(no title)

61 MIN · YOUTUBE · HTTPS://YOUTU.BE/CBALU0DDEY8
https://youtu.be/CBaLU0dDEY8

SUMMARY

Andreas Plottman, co-founder of Black Forest Labs, discusses the evolution of visual intelligence and the development of models like Stable Diffusion and Flux. He emphasizes the importance of multimodal learning, where models can integrate various forms of data (text, images, audio) to enhance their understanding and capabilities, ultimately leading to advancements in AI applications.

- *The transition from unimodal to multimodal models is crucial for developing advanced visual intelligence.*
- *Black Forest Labs has focused on creating efficient generative models, utilizing techniques like latent diffusion to improve performance with less computational power.*
- *The feedback loop from real-world usage helps refine models, leading to improvements in capabilities such as character consistency in image generation.*
- *Open-source models allow for customization based on user preferences, enhancing their applicability across different contexts.*
- *The concept of self-flow is introduced as a method for aligning multimodal representations, facilitating better understanding in AI systems.*
- *The importance of human feedback in training models is highlighted, especially for aesthetic evaluations in content creation.*
- *The discussion touches on the challenges of data labeling and the need for quality over quantity as models progress through training stages.*
- *Plottman argues against explicit 3D representations, advocating for learning through natural representations like video and audio for a more flexible understanding of the world.*

Quick show of hands. How many people recognize a song that was playing? So, my favorite song is called Bella Napoli. It has been added to the uh CS153 Spotify playlist. For anybody who has music requests for CS53 this quarter, also known as AI Coachella. We've [snorts] got an open playlist. Please feel free to add songs there. That one was a request from me in honor of our speaker today who I'm very lucky to call a close friend and is the co-founder of Black Forest Labs, Andreas Plottman.

Thank you for joining us, Andy. >> Thanks, Ans. Thanks, everyone. Thanks for having me. >> Andy is joining us from Germany in a little town called Fryberg, which I think a lot of you will be hearing about more and more as it becomes a hub uh for frontier research in Europe. If you remember in our first lecture, right, we talked about the anatomy of frontier AI progress and we talked about three or four important touch points in this class. You're going to be hearing about over and over again. One is that there's a a transition happening from the old systems, the old infra stack to a new one, right? And you got to be open to understanding what those rewrites are looking like and and our speakers are going to tell you which parts of the stack they're helping to rewrite. We talked about the basic AI scaling recipe, right? We've got two sort of loops that are important to run. Once

you do you get some compute, you get some data, you build a model, and then you do inference, right? That gives you revenue to buy more compute and then context feedback. We've talked about the bottlenecks right on that on getting those loops scaling which is context, compute, capital and culture. We talked about context and and compute. We'll talk a little bit about all four today.

And then the last was well for your projects which is the the part I'm sure many of you are anxious about is how do you get one of those scaling flywheels going right? And we talked about there being sort of three steps in the journey. There's an incubation phase where you kind of figure out which specific part of the frontier you want to attack with a state-of-the-art system, right? Then you land with a soda release, a state-of-the-art release, and then that allows you to expand to more and more capabilities on the frontier that you care about. And if you remember, we did sort of a a field trip into one of the frontier factories, right? Um in in our first lecture, which was anthropic, we talked about code as one domain. And today we have a chance to do a field trip into another Frontier Eye factory in Germany called Black Force Labs. And we've got here one of the factory owners, Andy Blottman, um who's a co-founder of Black Force Labs, also co-creator of Stable Diffusion. How many people here have heard of Stable Diffusion? All of you. Perfect. Great.

You've done some homework. And so today we're going to talk about the frontier here. asked uh on Tuesday we talked about the the audio and the speech frontier, right? What is audio intelligence like? What was it? Where is it going? With Maddie from 11 Labs. And today we have Andy talking to us about the frontier of visual intelligence, which I think is one actually one of the most exciting frontier, if not the most critical frontier to unlock more progress in if we really want to get um these models to work in mission critical contexts in the real world. And so we're going to spend some time talking about the anatomy of visual intelligence as as Andy sees it as one of the pioneers of the field. And then we're going to talk go back in time a little bit and zoom into how we bootstrapped the flux flywheel together a couple years ago.

Flux is the name of the flagship model family from Black Forest Labs. And then we're going to spend some time on the fun part which is future frontiers. Where are things right now that where where are the unsolved problems? where are we right now where you guys can step in and start co-creating this journey uh in the space. So this was the frontier factory right we talked about this is sort of the basic template again to be clear this is a directional heristic every team is different every research project is different but to kind of give you a grounding sense of repeating patterns about how um some of the best teams are manufacturing intelligence repeatedly remember this was the pipeline we had um pre-training mid-training post-training with agents in the real world there's a version of this that that Andy is going to walk us through but before we jump into that why don't we just spend some time on on you, Andy. Who are you and how did you get here?

>> Yeah, cool. Thank you, An. Thanks again for having me, everyone. Um, yeah, I'm Andy. Um, started looking into AI, I think, in 2019. Um, I I was actually originally studying mechanical engineering. It's classic German education. Uh, I think you go to a school and then you figure out you're kind of somewhat technical and what are you doing if you don't know exactly what what to do? Studying mechanical engineering in Germany, right? Um and then it yeah through through a couple of um I think coincidences I got into computer science into coding into already robotics back in the days. We talk more about robotics uh later. Um and applied at a PhD in

H Highleberg uh where I met my two co-founders Robin and Patrick. Um and that was a really like small lab. Everyone back in the day was doing representation learning with visual models or like for the visual domain and computer vision in itself back that was 2019 was kind of a a niche topic in this niche topic back then of AI. It was really like people saw the potential already but but no no one no one had an idea of how how that would uh explode then later, right? So it was really a kind of a yeah niche topic we worked on but we soon had a very good intuition about like how to train models to generate pixels mainly images back then. Um and we're competing on a research level as a very small lab with players that were much larger than us.

Uh and finally that already back in the day was Google and OpenAI their research teams and it was not about building frontier systems was >> foundation models. Yeah. Or or even before that uh who wrote actually the nicest paper to show that something was was happening. So back in the days it was like >> this was pre uh >> that was pre-stable diffusion. There was that was >> it was for the ones who remember it style GAN was kind of >> the images were most often generated with GANs because they had a kind of a good inductive biases for for for kind of this data domain. Um and it was generating a 256 x 256 pixels image was a challenge like not every algorithm could do that and yeah it was just a very different world. So um we competed with labs that were much larger than us and we had even back in the day way way uh less compute. So we had to come up with kind of more um efficient algorithms to solve that problem because images and not speaking of videos are so much higher dimensional than other representations say text or something text is uh much lower dimensional >> and and to anchor focus on time you were still this was when you were at the University of H Highleberg.

>> Exactly. Right. >> Exactly. Um so um yeah and then we we spent like two years investigating how can we actually find representations for natural data for images for video um mainly that are perceptually equivalent to the pixel space or to what matters to us humans in the pixel space but much lower dimensional and much more efficient because we didn't have the computer to train a kind of generative model on the pixel space and it's also super wasteful and that was what gave rise to a a series of papers on latent generative modeling. So you actually train a kind of a compression model um similar to a learned JPEG code you could imagine it to find that perceptually equivalent representation to the uh pixel space and you train the generative model there and that um helped us saving tons of compute training our models much more efficiently and with orders of magnitude less compute than our competitors put out like better like models that were on power even better than those competitors and that was what that algorithm latent diffusion also gave rise to stable diffusion. Then um so we proposed the algorithm saw the potential set out to search some compute luckily find that in the open source community um and train stable diffusion that was then released in 2022 um and pretty much surprised us as well like with all the hype it got and actually was it was fun. It was here in the Bay Area it was hyped much more than in Germany. in Germany. Still today, not a lot of people know about that model, funnily.

>> Yeah, it that there was a moment I remember Dolly 2 was in preview, I think. And and then you guys put out stable diffusion and I remember on Reddit there was somebody had sketched out uh they had taken one of their kids' like uh drawings. It was like a crayon drawing and had turned it had run it through the image to image transfer on SD. Uh I think it's for SD1. Um, and it and out out had come this beautiful illustration and I remember taking a screenshot of that cuz I was just blown away and I tweeted it and I think it was like a Monday morning. We went into our exec meeting at Discord and then I came out for lunch and the tweet had like 3 or 4 thousand likes and it was it like a for me it was a moment where I realized that the technology of

generative modeling at that point had crossed an inflection point where it suddenly became legible to people outside the machine learning community because it was so visual.

>> Yeah. >> Right. I I think it it might be worth spending a couple minutes here to just take people back in time because at at this moment in time in the ML community I would say there was a bit of a dogma that language modeling was the beall and endall of intelligence. You know the general consensus at the time was that language is the interface to reasoning to to for intelligence which that the way humans reason about intelligence the way we think is through language and I would say that's a philosophical belief that has come and gone in its sort of the strength of its religious zeal. Um, but for those who are in the computer vision community, and I I count myself as one of those because my last company, as we've talked about, was Ubiquity 6. It was a 3D mapping and computer vision company. We were working on 3D reconstruction.

It it was clear that language was extraordinarily valuable, don't get me wrong, at at at reasoning about certain tasks and fields. But for those of us who are in the computer vision community, it felt incomplete because language is is just one way we communicate, one way we reason about the world. For those of you who are visual thinkers or visual intelligence who believe in multiple intelligences, right, you just learn better when you see visual representations of things. And so it was quite um cool to see stable diffusion coming out and make progress of a different kind legible to both the machine learning community as well as the the broader developer community, the broader consumer community and and that's when I think we we reach you know we started working together because we were trying to get some of these um stable diffusion like capabilities onto discord but can you talk about it? You know, you said two things that I think are quite helpful to overlay for the students here, which is the difference between natural and unnatural representations. Could could you speak about that for a sec?

>> If we think about ourselves, if everyone uh you look at me currently hopefully and and um the medium through which you are perceiving this is clearly video and audio, right? you hear what I'm saying and you're seeing me um gesturing here or or talking to an um so these are these are what we what I call natural representations. If you think about the source of those representations eventually it's the sun or here we have some some lights that try to resemble what the sun sun does but it's electromagnetic waves of a source that we humans cannot control. We can shape obviously the the or we we we can we can control the shape of this world and we can build buildings but this the electromagnet magnetic spectrum that falls onto the earth we cannot control.

Same for sound like natural signals like a hearing a river flow or something that's just might some might call it noise or something that's just natural and it's there. Whereas text is inherently humanmade. You see this in so many different um occasions. If you just measure the the information per sign that text transports, it's so much higher than the information per sign per pixel in an image. And why is that? Because it's humanmade. It was evolutionarily very important for us to communicate uh efficiently. Um, and that there's I think that that's also at the heart of why we need to compress images and videos before we train a generative model on them because there's so much redundancy in in it. In text, you don't have this redundancy because it's humanmade. Throughout evolution, we reduce that redundancy and um and made it efficient for learning. However, it's

super important at least in how I see it and how we see it at like Force Labs to consider two things. First, if you think about yourself as babies, how you learn, it's first observing things, hearing and seeing and then interacting with things in the physical world, right?

This is pretty much the first I would say three, four, five years. I don't know when I learned reading or but it's I think maybe with five or something. And just the level of intelligence a three-year-old has compared to the level of intelligence a a language model has is very different, right? And I think that's what what what why we care so much about natural representations like audio and video because we we are like absolutely convinced that this will be the fundament of all the kind of higher intelligence that these systems will eventually have and starting from language and trying to to stack up a bit of additional uh kind of representations on top of that is in my um kind of opinion the the wrong way. You should start with from first principles how we humans do it and that's clearly learning our natural representations by first observing and second we'll talk about that later interacting these are just from how we think about it the main pillars of learning and also the main pillars of what we define as visual intelligence actually >> so I I I think um so this is pretty important because two years ago three years ago I would say the consensus was that the way to do generative modeling was roughly this, right? Where you had this these foundation models that were unimodal.

They were text to image, text to video. >> Looking at stable diffusion as a stable image model. Exactly. Unimodal based on images. >> Yeah. Could you could you just talk about what this the what the state of the art was then >> versus now? >> Yeah. >> So, yeah, stable fusion. I think it's a perfect example of that. Um, it was a text image model. You could you could do super nice kind of artistic things that have not been possible before, but was clearly made for content creation, right? It was a a unimodal model made for the purpose of content creation. You could you could yeah make artistic style transfer. You could you could do you could trail Laura and and do maybe some some kind of um character consistent marketing transformation or like yeah character consistency into the get character consistency into the model then use it for marketing or something.

But that's all content creation. We currently see that visual models starting to become way more than that. We don't train a single unimodal model anymore to just like fulfill the purpose of content creation. We're training a a unified a uni a multimodal model for natural representations or natural data that then can give rise to so much more. It's about physical AI. It's about robotics, computer use. We can do with these models. We had a couple of demos already like or there were recently a couple of demos that were super impressive. We can do world modeling and simulation and still content creation. Um but combining different natural representations or only training on one is the key ingredient because it will give the model a much more natural understanding. As one example, if I if I just see two rigid bodies colliding, I always have a sound attached to it, right? There's a correlation between that sound happening and a certain action in physical in the physical world happening.

And being able to observe this correlation for a model is super important because it will help it that model understand much better what's actually going on. Whereas if I only train at at one single modality, it's much harder to to kind of understand what's going on or >> just interacting with with this bottle. I think it's super hard for a model to understand what's actually going on if it's not if it's doesn't hear that sound. How would that be different for that kind of transparent body compared to to someone um putting their hand through water or

something which is also transparent. Right.

>> Right. So um these correlations between different natural data representations are super important for a model to learn kind of at a higher um representation of intelligence as well. Now this this idea you know for those in the machine learning community is not new. I mean, for a while there was an there was a sense that the progression of technology would be we'd have sort of state-of-the-art systems that were capable at individual modalities >> and then at some point to make them smarter, we'd have to give them the ability to reason across different domains, you know, transfer learning, so to speak, where you can reason about the physics of a of of of this bottle hitting that and and the the sound, the audio emerging. Um but you can't start with everything on day one. And so could you just take let's let's talk about how we bootstrapped the fly wheel because now today fast forward two years ago you know four years after stable diffusion came out you know flux is now uh used by millions of people around the world you guys have hundreds of millions in revenue blah blah blah but for the purpose of the students I think it's helpful to zoom back in time and say okay you guys have this clear thesis for eventually models will be good enough at reasoning about all kinds of all of visual intelligence But you have to start somewhere.

>> Yes. >> Especially when you have less resources than than the largest companies in the world. You're a smaller team. You have less data. So can we spend a little bit of time talking about how did you concretize where to start? >> Yeah. >> And then how did you initialize that kind of momentum flywheel we talked about from at day one. >> Yeah. Absolutely. Yeah. I think uh that's one of the most important things when starting a company. focus matters >> well or any research project right at the time actually SD was not even a company.

>> Yeah. Yeah. Absolutely. Absolutely. But I think I want to as an example I want to I want to take how we started the company because there we we had this kind of huge experience in uh image generation unimodal image generation. We've done stability fusion then we've worked for stability AI put out a couple of uh more models on that domain and we pretty much had the recipe to kind of train a frontier model for that domain >> right >> so when we started the the the company we clearly or we looked at the field and we said there's clearly a need for a next generation of image models because so far the models cannot say produce hands that that are actually having five fingers right that that was back in the day a So we attacked that specific field and said okay we want to be building a model for for specifically for image that is just 10x better than everything else and that's what that then we sat down together 3 months we had all the recipes we knew what to do we scaled it and what came out of that was flux one that initially had kind of >> product market fit you can say we even before we took our API public we had a couple of very large customers uh that that kind of helped just closed the feedback loop. Now talking about the feedback loop because obviously on once you can build a technology but setting that technology out to solve real world problems will give you the very important kind of data to actually learn from first what is an important problem to work on and also how to make the model better for that specific problem. Right? By that you you have the first kind of >> uh loop closure for the flying. I mean, let's let's break that um that release down. Flux one, I think this is the kind of pipeline we talked about, right? So, could you just go through sort of the BFL version of this and explain what's going on at each step within the the company of the BFL sort of pipeline? Yeah. So I mean this this is particularly for um for how we would define visual intelligence now but I think I can also for flux one it was clearly we trained only on unimodal like like text text and image right only on those representations so the pre-training was just a large corpus of text and image

for the mid training we added um higher resolution um and like a couple of more capabilities into and then we had this kind of post- training phase where we exposed the model to first we did a kind of offline post training before you release an initial model. Do some distillation to make the model more efficient. You you align it with uh your intuition about what customers would care about, >> right?

>> And then you expose it to kind of the real world, but then you get this feedback. And for for flux one, a very interesting um observation we made was oh wow so many people are using our text to image models to actually train a Laura and then do character consistency. Like they they want they they want to they want to have the ability to control the model with more than only text because text is obviously nice and easy and low-key. Everyone understands it. Everyone can use it.

that it's also very ambiguous if you like and and again the there's a kind of disconnect between this kind of artificial representation text and and the natural representation image. So if I say an image of a blue bird, there are infinitely many images that that give rise to this kind of um description, right? The bird could be sitting on a branch, the bird could be flying and so on and so forth. And it's actually super hard to apply precise control to um image to to the image you want to be generating.

So that was that was I think that's a perfect example of the benefit of the loop closure because what we learned was okay people want to actually do image editing >> right >> so what did we do we did a post train on partially based on the data we got partially based on new stuff to create an image editing model which was flux one context that came out I think pretty exactly a year bit less than a year ago um and that was the first image editing model where you could actually in a scalable and fast way get character consistency. So I now I could take an image of you an say uh and maybe of me and combine us two sitting together not in a lecture hall but um in a cafe having having a chat and that's that just has massive potential for everything content creation right marketing needs it to to like get get different product different products into different contexts and it just like supercharged or currently supercharging like a lot of different applications around the creative world. Yes, this may not be ob obvious, but I want to pause here because because you know Andy did this quite naturally, but for those of you who who were trying out AI image models, let's say 18 months ago, how many of you tried giving it a photo of yourself and then saying, you know, give this person a hat and then it came out actually looking like you?

Yeah. No hands are going up. One one hand is going up. It was a pretty basic capability gap. these image models just didn't have character consistency, right? You just give it like a photo of yourself and say, you know, give him a mustache or whatever and out would come somebody looking not like me. And um that you know that that for many people in the space that was just a I actually can't I if I had a dollar for every time uh people would would say you know that that's just that problem will not get solved like these models are so dumb like look AI is so dumb it will never like it'll never be able to surpass that capability threshold.

Um I I I would just sit there and the and these are very smart people including by the way some of the speakers in this class who over the years have realized they had to update their priors about the the speed at which you can update these capabilities. But it was common consensus at the time that these image models are just not

going to get that good. You know AI is dumb. It can't reason about the the way the humans do humans do about create like you know faces and specific characters. And I was just it was I was shocking to me how very smart people would very confidently proclaim in the industry that I was just not going to get salt. Meanwhile, you know, here we were in Fryberg um looking at the data where there were people using flux one which was not very good at character consistency but that that context feedback of seeing how the prompts people were trying to use with the model and the the feedback of them saying actually that that was not good. can you please try you know doing this better?

the the the multi-step sort of reasoning chain that we were getting from seeing people out in the wild using it. And it was an openweight model which we'll talk about in a second which is quite unique gave us a very clear path actually to improving the capabilities and then actually it was you know one of our team members Dustin uh in in SF who figured out that you know we should just make an update to this that's that's called context with a K because that's the German way to pronounce it. um that is specifically an editing model and I think between the insight that insight which was at an offsite in Spain where where we were we were in >> I think it was in um in Italy >> in Italy we were in Italy you know uh I I think Dali no GP GBD1 image had just come out we were literally all together >> and there was a sense you know this is an important thing as a as a as a new team or as a first-time you know researcher it can be quite daunting when some lab that has way more resources than you launches something that's that looks way better.

And and your first intuition is to go, "Oh my god, we're But you got to remember that the mark of a good leader is to not panic. Keep calm, look at the data, assess the landscape, and then come up with a plan step by step." And often you'll notice that if you're good at at mapping the domain you're you want to be an expert in somewhere in your intuition, you have a gut feeling that's telling you there's actually some unsolved problem still left. And you know Dustin did a great job at that. The team rallied I think within 24 hours we had redone the the staffing on the team. And I think what 60 days later context came out. Yeah.

>> Right. And revenue from context doubled I think within six weeks. In fact, I think soon after is when um this part is public now that Meta announced a partnership with BFL and said they were going to be using Black Forest models. This tiny team out of Germany. I think the team was you guys were like 25 people. >> Yeah. >> To drive image editing for all two billion Facebook and Meta and so on users. I mean this is not normal, right? And and observing my I was I was lucky enough by this time, you know, I I had been an investor with you guys for about a year and a half and so we would go to these offsets together and I I'd had a chance to sort of see how in real time you know this this is not the systems problem often is not just a technical problem because actually all the data was available to us the context was available to us. It's the human system of of organizing the team and the research sort of culture, right? In a way that is not you're not panicking, but you're still assessing kind of the the frontier meth methodically and being very honest with yourself about how fast capabilities are moving and where you can uniquely sort of contribute to that system is is the key to keeping that loop going over and over and over again.

And I think that's why you know BFL is where they are today. They went from zero to several hundred million revenue. So it's the company is now worth more than 3 billion. But it can be easy to forget that that wasn't

always the case. And there are these moments in your journey, especially in machine learning where things change so fast. It's very tempting to just give up, you know, and say, you know what, this problem is solved. I mean, it really it is remarkable to me how many teams in image generation just don't no longer exist because they just gave up and instead BFL just stayed persistent and today is one of the only leaders left. I would say an independent leader that's pushing the visual frontier. In fact, I I hope some of the projects here push the frontier too. But um that's that's I think a learning for me has been it's actually quite straightforward sometimes technically if you have the right leadership to keep advancing the frontier. But sort of a drift in your mission, a lack of conviction that you know it's worth attacking the problem you you are committed to over and over again in the face of crazy challenges is often just the difference between you know success and fail. How many people know that have seen that meme of the guy tunneling and giving up right? Yeah, you guys know that meme. We we'll put it in the class reading list. I'm dating myself with with Boomer memes here, but I can't tell you how many times it's felt like that at BFL. And then one release later, right, the world has changed. And that that's actually I think also a good segue into what's next. Like now we're seeing this this kind of these models being applied everywhere for content creation. But again, then you need to think more like like look forward and and and and think okay, what's next? And clearly we now see this insane potential of especially these combined multimodal video audio image models for the capabilities we just talked about right physical AI um computer use right world modeling and simulation and still also content creation and you can you can actually build a single model that is capable of all doing all of that together and you will actually get compounding effects based on this example with with the correlation of noise and the action in the physical case. Um, you can also make models that are much smarter for generating norm regular images or videos footage for say again advertising or something.

>> Well, could could we could you talk a little bit about that? So let's zoom forward to um actually could could you talk about you know how do you take an image or a content creation pipeline and add to it what you just talked about the ability to actually interact with the physical world or learn from the physical world you know what what does action prediction mean how is that done and then maybe you can talk about self flow a little bit since we're going to be assigning that as reading. Yeah.

Yeah. Yeah. Absolutely. Um so yeah, I think first you you need to go from we already talked about this unimodal to multimodal, right? Um and then you get this kind of Yeah, this is this one. And if you if you go back to the slide here, I think this this is a good one. So there's a large pre-training on again natural representations. These are the representations we humans used to to learn from in our first years of of of our lives. Um and there you you just combine everything together and you have these um this combined pre-training that gives you a very very general model. So pre-training for us means images, video, audio combining with a architecture or with an algorithm that we've also published um beginning of March selfflow which allows the model to actually get compounding effects by observing again I don't make this this uh example again you you saw it uh already a couple of times by observing correlations that that exist between those modalities that gives you a very very general representation What we add next in mid training is additional context. We do new tasks such as conditioning on I can condition a model on an input image and an audio track and I say I I want to I want to hear an saying XYZ in that voice model does this. This is additional context.

But importantly for extending the scope beyond pure um content creation, you also want to condition the model on actions. And you want to have the model predict actions. And then we can arrive at models like this computer use models for instance that that are conditioned on a video or an image and they predict the next move based on keystrokes or something to achieve a certain task. Say I want to be opening a new brow browser tab or something, right? So this is crucial to get to expand the scope um of this kind of very general representation that we get from pre-training but we actually want to be using it for we want to make use of this kind of general generality of the representation. So we add additional context and importantly actions. Um and what we then do this is very important or yeah maybe maybe to zoom out a bit more and come back to the human learning example.

pre-training, mid-training, this is all still observation. All the algorithms that we're training like foundation models with in the early training stages currently are models observing examples. We we're calculating a loss from that. We back propagate that through the network, but there's no interaction whatsoever. >> Right? >> So, how do we actually get the model to interact really in the physical world? That's super important for kind of learning higher forms of intelligence as we are all uh convinced of. So what do we do? We use this model that can actually given a video predict an action to do something and hook it up in the real world on a say on a robot for instance, right? And then that allows us to inter like allows this model to through a robot interact with the physical world create data for that again and we can pipe that back into the model training and that's when we close this feedback loop. Uh so our post training looks or means interacting with the physical world right so this is important if you guys remember we talked about physical verification right as a key predictor where frontier progress is going to continue wherever you have context that can be and performance that can be verified progress can quite reliably be made there right so in software engineering that's verifiable because you can write unit tests in image generation not very verifiable right because one be beyond the basic tasks that that um Andy talked about which is accuracy right six five fingers instead of six character consistency which is more a preference uh >> but even that example how would you measure that at scale um without having a human telling you no exactly >> is that actually five five fingers >> well I think you should talk about how how that verification works and then What is the you know in the new world where you have to verify physical task robotics what does that look like?

>> Yeah. Yeah. So I I think in it's fun because if you if you Oh yeah. So verification for for for images are super is super tricky especially when it comes to kind of or for videos when it comes to physical things. But once you hook that up in the real world, there are just certain things that go that that that you can do and certain things that you can't do because a robot arm can cannot just just do certain certain joints. So it's like exposing it to the physical world naturally um applies the boundary conditions that we would expect. So that that's a very important step and by that you you have the perfect uh kind of environment >> right to to directly inherently model these kind of restrictions.

>> Whereas in the case of aesthetics or visual preference how did you guys verify that? How how did you get a model to be better when just doing content creation? >> Yeah. >> Uh that that involves like massive mass massive amount of human judgment and then uh feedbacking that signal through the model again. Right. Okay. But that's like of often very tedious and also often very dependent on who you're asking. It's like a uh if I ask you, you've looked at so many images by now. You're I would consider you an expert. Um because we always

show that guy our models, but you're also enjoying it. I guess >> depends on how much Betsy I've had that day, [laughter] >> you know, but but but like showing an image to someone who has no idea of like of like of of of say image generation versus to myself and I've looked at so many images, it gives you a very different signal. I would rate something as good or bad that looks very different from what a kind of a another person uh would do, right? It depends on the crowd who you're asking. So it's you can ask people that is very unam very ambiguous in a way.

>> So this is a key insight I would say because anytime the answer to the eval question of how do you verify is it depends on the audience or it depends on the person consuming the system. It should trigger a light bulb. At least it does for me. That the value that you get from the system varies a lot by how much the model can be customized for a particular audience. And that is where open source comes in because the beauty of open models is if you give away the weights and they're good general weights, right? Then you can tell Meta, hey, you're welcome to customize the preferences of what of this model as you see fit for your users. And you can tell another government that has different cultural preferences and biases that wants to, you know, be able to deploy content creation for let's say internal teams in a completely different culture and say you can have a control over that last mile. And I think that's turned out to be a very critical part of the open ecosystem where I often get asked an you know why did BFL open their models up and just give away all this research for free. Is it just that they want to save the world? Well, you know part of it is cultural as you can tell Andy you know Andy was a came from the academic community enjoyed and and benefited from open publishing but at the end of the day you've got to turn these research products into businesses right and it turns out there's extraordinary value in producing state-of-the-art systems that are then open and customizable when the consumer of the system, the person benefiting from the system has a very different preferences from other people who might be consuming the system. Does that make sense? Have I lost you guys? Can I get some nodding if that's making sense? Yes. Okay. This will be a theme consistently okay in the class which is anywhere you have consumers of a system or customers or the people benefiting from the system wanting more and more personalization customization of the system for themselves. That's where open models become extraordinarily valuable and you can actually build it turns out a very large business very quickly doing that.

So I actually think there's a false trade-off in the space a little bit about open versus closed. These are both just techniques or tactics for how to deliver value. They they some sometimes they get politicized philosophically but actually just from a very bas basic first principle of commercial perspective you know open makes a lot of sense in some domains where where the aesthetics the preferences and so on are quite there's a long tale of the distribution is quite wide and and and um and heterogeneous versus domains where you know preferences are actually quite narrow. If there if there's a pretty narrow distribution then I think you know closed models and so on are quite valuable in that case. I think there's one last piece that we haven't covered which is the state-of-the-art today because as Andy said now the state-of-the-art right is about how do you get these systems to reason in a unified fashion across text image video and so on in a way that's that has cross sort of transfer learning across these different modalities a very hard problem very very hard problem but as is the case with BFL consistently the team has you know makes these sort of research advancements and then gives the technology and so this this was actually one example I think what two months ago no a month ago >> month ago >> self flow which has turned out to be a technique now that suddenly magically all my friends at all the labs are calling and saying an have you heard of self flow I was like yes we have heard of self flow you it's on archive cuz

Andy and team published it so maybe you can talk a little bit about what the intuition behind self flow is as a mechanism to to solve the multimodal reasoning problem >> um so when training uh visual generative models it's always been historically um a bit tricky to get representations into the model where they don't only generate pixels in a way, >> right?

>> But also understand what's actually semantically going on. And there has been a body of work on um so-called or that that worked on aligning representations of generative models with um representation learning representations. Um that >> getting mad at there >> to to to to get them a bit more understanding of what's actually going on and not only make them stupid pixel generators that just learn to kind of um resemble what what looks consistent in an image. um that has been in the last two years always been focused on single modality. So what people did they used a pre-trained representation learning model like maybe some of you might know the dyno model for images um pre-trained model and they try to just make the representations that the transformer that is a backbone for image generation has internally aligned with the representation that this representation uh learning model um had and that is clearly restricted once you want to go multimodal but as you saw in the last slides going multimodal is super important to learn kind of higher form of intelligence. It's we want to be multi we want to have models learning multimodal representations because that's frankly what we humans uh have right so we it's crucially needed. So how can we actually combine learning having these so-called alignment losses with multimodal representations and the cellflow paper solves exactly that problem in a very natural way. Really recommended read for everyone. It's it's a sign super super nice. Okay, good.

Shall we transition to questions? Okay. >> Yeah. So, the question was um when when when we close the feedback loop, how do we ensure to to compress this? How do we make sure that actually uh personal data is respected um and um that no no harm is is generated based on those models. So first um we have a lot of content filters on on our API obviously because we our our belief is that these models are powerful tools for humans to to create super super nice and creative outputs and also much more than only content creation as we just saw. Um and we don't want to have them misused.

So we we add a lot of content filters that actually um make sure no harm is is generated um on the personal uh information. Obviously being being based in the uh European Union, we comply with the AUI act and there's actually a a kind of law that that um we also follow that you based on a request. So if if you put in a um a an image of yourself on our API and you say hey look I I I don't want to you you to you to um to store this kind of data we have to delete it. So we have systems in place to actually make sure this this basically happens. So the question is like we had a lot of partners um large companies that we worked with like XAI, Meta um backed by Nvidia and the the question was how do we evaluate um with whom we work and with whom not. Um I think maybe as a general statement we're working on building visual intelligence infrastructure for um everyone basically. So from an infrastructure perspective, you really want to make sure you put guard rails around your models um that people cannot misuse those. But then infrastructure is there for basically everyone, right? And and that that that's that's the the standpoint we're also taking. We care a lot about the safety of the models.

That's important and we do everything we can to prevent misuse. But then um I think it's also us provide putting out a technology there and providing it to to to everyone and it's always hard to take a certain

standpoint on like who you're working with, who you're not working with. Uh because it you get it gets very tricky to justify in the end. Uh >> let me try and translate what Andy is saying. [laughter] The company basically applies its guard risk to everybody.

So no matter who you are and how big you are and how much money you've got, if you want us to remove our guardrails, sorry, those guardrails apply to everybody equally because being a standard and being infrastructure that people can rely on means you don't treat different people differently [snorts] and everyone can rely that they're not getting, you know, just because they might have more money or they might be more politically influential, whatever it might be, that they can get the same quality of service as everybody else.

And so that's the position BFL has taken as an infrastructure provider is that doesn't matter who you are. Now sometimes you have custom needs because you have of the scale that are technical. Hey, we need it to be deployed in this way. We need some latency requirements that are more technical. But when it comes to guardrails, that applies to everybody. And so when some partners say we want you to move those guardrails, say sorry, you can go elsewhere. And that has resulted in sometimes the company losing meaningful amounts of revenue. And that's okay because in the long term as we talked about in the first lecture the way you get infrastructure to move stably is you have trusted standards and trusted institutions to enforce them and sometimes you got to enforce them yourselves.

>> Would you say that's roughly correct? >> Thanks sir. Yes. [laughter] >> We've had some we've had some spirited debates I would say at the company and you know we've talked about culture as a bottleneck on on progress. You know one of the most one of the secret sources of BFL is a very united culture. where there's a lot of debate and dissent on what to do and not to do but then when they commit they all commit together I mean how many people have left the company in the entire lifetime of the company like two >> uh um one >> one they've [snorts] had one person leave in the entire history of the company not common in the AI space where sometimes you have like co-founders leaving six months in I'm sure you this is the thing this is my one issue with the Bay Area the culture's forgotten that sometimes to keep to make progress on long-term ambitious goals, you got to stay together as a unit. And and that that's a great question. I think it challenged, you know, I think the culture at several points and I think they turned into uh sort of moes in a sense.

>> Yeah, absolutely. You debate, then you you disagree and then you commit. >> You commit. >> Uh and onwards then >> and there'll be more, I'm sure. Uh next question. Yes. question is how do we deal with um the insane amount of data labeling that has to be done and other than for for text uh images are just like not not not not that straightforward to label um I think two answers first when we train a model we start obviously from we just saw this kind of pre-training mid training post-training uh stages we start with more data and also more noisy data and pre-training and then we like as you progress through training you reduce the amount amount of data but you increase the quality. So for um in pre-training it's enough to do automatic automatic like labeling that you can automate and then really apply it at at massive scales. There are systems that that that are available to do this uh also publicly some but obviously also uh we have some internal stuff that I can talk too much about now. Um but then the more we approach later stages in training, the more we also involve um say human signals and stuff like that because you want to make sure as you say that in the latest stages of training where you

actually then again align this kind of very broad and general representation your model learns with what actually matters most to everyone out there. You want to make sure that this is actually you have annotations that reflect exactly what you want. And that's when you when still the the gold standard is involving uh human labeling. Then where do we see the the field going in terms of um denoising it iterate like just just in general iterative denoising? Uh is is it will it still be needed in the future? There are now other um probabilistic approaches that such as drifting models that allow us to do maybe a single step. Um and yeah, I'll answer that very generally. I think it's super interesting if you compare these kind of flow matching diffusion models uh with language models. Both are iterative. Both are iterative models.

But flow matching models or diffusion models are iterative in a dimension that is orthogonal to the data and this kind of time dimension that we artificial time dimension that we apply that goes from pure noise to to kind of the data you want to be generating. Whereas language models are iterative in the direction of the data, right? You generate token by token and that that has very interesting implications for both the training and the the inference um kind of properties these these models have. Um, for diffusion flow matching type models, you have you're actually pretty data inefficient because every training example gives rise to infinitely many kind of potential losses because you can pick every kind of um point on the continuous trajectory from clean image to noise and say I want to den noiseise from here to say the next step, right? And then I can do this super often.

So that that that tells us it's super data inefficient in a way compared to language models where we can train on all tokens par in parallel or let me specify language models a bit uh more autogressive models uh where we can train on all tokens in parallel on the other side we have at inference it's like the these two properties being switched so um or the effects of these prop two properties being switched when you see language models you have to generate token by token and there are some hacks such speculative decoding and stuff like that. um that maybe can can help you. But essentially you still have to like you cannot just miss data.

Whereas for diffusion models or flow matching models you can actually distill a model down right what we do when we do post training we do distillation. We've wr we've written a bunch of papers on uh adversarial diffusion distillation where you get down the kind of number of steps from flow matching models from 50 say to four or two and then it actually doesn't make a a real difference anymore if you if you then do a drifting model and you have this then directly at at one step maybe or you maybe take two steps but the pipeline is just more stable and mature when you distill a diffusion model down to two steps using adversarial diffusion distillation right so I think it's it's two things of the same uh yeah of the same side of the coin but coming back to auto reggressive models that that that's not really the like like possible for like get getting these insane speed ups by just distillation in in in the or using the iterative nature of the model that's not possible so I think a very interesting research problem that I'm thinking often how can we combine the data efficiency of auto reggressive models with the kind of inference capabilities or inference properties that these kind of diffusion flow matching type models have. So everyone who's who's doing who who likes to do research that that's a super interesting problem to work on. Uh, >> are you guys hiring >> and you Yeah, always always >> and yeah I could not I could not have spent the next half hour talking about this but we >> this this part is a the you know latent adversarial distillation is a very um it's a part of the the pipeline at BFL that I would say is is very near and dear to the to the core of the company not only for two reasons. One is because it actually makes these models extraordinarily efficient. And for those of you who have German friends, you know that efficiency is top of mind. And I think

that's that's that's a through line through everything uh BFL does.

It's high quality. It's efficiency. But it also ended up being a key unlock for our business model. Because early on, you know, a big question was well, we have this philosophy of we want to be open. We want to produce open weights, but we got to find a way to make it commercially sustainable. because there's a lot of projects that open models up and then they they just die and then that's not stable infrastructure either that you can rely on. And so one of the key differences between diffusion models and autogressor models as Andy's talking about is that a you know the the the model size is actually the same in a diffusion model in the if you look at the first flux family we we released we didn't release flux as a single model. Uh it was actually flux one was was um packaged into three different models. Flux uh Schnel which is German for fast, right?

Uh Flux Dev and then Flux Pro and Flux Pro we put behind an API whereas Flux Schnell was full Apache 2.0 open Weights and then Flux Dev was open weights but a commercial license where any you were welcome to look at the weights use the model but if you wanted to make revenue off of it you had to pay. And the key distinction between these three was actually they were the same size model unlike for example language models where you have flux uh like if you have claude haiku sonnet opus and so on they're actually different sizes. So in autogressive land you distill down the model you you train a big model then you distill down to smaller and smaller sizes. In flux one which is a diffusion model family it was the same size but fewer steps.

>> So you you can still do size distillation. You could say the size is installation but I think it was they were all at this point >> for flash one it was yeah yeah absolutely >> and so we we distilled it down to Chanel which was basically a singlestep model at that point so it's four steps >> four steps sorry fourstep model super fast super lightweight um lower quality pro more steps super high quality slower right because you're iterating over more diffusion steps and so that turned out to be this very beautiful kind of packaging of the core technology in a way that was also commercially sustainable because the open source developer community was was thrilled because now they had this really fast model for a lot of use cases that you can run locally and all the enterprises who didn't want to deal with customization had a high quality model that was behind API and developers who wanted the mix sort of a mix of both got a pretty high quality model that was also open weights that was fast and and that trade-off you know is is a hard one to make if if you don't sort of foresee the fact that you want to close this loop that we've talked about of Frontier Research arch repeatably.

You know, I would say 2 years ago the state of the art was train a model, put it out, put the weights out there, let's see. But when you start thinking long term, then you're not thinking in terms of a single model release. You're trying to think about it as a system. That's capital cont. You know, you know, we've talked about all the bottlenecks and you want one iteration to help you unlock the bottleneck for the next run and the next run. And adversarial distillation, laten adversial distillation turned out to be a pretty key unlock for that part of the bottleneck two years ago. Um, next question. spatial intelligence.

Yes. And whether it's more 3D or um or like how how I see the the the 3D space where some companies are working versus our kind of more video based um approach um going forward in the future. Um I think I'll I'll I'll take a a kind of opinionated uh view here. >> Again, it comes back to how how do we humans learn uh in in our

childhood? I think we don't have explicit 3D representations of anything in our head.

We just learn based on video and audio. And I think that's the way that that will or I I don't know if anyone have at least myself. I I I can I cannot look into people's heads, but I I don't have an explicit kind of 3D coordinate representation of something in my head. I just I see you. I have a kind of my my eyes are doing a bit of triangulation obviously, but it's still pretty much a a kind of projection onto onto uh those two eyes. Um so I think it's it's I don't have an explicit 3D model of of like say an this bottle or something else. I just learn it from video and I can obviously move my head around or I can interact with this object and that's all we need. So I'm not a I'm not a kind of um so like I don't believe too much in these explicit 3D representations.

It's also a data problem at some point. Do we get all the data kind of labeled with 3D representations? I don't know. Just learning based on natural representations and then interacting with it as I already said like before is in my view the the kind of way to go forward. I I I I'm going to I'm going to express a somewhat contrarian opinion to it and which is which is actually not totally disagreeing with you but I I would sharpen it to say I think what we have learned empirically is that these static 3D representations and I have a strong point of view on this because I I I started a company that failed at this [laughter] ubiquity 6 was a 3D mapping company we tried to map the world with uh reconstructions in 3D using a bunch of deep learning priors and what we learned was it actually don't get me wrong the technology worked on it is valuable, but explicit 3D representations are very narrow, inflexible, and static, especially when you take the temporal the the time element out of it. Their uses are quite limited and niche. And so there are these applications for 3D representations that are useful, don't get me wrong, especially when you're getting human like machine perception.

You know, point clouds, for example, are great at having robotics, you know, do pos indoor positioning in GPS denied environments. But when it comes to when you're trying to build a system that can be interacted with by humans, it turns out point clouds, 3D meshes, the these are all sort of intermediary representations and in a sense hacks to you know that that are less general and flexible than representations that can integrate naturally time and audio and all these other modalities because that's how we reason. Now I because I actually disagree with you Andy. I do think I have a 3D representation in my head of certain things like I if I if I >> but is it explicitly are you thinking in in in terms of coordinates? I don't I don't think so.

>> Well, I think about this lecture hall. >> Yeah. >> Like when I'm planning a lecture, you know, I often have a 3D representation spatially, but it's it's just one input representation as part of a broader set of >> but but you're not like that there's there's no kind of prior that enforces you to to to think think about this uh kind of explicitly. It's implicit. You learn it. You learn it based based on like what you're perceiving and what you're interacting with.

>> That's right. >> And then and obviously we could we could maybe maybe a network could in in its weights represent this kind of uh implicit 3D structure if it needs that. >> But I think the interfa like we're talking about the interface, right? Do we do we need these exp? >> Uh okay. Yes. At the human interface level, no. It's it's it's quite unnatural to reason in in explicit 3D prayers. I would agree with that.

>> Welcome to our late night conversations. [laughter] >> Yes. Thank you so much Andy for coming. Thanks for having me everyone.