October 31, 2024

Oasis: A Universe in a Transformer

Decart, Etched

Try demo

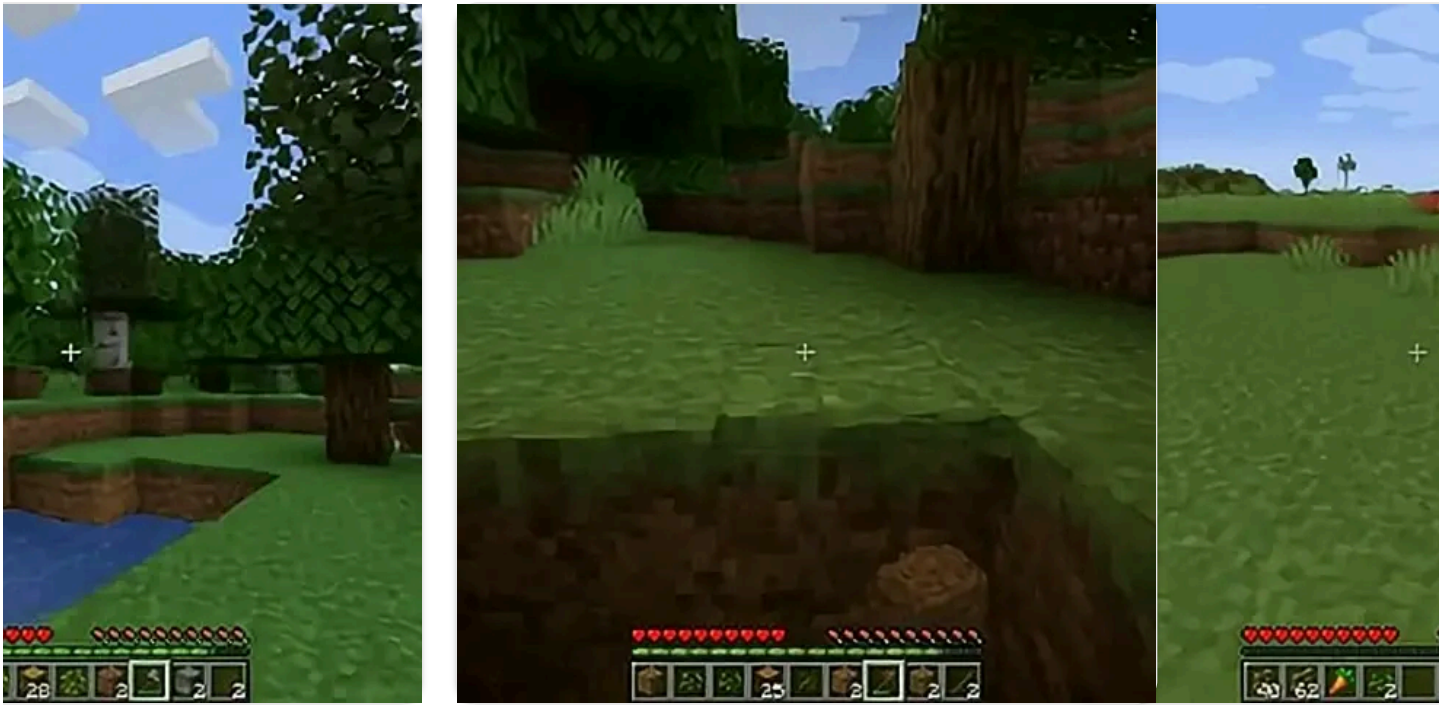
↗ Decart Blog | View code | Model weights | ↗ Etched Blog

We're excited to announce Oasis, the first experiential, realtime, open-world AI model. It's an interactive video experience, but entirely generated by AI. Oasis is the first step in our research towards more complex interactive worlds.

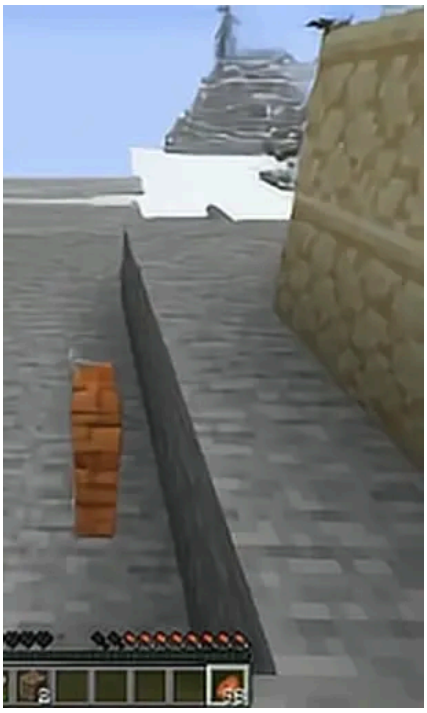
Oasis takes in user keyboard input and generates a real-time experience, including physics, rules, and graphics. You can move around, jump, pick up items, break blocks, and more. There is no physics engine; just a foundation model.

We believe fast transformer inference is the missing link to making generative video a reality. Using Decart's inference engine, we show that real-time video is possible. When Etched's transformer ASIC, Sohu, is released, we can run models like Oasis in 4K. Today, we're releasing Oasis's code, the weights of a 500M parameter model you can run locally, and a live demo of a larger checkpoint.

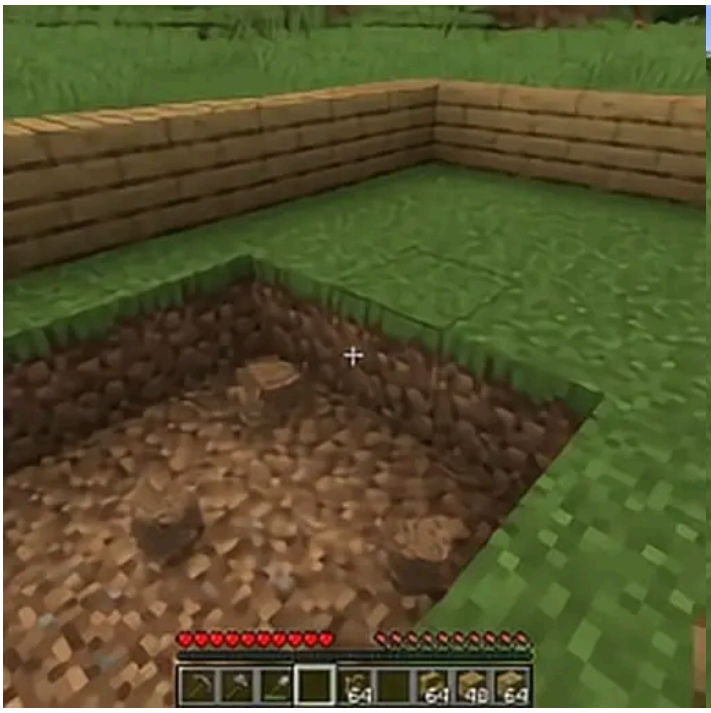
Results



Oasis understands complex internal mechanics, such as building, lighting physics, inventory management, object understanding, and more.



g health when eating



Shovel is faster than hands

Placing non-c

Oasis outputs a diverse range of settings, locations, and objects. This versatility gives us confidence that Oasis can be adapted to generate a wide range of new maps, experiences, features, and modifications with limited additional training.

rs open inventory chests

Exciting animals and characters

Space-like da

Oasis is an impressive technical demo, but we believe this research will enable an exciting new generation of foundation models and consumer products. For example, rather than being controlled by actions, an experience could be controlled completely by text, audio, or other modalities.

Architecture

The model is composed of two parts: a spatial autoencoder, and a latent diffusion backbone. Both are Transformer-based: the autoencoder is based on ViT[1], and the backbone is based on DiT[2]. Contrasting from recent action-conditioned world models such as GameNGen[3] and DIAMOND[4], we chose Transformers to ensure stable, predictable scaling, and fast inference on Etched's Transformer ASIC, Sohu.

In contrast to bidirectional models such as Sora[5], Oasis generates frames autoregressively, with the ability to condition each frame on user input. This enables users to interact with the world in real-time. The model was trained using Diffusion Forcing[6], which denoises with independent per-token noise levels, and allows for novel decoding schemes such as ours.

One issue we focused on is temporal stability--making sure the model outputs make sense over long time horizons. In autoregressive models, errors compound, and small imperfections can quickly snowball into glitched frames. Solving this required innovations in long-context generation.

We solved this by deploying dynamic noising, which adjusts inference-time noise on a schedule, injecting noise in the first diffusion forward passes to reduce error accumulation, and gradually removing noise in the later passes so the model can find and persist high-frequency details in previous frames for improved consistency. Since our model saw noise during training, it learned to successfully deal with noisy samples at inference.

To learn more about the engineering underlying this model, and some of the specific optimizations in training and inference, check out the Decart blog post.

Performance

Oasis generates real-time output in 20 frames per second. Current state-of-the-art text-to-video models with a similar DiT architecture (e.g. Sora[5], Mochi-1[7] and Runway[8]) can take 10-20 seconds to create just one second of video, even on multiple GPUs. In order to be interactive in realtime, however, our model must generate a new frame every 0.04 seconds, which is over 100x faster.

With Decart's inference stack, the model runs at live framerates, unlocking real-time interactivity for the first time. Read more about it on Decart's blog.

However, to make the model an additional order of magnitude faster, and make it cost-efficient to run at scale, new hardware is needed. Oasis is optimized for Sohu, the Transformer ASIC built by Etched. Sohu can scale to massive 100B+ next-generation models in 4K resolution.

In addition, Oasis' end-to-end Transformer architecture makes it extremely efficient on Sohu, which can serve >10x more users even on 100B+ parameter models. We believe the price of serving models like Oasis is the hidden bottleneck to releasing generative video in

production. See more performance figures and read more about Oasis and Sohu on Etched's blog.

Future Explorations

With the many exciting results, there come areas for future development in the model. There are difficulties with the sometimes fuzzy video in the distance, the temporal consistency of uncertain objects, domain generalization, precise control over inventories, and difficulties over long contexts.

ɹ precise object control

Struggles with domɑ

Fuzziness of distant sand

Following an in-depth sensitivity analysis on different configurations of the architecture alongside the data and model size, we hypothesize that the majority of these aspects may be addressed through scaling of the model and the datasets. Therefore, we are currently developing this direction alongside additional optimization techniques in order to enable such large-scale training efficiently. Further, once these larger models are developed, new breakthroughs in inferencing technology would be required in order to ensure a sustainable latency and cost trade-off. If you're interested in collaborating, reach out to contact@decart.ai and contact@etched.com.

[1]: Dosovitskiy et al., An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale

[2]: Peebles et al., Scalable Diffusion Models with Transformers

[3]: Valevski et al., Diffusion Models Are Real-Time Game Engines

[4]: Alonso et al., Diffusion for World Modeling: Visual Details Matter in Atari

[5]: OpenAI, Video generation models as world simulators

[6]: Chen et al., Diffusion Forcing: Diffusion Forcing: Next-token Prediction Meets Full-Sequence Diffusion

[7]: Genmo, Mochi 1: A new SOTA in open-source video generation models

[8]: Runway, Introducing Gen-3 Alpha: A New Frontier for Video Generation

* Estimated throughput figures - Sora reported, Mochi-1 from FAL.AI endpoint adjusted for parameter count, Runway from Gen-3 reported throughput

BibTeX

If you'd like to reference Oasis in your work, please cite it as:

```
@article{oasis2024,
  author    = {Decart and Julian Quevedo, Quinn McIntyre, Spruce
Campbell, Xinlei Chen, Robert Wachen},
  title     = {Oasis: A Universe in a Transformer},
  year      = {2024},
  url       = {https://oasis-model.github.io/}
}
```

Contributors

AI Team at Decart

Etched: Julian Quevedo, Quinn McIntyre, Spruce Campbell, Robert Wachen

Thanks to: Xinlei Chen