

Yume1.5: A Text-Controlled Interactive World Generation Model

Supplementary Material

7. Vbench score in text-to-video

Model	Quality Score	Semantic Score	Total Score
Wan2.2-5B (20)	0.8401	0.7357	0.8192
Yume1.5 (4)	0.8470	0.7380	0.8252

We used Wan2.2-5B with 20 sampling steps (using CFG) and Yume1.5 with 4 sampling steps to generate videos for evaluation with vbench. We found that after training with our method, Yume1.5’s text generation capability did not significantly degrade while achieving substantial improvements in sampling speed.

8. VRAM usage statistics

Time (s)	5	10	15	20	25	30	35	40
Mem. (GB)	36.3	40.1	40.5	41.4	42.7	44.7	45.3	45.9

As demonstrated above, memory consumption scales sub-linearly with the number of frames.

9. Limitations

Yume1.5 still exhibits certain generation artifacts, such as vehicles moving backwards and characters walking in reverse. Performance tends to degrade in scenarios with extremely high crowd density. While increasing the resolution from 540p to 720p provides some mitigation, these issues persist to some extent. We attribute these limitations to the constrained capacity of the 5B parameter model; however, scaling to larger models would lead to prohibitively high generation latency. Inspired by Wan2.2, we consider exploring Mixture-of-Experts (MoE) architectures as a promising direction to achieve both larger parameter counts and reduced inference latency.