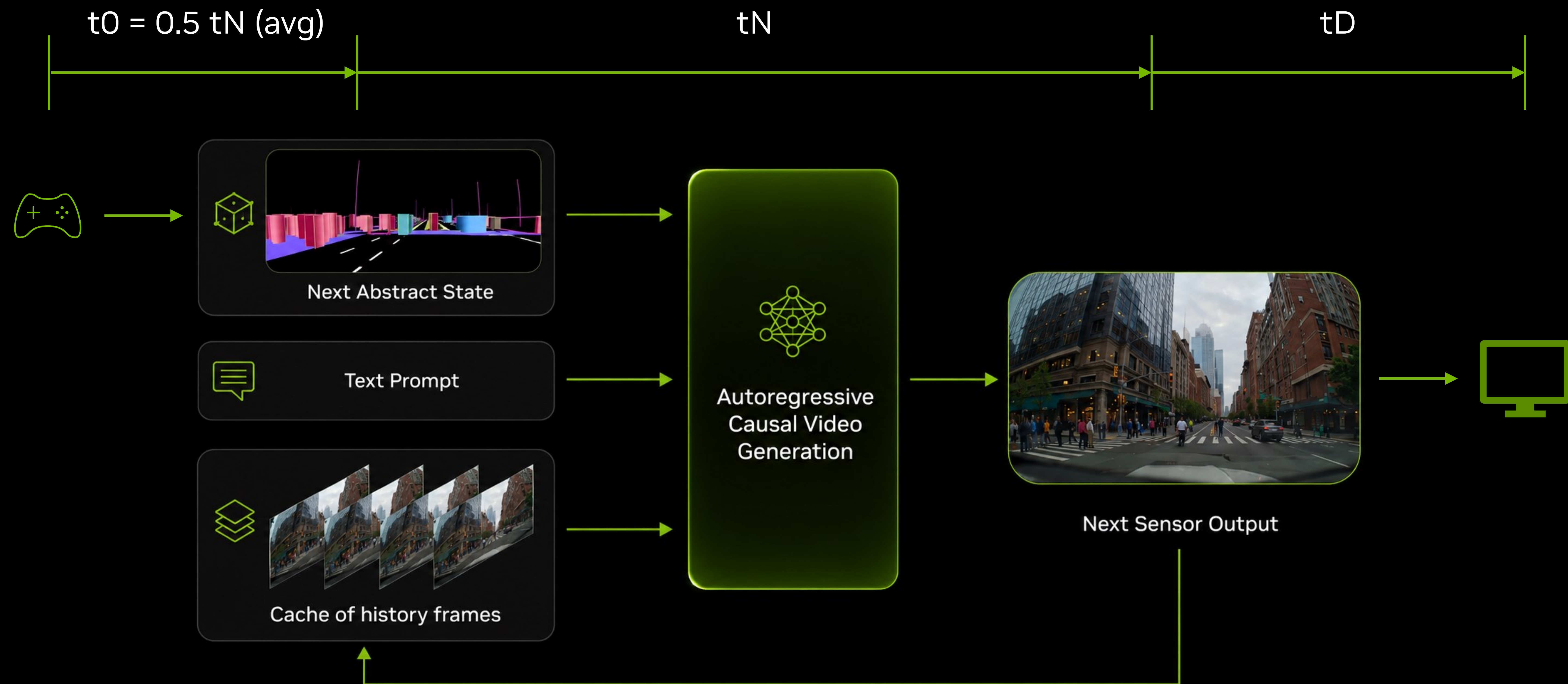
Revisit autoregressive video diffusion



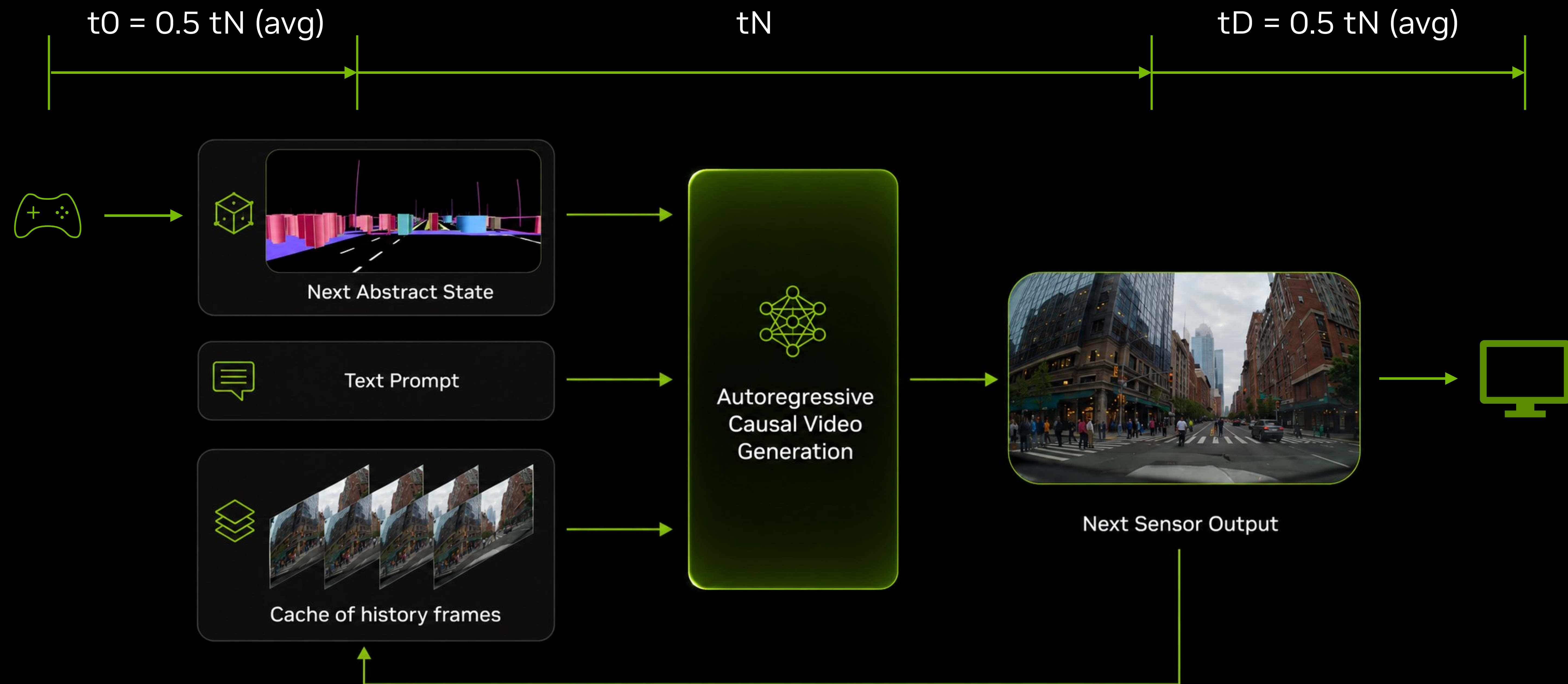
Revisit autoregressive video diffusion



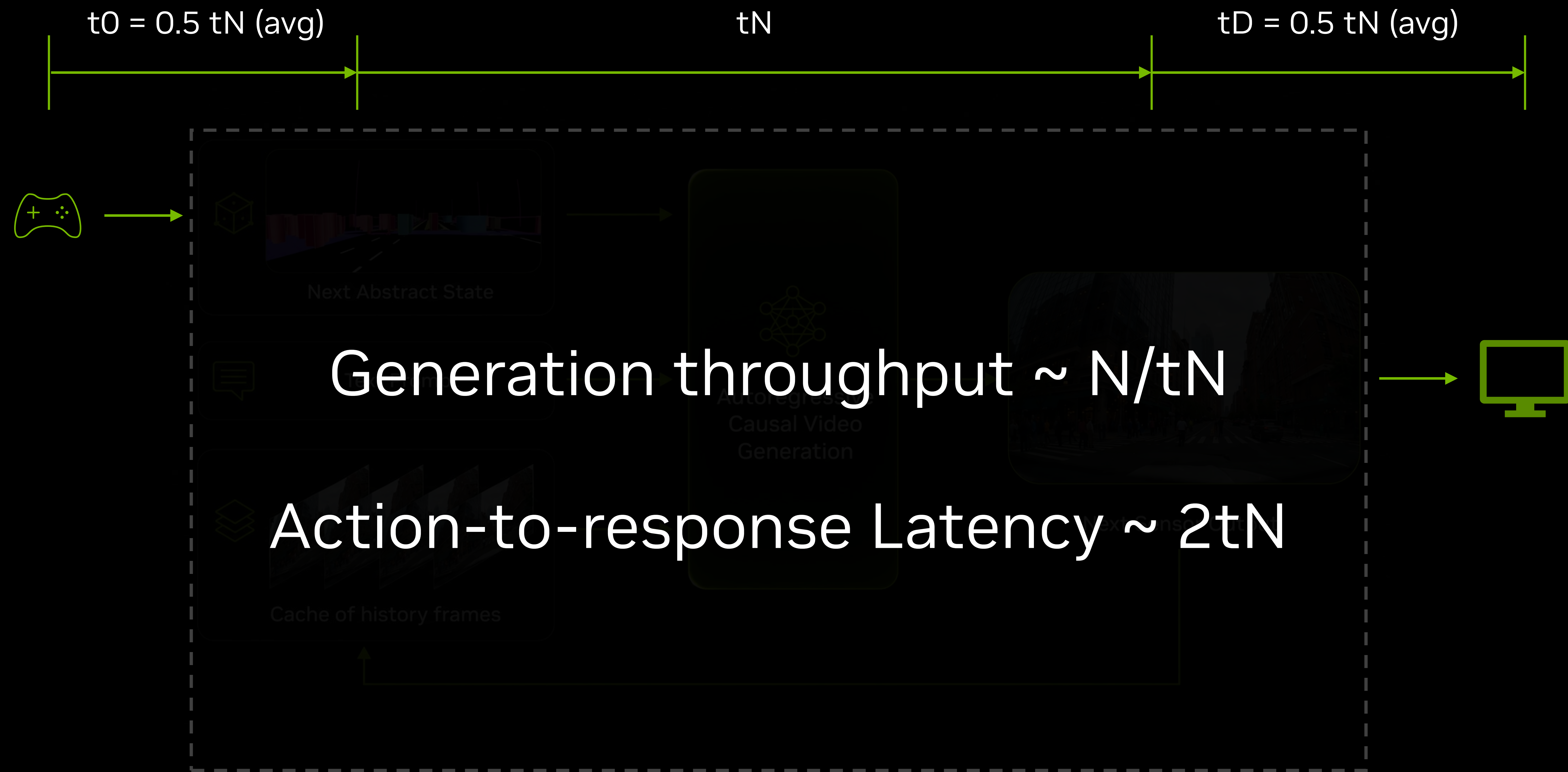
Revisit autoregressive video diffusion



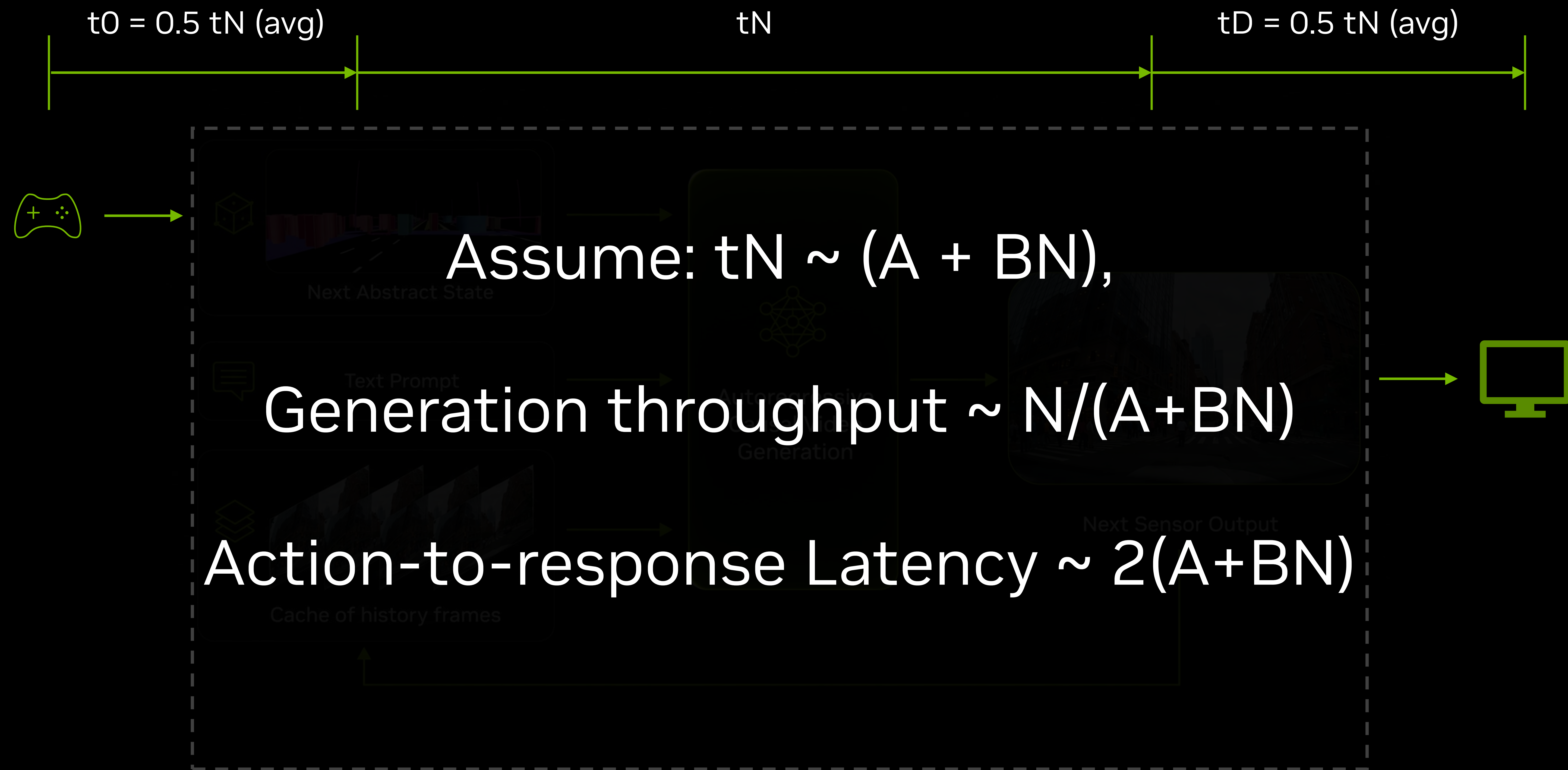
Revisit autoregressive video diffusion



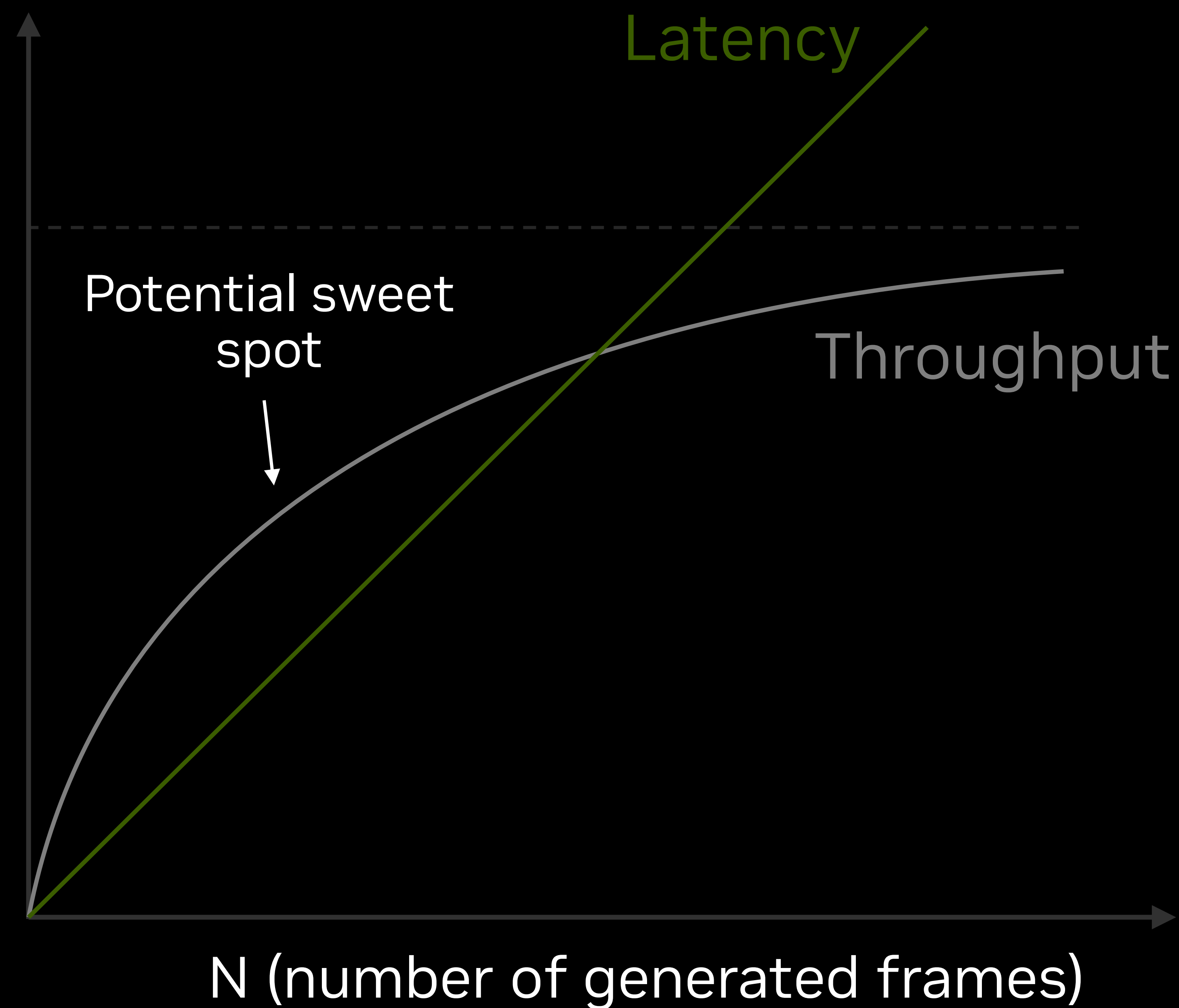
Revisit autoregressive video diffusion



Revisit autoregressive video diffusion



Not all throughput improvements are helpful for latency



An outline for joint optimization

Optimize A and B to improve both latency and throughput

Maximize throughput by increasing N while latency is acceptable

Summary and Conclusion

Generative world models will redefine computer graphics for video games

Achieving interactivity will be a fundamental challenge

- We are starting to achieve acceptable throughput

- Latency problem is largely unsolved, requiring speedup across the stack

Not discussed today:

- VRAM challenge and joint optimization of VRAM, latency, and throughput

- Startup/load time efficiency