Stefano Ermon's startup Inception uses diffusion models to 10x AI response times

bizjournals.com/sanfrancisco/news/2026/04/23/stefano-ermon-inception-ai-diffusion.html

April 23, 2026

By Simon Campbell – Special Projects Editor, San Francisco Business Times

After years of artificial intelligence research, Stefano Ermon heads a company that could change the field.

Inception's approach stems from research Ermon pioneered as a Stanford professor. Unlike traditional large language models like ChatGPT or Claude, which generate responses token by token — essentially word by word — Inception's Mercury model uses a technique known as diffusion to dramatically reduce response times. Diffusion models have historically provided the underlying technology for AI video and image generation. Inception is applying it to text and code. The results have revealed a potentially dramatic reduction in response time and associated costs.

If ChatGPT works like someone assembling a jigsaw puzzle one piece at a time, Inception's system is more like a potter at a spinning wheel shaping clay. By refining the whole form at

once, the savings in time can be 10 times faster than rival services, according to Inception research.

Inception released Mercury, the first commercial-scale diffusion language model, in February 2025. Later that year, the company closed a $50 million seed round featuring Silicon Valley heavyweights Nvidia, Microsoft, Menlo Ventures, Mayfield, Snowflake Ventures, Innovation Endeavors and Andrew Ng.

What is the Inception AI founding story? We founded the company in the summer of 2024. We got some funding from friends and family and a small preseed round. We'd been working on core R&D for a while and eventually we figured out a way to get that approach to work at a commercial scale.

So no longer tiny little models, but something that was actually useful in practice and people could deploy to solve real world problems, build apps, build agents on top of it.

That was launched in February 2025 and that was the first commercial-scale diffusion language model ever released.

What was the big breakthrough that convinced you to start the company? In 2024, we published a paper showing that for the first time, diffusion-based language models were competitive with the traditional GPT-style models, autoregressive models.

We were matching the quality, but we were like 10 times faster compared to the traditional approach. That was the thing that really convinced me, "Yeah, there's something here. You really need to scale this up." And I basically recruited two former Ph.D. students (Aditya Grover and Volodymyr Kuleshov) who are also leading experts in this area, and we started the company.

How is what Inception is building different to OpenAI or Anthropic? Whenever people think about using an LLM to build an app, to build an agent, to build the customer support voice agent or a coding agent, they only care about three things: quality, speed and cost.

There is a certain fundamental limit to what you can do with an autoregressive model, the kind of thing everybody else is building. What this means is that when they ask a question, (the LLM) will give the answer and it will generate it one word at a time.

So it's very sequential, which means that it's kind of slow. And what we have is an alternative approach.

In May 2025, Google announced its own diffusion model, Gemini Diffusion. What is it like operating in a space filled with such large competitors? It's challenging, but also

exciting. We're probably working on the most important technology that humanity will ever build. I think it's such an important technology that's going to change the world in so many different ways, and being able to play a part and compete in the game is just so energizing.

At the end of the day, I think we have a lot of advantages as a small startup. We are very, very focused on our mission and this approach, as opposed to bigger players that have a lot more bureaucracy. We have the best talent in the world, people who invented diffusion models.

You've been studying AI and machine learning for decades. Why did you choose this career at a time when that wasn't such a popular choice? I studied electrical engineering, and then I went on to do a Ph.D. in computer science, specifically AI, at Cornell. Back then, AI was almost like a dirty word. Nobody was using it. Social networks were the main thing that everybody was talking about and excited about. But I stuck with my dream of building a machine that can think.

Why do you think interest in generative AI was slow to develop initially? The thing that always stuck with me was this idea of building generative models, things that can learn from existing math literature to generate new theorems, or look at existing images or diagrams and create new ones. So I was very early on working on what's now called generative AI.

Back then, it was called just generative models, and it was a very niche field. Back in 2014 people didn't believe it could actually work but eventually that's the thing that led to the things we have access to today, data, LLMs, generative models of images and videos. It's all rooted in those early approaches that people were trying out back in the day.

How significant was it for you to join the Stanford faculty? My group at Stanford made quite a few key contributions to the technology that are powering a lot of the state-of-the-art generative AI systems that everybody's using. In particular, one thing my group is famous for is basically coinventing diffusion models. It's behind essentially all the state-of-the-art systems that can generate images from captions or generate videos like Midjourney, Sora — they are all based on ideas that came from my group back in 2019. Everyone else was building a different kind of approach that we felt was not the right one.

We thought it had a lot of issues and we pushed for this new paradigm.

Are you surprised at the speed of progress in the AI space? When I picked this research direction, I thought it was going to keep me busy for my whole career. I thought it would be much harder to crack these problems. I thought that was the right way to get artificial intelligence to these really humanlike capabilities that were so hard to bake into a computer program if you were to organize things by hand.

What do you think about the pace of change and progress in AI technology? I've been surprised, maybe that's an understatement, by how fast things have moved and how quickly the field has made progress in this space. I never would have guessed that we would be where we are right now. Based on the results we were getting maybe 10 years ago. Just looking how far the field has come is pretty unbelievable.

What do you think about the future of AI and its impact on society? I often think about it. I think about it for myself, but honestly, I think about it more for my kids. Like, OK, what should they even learn at school? And what kind of jobs are going to be there by the time they finish school?

Even though I'm so close to this technology, maybe I should know the answer, but I don't. Things move so quickly, and I've been wrong so many times. I tend to generally be optimistic, but my concern is that things will happen quickly, and there will be sudden changes.

ABOUT ERMON

- **Born**: Italy
- **Education**: Bachelor's and master's, Università degli Studi di Padova, Italy; master's and Ph.D., Cornell University
- **Resume**: Stanford University
- **Hobbies**: Playing and watching soccer, a fan of Italian team Juventus
- **Favorite Bay Area restaurant**: Impasto by Terun, San Carlos

ABOUT INCEPTION

- **Founded**: 2024
- **HQ**: Palo Alto
- **Employees**: about 30
- **Total funding**: $50 million
- **Investors**: Mayfield, Innovation Endeavors, NVentures (Nvidia), M12 (Microsoft), Snowflake Ventures, Databricks Investment and angels Andrew Ng and Andrej Karpathy
- **Customers**: Several Fortune 500 companies