The pivot that paid off: How fal found explosive growth | Gorkem Yurtseven (Co-founder and CTO)

59 MIN · YOUTUBE · HTTPS://WWW.YOUTUBE.COM/WATCH?V=E_AG1C4W138
https://www.youtube.com/watch?v=e_AG1c4w138

SUMMARY

Todd Jackson interviews Girkham 7, founder and CTO of FAL, a generative media platform that pivoted from a data infrastructure focus to AI-driven image, video, and audio models. Despite initial challenges in raising funds and transitioning from a working product, FAL has seen explosive growth, achieving over $100 million in ARR within a year by serving 2 million developers and over 300 enterprises.

- *FAL pivoted from a data infrastructure product to a generative media platform after the emergence of stable diffusion models.*
- *The company grew from $2 million to over $100 million in ARR in just one year, raising three funding rounds in 18 months.*
- *Their focus on developer experience and easy-to-use APIs has attracted a wide range of customers, including enterprises like Adobe and Canva.*
- *The company emphasizes the importance of optimizing GPU usage and model deployment to handle a diverse set of models efficiently.*
- *Girkham highlights the significance of building a strong sales team early and securing annual commitments from customers to ensure revenue durability.*
- *FAL's marketing strategy is tailored for developers, utilizing memes and community engagement rather than traditional marketing methods.*
- *The team culture promotes collaboration without formal engineering managers, fostering a hands-on approach to problem-solving and innovation.*
- *Girkham sees the generative media space as a net new market with significant growth potential, differentiating FAL from competitors in the AI landscape.*

We even had paying customers. We did both products for a while at the same time. We tried to convince each other like the pivot is not as drastic as it actually is. Hey everyone, it's Todd Jackson. I'm a partner at First Round. For today's episode, I'm excited to sit down with Girkham 7. He's the founder and CTO of FAL, a generative media platform for building with AI image, video, and audio models. I invested in foul seed round in 2022, but back then Girkham and his co-founder Burkheai were building a completely different product for data teams.

>> We were doubling down on this data infrastructure. He thought at the time compute was going to be big. >> But when stable diffusion emerged, they made the tough call to pivot, walking away from paying customers. >>

We tried to explain this to people because it was so new. No one got it. We actually had a hard time raising our series A. They made these models run so fast that when they showed me a demo turning my face into George Clooney, it literally looked like a video. Since then, they've grown from 2 million in ARR to over a 100 million in just one year, serving 2 million developers and raising three funding rounds in 18 months. Investors criticized us that this revenue may not be durable. So, we took enterprise sales very, very seriously. In our conversation, we dig into their pivot story, how they stay ahead in this fastmoving market, and what's next for AI video. Traditional marketing doesn't work for developers.

People think it's cringe. We just jumped onto the meme and created these two hats. We ran out of the GPU poor hats like way before the GPU runs. Let's dive in. [Music] >> Girkham, welcome to the show. >> Thank you so much. Thanks for having me, Todd. >> We're going to dig into a bunch of different topics today, but just to make sure everyone has enough context. Could you start by explaining at a high level what FAL does? FAL is a generative media platform for developers. Uh we host image, video and audio models as easy to use APIs using our inference engine behind the scenes and other developers working at other big companies or solo developers build products on top of these easy to use APIs.

>> You know, I think that you guys really pioneered this category of of a generative media platform >> and it's just been an incredible year. I know, you know, over 100 million in ARR now, over 600 models, 2 million developers, 300 plus enterprises, and a series A and a B and a C like all in the last 18 months. It's been a crazy year. Can you share a little bit about what the business was like this time last year, one year ago versus versus today?

We thought we were moving fast, but we were only at like 2 million error maybe in in August. And we had we had a slow summer. Uh I think the in general the space the image model space had a slow summer and it was right after stable diffusion XL uh was was released in April I believe and then for the whole summer it was very very slow and there was basically no new model releases. Everything was getting cheaper and cheaper like our usage was was growing a lot but because you know everything was older people were able to run them more efficiently including us. So our revenue was hovering around 2 million. And then first the flux model family was was released. That was really good for open source and and for us and then many other models like right away were released. Uh we had you know a really good launch with flux models and then video models that that's really big for our business. I think around October was the first video model that was commercially available in the API. Since then we've been we've been growing has just been crazy since then. Okay. Well, let's let's go rewind the clock a little bit because >> one of the most interesting parts to me of your path to product market fit that I think not everyone knows about is that you you know you and Burkai started off with a completely different idea. So, I'd like to go all the way back to the beginning and and just sort of tease out some of the lessons from the pivot that you guys did.

>> Um, can you just tell the story of of how you met Burkai and how you decided to work together? >> I met Burkai when I moved to San Francisco over 10 years ago. Uh we are both from Turkey and we went to high school there and came to the US to to study in college and then we both moved to San Francisco right after college and we were working at you know big Silicon Valley companies completely independent from each

other. So we we met socially by just being in the city being Turkish together through through common friends and we actually never worked together until fall. He was working at Coinbase at the time. I was at Amazon. This was this was during COVID times. We went to Palm Springs together when everything was in a lockdown in San Francisco and we rented the house there, stayed there for a couple months just so socially. We were we both had our own jobs and then you know we started like discussing some ideas and that's when like the the first seeds of of fall was planted and then it was like maybe seven eight months after Burkai quit his job and then shortly after I followed doing the same and we we kind of went around the idea maze tried to do like something in the machine learning space. We we knew building for the future, building for developers was was always the right track, but we had to do some exploration around like what ideas that might catch on and like we had to get a team of I think at the time five people to you know try some things open source tried to work with some of the enterprise companies that we had connections with and we were doubling down on this data infrastructure area because we thought at the time compute was going to be big.

We were following footsteps of data bricks and and snowflake. They were able to like monetize compute in the cloud very well. Turns out I our bet was correct. In a way we are still doing the same thing. We are doing compute in the cloud but the workload we were targeting was specifically for data transformation in in big companies where there's a lot of data so that people can transform this data to use it either for AI in in the past and and maybe analytics. It was first Delhi 2 and then stable diffusion and then chat GPT and then LMA 4. When all of these things released within months of each other, the the whole AI world kind of came backwards in a sense that you don't need data anymore to actually train these models. The models are already trained for you and all of a sudden like okay, you can do a lot more with an readymade model for you. So that's that was the spark for us to go for the pivot. If there's a readym made model that changes everything, this whole like data preparation stage can be skipped and only like the biggest of the companies are going to do that. Everyone else they'll just use something off the shelf and that attracted us towards inference.

>> The thing I think is really interesting because I first round invested in 2022 on the original idea, the product for data scientists and one of the things I think is so interesting is that that idea was in my memory was was kind of working. It's interesting to pivot to something else when when the product the initial product that you have is actually kind of working and you had customers. Was that a hard thing to do to to to walk away from something that was working to something else that you thought might be better? We had customers. We we even had paying customers. We did both products for a while at the same time. It's very hard when you are not screaming exactly what you are doing to to your customers, to the potential customers, to people you work with. uh it's really hard to sell because you know they look at your website they see something else like so it is really hard to do both of them at the same time probably did it for like two months or 3 months uh maybe 2 months we saw the the revenue growth for AI influence was growing much faster than data transformation so we decided all right maybe it's time to say bye to our old customers and double down on inference >> I mean is that a psychologically hard thing to do when you you have users, you have customers, you have investors who invested in that original >> everyone knows us for something, right?

There's there's the social aspect to it as well. Uh yes, it was it was very challenging. >> What was the moment where you just sort of said hey and and did you both kind of agree at the same time like we got to just do the new thing? >> We always tried to convince each other okay maybe that's this is not a huge difference. Maybe

this is not a this is not a big big pivot we are doing like we are still doing compute in the cloud.

It's just a completely new work workload. So maybe it's not that bad, right? Like we tried to convince each other like the pivot is not as drastic as as it actually is. So we we lost some time actually in in that phase where okay maybe this is not a pivot maybe we are just changing direction a little bit. >> So if you could go back to that moment before the pivot and kind of knowing what you know now what would you tell yourselves or or what advice would you give to other founders? So one thing actually you told us resonated a lot with us. You said okay which of the idea you think you're going to reach 1 millionaire first and then which of the ideas you think you're going to reach 10 million error first. Right. Uh we remember this very well and >> and the data science idea was faster to a million but the generative idea was faster to >> faster to 10. Yeah. And so it was like okay it's interesting but but then we actually reached 1 million with inference and 10 very quickly with inference as well. So our predictions were wrong but the framework you you told us actually helped us a lot to in in that decision-m process. The other thing I was thinking about is like when you first saw this idea where where did you see the opportunity like what was the problem that was going unsolved because at the time there were other inference providers there you know there was together AI or there is together AI there's base 10 there's a bunch of them they were focused on on language models and text what gave you the conviction that there was like a big enough other marketing media there is obvious value we saw in inference because all of a sudden you don't need a lot of data yourself to train a model you can just use something off the shelf And this all of a sudden increases the number of AI product users by we thought maybe it's 100x thousandx maybe it's a lot more than that maybe everyone becomes a AI user because it's so easy to create an AI product that value was obvious and we thought okay this changes everything then within that idea we had to okay why why stick with image inference and not do LM inference um so I think you know that wasn't easy as Well, image came first, right? There were there stable diffusion came before LMA 2. All the inference providers before some of them were doing other things as well like Bay Stan was was also doing more traditional machine learning. So, a lot of people saw the opportunity around the same time than us. Maybe they were able to pivot quicker than us. But either way, our first 10 customers or something like that were trying to build products on stable diffusion. So we had all of our customers doing that. But also within the stable diffusion world, we had to make some decisions as well like are we going to just provide GPUs for people to for them to just deploy any workflow they want or are we going to build easytouse APIs for them and they're just hitting an API point endpoint. This was a big dis discussion within the company and we ended up building an API endpoint which which we called it like an inference endpoint rather than doing GPU infrastructure. So first we we got over that hurdle and went with APIs and optimizing the inference process and then we were like so good at that uh we decided to stick with optimizing image inference because LM inference was way different at the time. There were different technical trade-offs. We we thought the buyer is is very different.

The market is shaping up to be different. Obviously, everything was happening so fast. >> You thought it was a very different customer, different customer, but also a different set of optimizations to do behind the scenes. >> Correct. Correct. And around that time, you were actually okay maybe it's it's a good time to start raising our series A. And we we tried to explain this to people because it was so new like no one got it. Everyone thought an inference platform is an inference platform.

Doesn't matter what kind of model it is. There are other people who are more qualified or more prepared to do

this than than us. And we were trying to explain them how our focus on at the time there were no video models. Our focus on image models is going to actually make a difference. And people thought the market was smaller as well. So all these things on top of each other, we actually, you know, had a hard time raising our series A. It wasn't it wasn't that easy. And the the fatigue was uh because all these big inference providers had raised maybe like a month before or around the same time we were raising and all the investors they were pitched the same thing over and over again and it sounds the same and that's something we underestimated how how disadvantage o of a situation it is to to raise at the same time with seemingly all your competitors because they all say the same story. investors hear it over and over again and there's some fatigue of hearing the same thing, but you guys definitely had a lot of conviction that the actual stack that you needed to build, the problems that you need to solve were very specific to what you were doing. So I cuz I remember in 2023 when you came to our office, the first time I had seen kind of this new idea and I think it was on stable diffusion SDXL or something where it was on your laptop where you had this, you know, through the webcam this video of me like turning me into George Clooney and like waving my hand. It was like a video, but it was actually like frame by frame images that you were generating >> and it just blew my mind how fast it was. What were some of the things you were doing behind the scenes to make that possible?

>> That demo is when we announced this new iteration of the company to the to the rest of the world. >> Still, we didn't make any money from that demo. So it makes a really impressive you know technical demo for people to see how fast we can run inference but we couldn't find any any use case for that like very fast image to image inference. Still to this day it is very impressive but we couldn't find any commercial value with it. There was commercial value in image inference just in general but that was very good for marketing but not so good for actually monetizing.

>> How did you get it to be so fast? At the time we had a twoperson inference team that they you know Banan our our VP of engineering now he has a compilers background he loves optimizing things at a systems level and then we had another engineer who who was really into like writing Triton kernels. So we looked at the program like what are the parts we can optimize how can we run the program more paralyzable and you know me the other engineer we have Burkai we all came together and obsessed over it over a couple weeks and we were able to at the time it was like SDXL and the uh SDXL distillations that's that was a distilled model there was one model that we were able to optimize and that was it that's what people used and to build products. So that was it.

>> I and I remember a bunch of uh a bunch of stuff on Twitter like you know some of some of the demos that you said you know kind of went very viral. But then let's get to the first actual like product when you started you know releasing and had developers using it. What was the first version of the product that you were giving to developers and what were the shortcomings of it like cuz yeah anytime you you launch a brand new product usually it can be very good at some things but probably not good at a bunch of other things. What was the first version like? I mentioned this before instead of focusing on like GPU orchestration and letting people deploy whatever they want we decided to build you know APIs every single uh code that's deployed is owned by us and we control the whole process uh and that becomes okay if the API can can do what you're trying to do yes you can do it if not you you are restricted by what the API can do so there were other competitors they they were allowing any workflow, any code to be deployed.

So that was a big shortcoming. But we we knew what people wanted to do. Everyone wanted to do the same thing over and over again. And therefore we thought there is value in actually optimizing the most common workflow and then you know people will realize that this is actually what they need and it ended up being like that. >> So let's talk about some of the earliest customers and their use cases. How did you get like your first 10 customers and how are they using it?

>> I think almost all of them were horizontal like design or image generation products when when a technology is so new. I think it's hard for products to be specialized. So a lot of them were general either consumer AI applications or you know web- based image generation apps. >> And I remember them some of the early ones being like I thought of them as a little bit like indie dev kind of hobbyist stuff. Sometimes when you see a new technology, it's a little bit like a toy. And some of your early customers were these kind of indie devs. Did you have concerns about that or did you think this is just how it's going to start? I think one of the biggest signs of how this is going to be around is the amount of money they spent, right?

Everyone was spending serious money on the platform. Tens of thousands of dollars a day all of a sudden. And you know, maybe this is temporary, but as long as it's going, it's definitely not a toy. people are spending serious money making serious products used by real people and to this day it's growing as we speak. >> Did you have a sense at that time that eventually like enterprises are going to want this? >> I think so because every every piece of technology was like it was just too magical to be ignored. The models just had to be a little more capable and we could see how like every month every 3 4 months the models were getting more more capable and people were moving into video. A lot of money was poured in into training video models, so we knew things were going to get more serious. The timing, none of us expected it to be as as quick. I remember Flux, it was sometime in 2024 just being a very big moment for the company and you guys had day zero support for Flux. Like the Flux moment, you know, in 2024 was kind of like the nano banana moment I think feel like we're having now for a video. But but how did you get day zero support for Flux? The Flux team worked at stability before. So we had a relationship with them already through their times at stability. Of all the demos we did attracted their attention and we were able to be become friends and colleague working partners during their time at stability. So they were pretty under the radar. No one was expecting the flux model, but we knew they were doing like doing training and we really respected the team and we we knew that they were going to do something awesome. We got in contact with them throughout during the summer and you know we planned a big release together and it worked great for us.

>> Okay. So video, you know, has been a huge part of your focus over the last 12 months. And I remember last summer you guys were telling me 2025 is going to be the year of AI generated video before. I mean it feels obvious now, but a year ago at least to me it was not that obvious. What were the signals you were seeing that were like giving you that conviction? So first of all like researchers always want to be in the cutting edge and all of a sudden like I would say couple months after flux we we realized we looked around all the researchers that were good at like the diffusion space they started working on the video problem all of a sudden everyone left pre-Sora or Sora had already happened in August Sora was February before that and I couldn't run it maybe some people did but it was just a demo after Sora uh serious money was put into training video models. Yeah, exactly. And researchers left the the image field and started working on on video, which is interesting. I don't know if there there's a term that describes the situation because not all problems were solved for image by any

means. There were so much to do there, but the price for video was much shinier and there was a lot of VC money being poured into it. all the researchers left doing image research and then focused on video. So you guys saw those signals happening and you started to prepare kind of the company right like what were some of the new challenges that you had to solve for video models were a lot bigger which which was to our advantage like the same reason why a bigger model like flux compared to stable diffusion XL was to our advantage same way a bigger model that requires more compute power requires multiple of the fastest GPUs to run in the cloud so the optimizations we make matter a lot more.

If something takes 1 second, if you can shave off 20% of it, maybe not enough people care about it. But if something takes a minute and you can shave off a similar percentage, all of a sudden that's a lot more meaningful. So now, you know, you guys find yourself kind of just at the center of this really interesting fastmoving market. It feels like every week there's like something new that happens. And so, you know, my question is how do you organize yourselves internally to be responsive to that? You guys are 40 people now.

>> 45 maybe. >> Like how do you operationalize a 45 person company to be this responsive when the market is is changing every week? >> Among many things, positioning ourselves as a generative media company helped us hire the people who are the most interested in this, right? We we have a applied ML team. It's around 15 people right now and you know all they do every day is either deploy these models, optimize them, play with them and they're obsessed with it. It's it's as if this is their hobby and if they didn't work at file they would be doing this anyways. So they are extremely on top of the market and now with FA's position in the market we also get some early information from the research labs. everyone like tries to talk to us and release their models on file on day zero. So we get some additional access to information as well. These two things combined, you know, we have some advantages over others in in the market, but unexpected things happen as well.

And you know, we have to react to it really fast and we were able to build that muscle very well in the team like something drops immediately. We drop everything and try to release that model if we don't know about it beforehand. >> How do you know if it's like the you know a new model is working? Exactly. Like the first phase is okay, how can we evaluate this as fast as possible? Is the demo video that this model is released with is is that accurate? Can we actually compare this to like other similar models? Is it actually worth it?

But if you if you get a sense like this thing is good, then we we we drop whatever and try like usually we open on Slack, we open a huddle, eight, nine people get in there, someone shares their screen and we try to get it out as fast as possible. Some of it we actually did it publicly too. That was that was interesting. Oh, what what did you do publicly? >> When when a model first drops I think BAN did this couple times or I'm doing a screen share. I'm speed running deploying this model on file and like people watched it. You definitely do this internally like Btoan would share his screen and others will watch to speedrun to deploy the model to be ready.

>> Are there other things that sort of go along with that? Do you have huge demand spikes when a new model drops? Yes, there is. And we we try to allocating GPUs to a model and doing that as elastically as possible. It's a problem that we have to think about all the time and we spend a lot of time managing our GPU capacity, how to do it the most efficient way. Sometimes we have to get GPUs really quickly. Sometimes we are able to plan it.

But yeah, this is an ongoing ongoing problem that we have to solve. You know, one of my favorite things about working with you guys is that like every time we meet, I feel like I get a crash course on kind of the current the current state of the market and just a lot of opinions and insights on on where the market is headed. So, for example, I remember talking to you guys and you're like, >> "Yeah, Netflix is starting to use Genai.

We're we're not far away from 2 and a half hour feature films being completely generated." >> Give us a picture like what does the what does the world look like in 2027 to you? I would say like six months ago, I wasn't sure if studios were gonna actually use this or is it going to be so fast that all of a sudden you'll have independent movies on YouTube and like studios completely missed this trend, right? Like sometimes technology happens in in light speed and you're stuck in innovators dilemma and other you know startups upstarts whatever companies leaprog you in in a way that like you don't get to even compete with them. I thought maybe something like that was going to happen to studios because they're already having trouble with some of the labor situation they are having already like the the film and movie industry is having trouble in terms of the revenues they are making. There's like very successful couple movies a year but the long tales of movies they've been having trouble. So, I wasn't sure if studios were going to be able to actually use this technology and it looked like they were completely avoiding it for a whole year, but something changed this summer and we are getting a ton of interest from basically all the studios in in in LA or elsewhere. Everyone is really interested to at least do something about it because now they they understand this is good enough and they can actually save money by by utilizing this technology. The creative people inside the studios I think they've spent enough time with this tool that okay now they see this as something that enhances their creativity rather than just like replacing them. So they bought into it as well. So this this summer I I would say that was the biggest difference.

>> Are there other uh industries, genres, games like other things you think will be really you know again two years from now? >> Yeah, gaming industry is is a little more sensitive. People really care about things being handdrawn or the the creativity aspect is really important. But yeah, I think it's very suitable for gaming industry as well. In thinking about the market, one of the things that we've talked about before is that image generation, video generation in many ways is a is a net new market. It's a it's a new opportunity. It's a green field opportunity. Um why why is that an interesting idea for you?

>> Yeah. So this is one of the reasons why we we decided to actually brand the company around generative media. We believe this is net new. Anything we are creating is not taking market share from like a giant company. We are not a database company taking database revenue from Oracle or we are not like in in the traditional LLM market. The the biggest use case seem to be search any LLM inference is taking market share from Google search and you know this is really important for Google. They would give it away for free if if they had to.

So they're protecting it really really close to themselves and that's why I don't think it's it's a startup's game to play in that race. But for video it's it's completely net new. So we believe it's very suitable for a startup. And one of the other things is it it was a small market in the beginning but very fast growing al that also makes it really really suitable for a startup as well. And even if like Google, OpenAI, even if they're active, they have their own models, you know, but because there isn't a clear target in front of them, they are not as as nimble as a startup.

So it's it's really hard for them to go towards that target because that target doesn't exist. They have to reinvent it every single month as as we do. So that puts us and the market in a very unique position for a startup like us to stay very nimble, change direction very fast. you know, get close to maybe advertising, maybe maybe studios, maybe more on the design side, maybe e-commerce and retail and build the whole team around around this rather than just having a single target and going towards it. I mean in some ways I think these market conditions ended up working out so well for you in that it was an overlooked market at the beginning or people thought it was too small which is which is great as a startup because it means you know you enter it and there isn't a lot of other competition and then even as it's grown there's no clear like sort of giant that you are threatening and who's going to come after you did you know did you predict all did you know this was >> maybe maybe we got a little bit lucky and in the whole AI market it looks like it's very hard to differentiate with models if you have a good model, you maybe have a 3-month lead, maybe 4 months. Uh people catch up for two reasons. Number one, if you had a unique research inside that leaks and other people do it. Number two, once people see something's possible, it's way easier for them to actually go try to do it because now they know it's possible.

Someone else did it and other people go do the same thing or try to reach to the same same idea. And number three, these models have APIs. people distill it. If you have a strong model, you can create similar models for way cheaper. So, it looks like it's going to be really hard for someone to differentiate with the quality of the model. And we believe this is going to lead to even even more fragmentation. So, a company like FA where you get to access many different models at the same time is is going to be positioned very well in the market going forward, the better better it is for us. Yeah. I think a lot of people don't fully understand what you guys do at fen why and why it's so hard. Part of that is you've done so much work on GPU infrastructure on model cold starts. You rarely talk about this stuff publicly I think correctly because you put the focus on the models you put the focus on what creative things people are doing with them ju just for a second because I think the amount of systems work and all the creative engineering that you guys have put into the platform is is wild.

So you don't have to give away your secret sauce, but what is what are some of the biggest optimization wins that you've had that that people don't know about? >> Yeah. So one thing people don't appreciate, it's a different problem to host a single model and optimize it and just serve that versus hosting 600 different models at the same time. All various different architectures, various different problems and all of them have different uh traffic patterns. It's a much much harder problem than hosting a single model. Any research lab either hosts their like marquee model and maybe a couple other. So the number of models they have to host is like less than 10 and and for us it's like around 600 which is complicates things like a lot and we had to build very very serious systems to to support this and one of the things that we have to pay attention to is how we utilize these GPUs. We can't just say all right we are going to equally deploy these models to to all the GPUs and you know hope that the traffic patterns are equal like everything has to be elastically scaling up and down. Uh and to be able to do that whenever a new request comes in we should be able to autoscale up and down really fast and that that is only possible with couple things. First cold starts is important. Starting a new model as fast as possible when it's a completely new start >> and we're talking about like seconds or milliseconds.

>> Seconds uh when everything is fresh but then we have to be really smart about how we cache these models. That's part of the strategy. We are running in I think 28 different data centers. So when a request comes in, we have to make sure the request goes into a node that has this model cached locally or or at least close enough so it

can can load into memory as fast as possible. And we do cache these models in memory sometimes too even if it's not being used at the time it stays cached while serving another model. called starts the caching strategy and if the model is latency sensitive for example audio models it matters a lot if the the model is in a data center that's close to you because you you care about milliseconds then routing the request to the closest GPU cluster is important as well when a node serves a request how long it's going to stay alive and when it's going to like shut down that's an important parameter as well um so we have to optimize all these things to serve these models as efficiently as possible. We need to be so good that like if you were to just host this one model on your own, we need to like host 600 of them but still be better if as if you you are doing one.

If you could wave a magic wand and solve one technical problem kind of in generative media right now, what would that be and why? One really important problem with model optimization, you you want to be able to run the models in many GPUs because let's say given that things scale linearly, instead of it taking 1 minute on one GPU, you want to run it on two GPUs and let it take 30 seconds. But in reality, every time you add another GPU, it's the gains are not linear. It's a little bit worse. You got to be able to do it as linear as possible. So if the optimizations work linearly I think it would solve a lot of the problems we are facing. We should we could be able to connect many GPUs and do like really really fast inference.

This is an active research and active engineering problem. Um you know we are getting close to make it linear but the more GPUs you add to the end it gets worse. So in the in the beginning it's close to linear and then it gets worse and worse. Another thing I've noticed is that you're very obsessed with developer experience and developer support and like I remember, you know, one of the founders that I work with was having an issue recently with Foul and I connected you guys on email and your response to him was like instant like it was faster than than the time you replied to my emails. Um, is there something that you guys do to make that customer obsession, the developer obsession scale?

>> From the beginning, Burka and I, we we we really cared about building for developers. We believe that's the best way to multiply our input to the world. Like we build for smart people who then build very smart products and then the the impact is multiplied. If you look at some of the newest software uh products like public public companies, they are all one way or the other like developer platforms or infrastructure platforms which is like a subset of developer platforms. So we we wanted to build a company for developers. We knew that was the biggest force multiplier we can get on our work. Once you decide to do that, you got to do your marketing towards developers. You got to make sure everyone gets the right amount of support. You got to care about things that developers care about. And you know, you can only do this if you obsess over the developer experience. And I think we are doing an amazing job at File at that. We have, I don't know, 500 different Slack channels with all the engineers from companies we work with.

and the the response rate of those Slack channels. We measure that daily and like we obsess over that. >> The market is changing rapidly. We've talked about this. You guys are are growing. Um and so naturally, you know, there's competition and I think you're you both have been incredible chess players and sort of navigate these changes. Well, how do you think about kind of competition in the space and and future chapters of FAL and and how you navigate that competition? One thing that helped us a lot was okay we had a technical advantage a

technical mode or how can we turned into a business advantage and like a marketing advantage calling ourselves a generative media platform and owning that term was was really helpful as biggest of the biggest enterprises were adopting this technology because when you are screaming what you do well people believe you right no one else is actually calling themselves a generative media inference platform. it to this day no one is defining their company as that and I think that positioning was a big advantage and now we can go out there talk about generative media and talk about the industry itself but indirectly we are just talking about ourselves being in that position is really really valuable and I think that differentiates us from our competition we are almost defining the industry while we are basically just talking about our own company and so that is a unique position and you know VC's always say this being the market leader you know there's a premium to that and yeah we are seeing that premium because whenever a big company wants to get into this we want to be the first phone call that they're making so speaking of of kind of big companies enterprises you know you've got Adobe as a customer canva what have been kind of the biggest lessons in in building for individual developers who you still serve serve and enterprises at the same time.

>> Our head of operations jokes about this. We we became a legal tech company with all these different AI models. The enterprise companies have a lot more concerns about all right how are these models trained? What happens with the inputs and outputs things like that versus like our initial set of customers? They just want to go to production as fast as possible. Um, so we had to build that muscle and being able to, you know, accommodate the requests coming from the enterprises about like the legal side of things and making sure they're comfortable with using this in production.

>> And so, do you feel like you're able to scale and serve both of these kind of different parts of the market equally well? >> 100%. Yeah, I'm not scared of any security review anything like we we've been through the hardest ones, so we could we could do anything else. >> And you guys are just scaling so quickly. It's it's one of the fastest early revenue growth rates I I think in in startup history. As you said, 2 million last summer and now it's >> now it's over 100.

>> And I know that model launches also make things a little lumpy. It seems like a hard to forecast business. Is there anything you've had to learn about managing this explosive growth and forecasting for what you need? >> Not forecasting, but you you mentioned model launches. It's it's just an incredible situation because every single model launch that happens in the platform is an opportunity for us to to do more marketing hopefully with with the research lab who was launching the model and then like go out there talk about this model, talk to more customers, an opportunity for us to talk to a customer we were trying to sign or get get them to use the platform. And now it's another opportunity for a touch point for us. Look, there's a new model and this is happening. I would say it used to be once a week, now like three, four times a week. Like so many model releases are happening. It's out of control. And all of them is an opportunity for us to get another big customer or get some more eyeballs on on X or on Reddit. uh it's it's just an incredible tool for us to be on top of all these model releases and be the first platform uh to support them.

>> Do you think about revenue quality and revenue durability like the revenue from one customer is different than the revenue from another customer? >> In the beginning mostly investors criticized us that this this

revenue may not be durable. You know the quality is not high enough. So we took enterprise sales very very seriously. We built a sales team early on, maybe earlier than some of our competitors, and we we tried to get as much of this revenue in form of yearly commitments rather than pay as you go. To this day, I think we are doing an incredible job at that. And that protects the revenue in a bit but also proves that you know these are serious companies doing serious business and willing to you know commit for 12 months or millions of dollars in in some cases. So this this was our reaction to that criticism. What are the top handful of metrics maybe the top two or three metrics that you watch internally the most obsessively?

>> Revenue is number one. anything else we've tried to do kind of became useless. Like number of requests was something that we cared about. All of a sudden when video models were part of the one video model request is a lot more valuable than image models. So that that doesn't matter anymore. Like the number of team accounts if you sign up as as a team and you have like many members in your team that means you are probably going to start spending more.

So that's one thing we pay attention closely. Other than that, like revenue has been has been the metric to to follow. I know as a PM you don't like that. >> I think it's incredible that as founders and especially as technical as you guys are that that revenue has always been your northstar. I mean as an investor you like to hear that >> PMs always want to find other metrics that are not correlated to revenue to to track. Maybe it's like revenue in the future but like not revenue right now.

It's been hard for us to find what that is, but maybe people do that when there isn't enough revenue. So, they're looking for other things to to track, but in our case, there is enough revenue and that's what we've been tracking. Okay, so let's talk about the team. You guys have built an insanely talented technical team. Have you been able to tap into talent that other people overlook or like how did you find some of your key hires? I think one of the advantages of working at a big Silicon Valley company for five plus years is you get to meet a lot of other very talented engineers. You know, you could do that at at at a startup too, but you don't meet enough people there. And in a big company like Amazon or Coinbase, we bought like met a ton of talented engineers and we were able to bring the most talented ones into into fall. It was their first startup job maybe, right? they they trusted us to work at like in this environment and that's how we were able to bring them in into fall.

So that worked great for us and then yes we are both Turkish and our early engineering team had had some Turkish members as well. We were able to attract incredible talent from back home. That's cuz people trust us more and it's just a very good company for them to work at. That's that's been really useful for us. How global is the engineering team? >> I would say more in person in the six months. People are willing to move here.

You know, we've been trying to get both our American engineers who live in other cities to move to San Francisco, but also some of the overseas engineers to move to San Francisco as well. The center of gravity has definitely moved to San Francisco, but we we still have like 15 people maybe in remote. What are you looking for either in the ML engineers that you hire the systems engineers like is there is there a specific thing that you evaluate?

>> It's either obsession with optimization. So if they worked at database company before or if they did any like low-level systems engineering that is a big plus even if they haven't worked with a GPU before we believe they can learn very fast or obsession with the space. In the beginning, this was harder to do because the space was so new. There wasn't enough time for them to get obsessed with. But now you can five minute conversation immediately understand if they they love video models, image models, they know what's going on. And if you can feel their obsession, it's extremely useful for us. And those are the two things we are we are looking for when we are hiring ML engineers. Is there anything kind of unique or different that you guys do through interviews or the way that you hire folks? This is an evolving process. In the beginning, we were able to get to know people, spend more time with them somehow with with the amount of time we spend in the in the ecosystem. So, we were able to know someone for a longer period of time with the work they do out in the open. uh and you know we had more more things to look at before deciding if they should join or not and now we have a recruiter we have a bigger pipeline so some of the people I'm meeting them for the first time at the interview so that process has been evolving like what are the things we can do differently than Meta or Google that we can identify the unique things that these these people can be valuable to to a company like FA rather than any any big software organization.

>> How do you recruit an amazing ML engineer who's also talking to the various large labs that's got to be hard, right? Is there is there a thing that um is it just obsession with the with the category? So this question used to be like how do you convince someone who has also an offer at Google or Facebook? I think we are past that. Like you said, we need to look at some some overlooked career paths. Find people non-traditional backgrounds, maybe younger earlier in their career that might help. Open AAI and anthropic there. You know, young companies too, they can't be interviewing everyone. So, you know, there's enough talent in the world if you know where to look.

>> And speaking of of just talent and how concentrated the talent is at FA, like you guys have kept the team lean. I mean I think you were at 25 people when you crossed 50 million and now you're at 45 crossing 100 million and and how many like folks do you have on sales now in go to market? >> We have around six maybe maybe 10 if you if you include CSM and like all that. >> Okay, but still like how many companies you know had 100 million revenue with like you know 10 people on the go to market side.

>> I think AI kind of changed a lot of things in how software companies did sales because before the market was more stagnant. you had to actually go like convince people to use your product and they were choosing between like five different options and the sales process was really really long. But with AI you have so much demand coming from the market you have to qualify like your problems are very different. Your problem is you have to qualify who to spend time with. You have to qualify who's going to have the most spend among these companies for you to actually like go talk to them. So the the go to market motion is very fast-paced like in a way transactional. So it's it's very different than the traditional SAS sales cycles and yeah we have to hire for that type of AE profile which has been an interesting challenge for us as well.

>> You know as technical founders has been your biggest lessons kind of in this area of of teaching yourself sales. I know you did a bunch most of the early sales. I think the biggest lesson was starting early to get people

into commitments because immediately that's a measure to understand how serious the the person in front of you is as well if if they are willing to uh invest in your relationship and you know do they trust you? Do they think you can do a good job and make them successful?

Getting people to commit as early as possible and giving them that option. I was a huge skeptic. I was like, "No one's going to commit to this." Like, >> you you didn't do sales, right? >> No, I didn't do sales. No, Burkai didn't either. >> So, how do you how do you sell the product, especially to like a large enterprise? What is how do you sell the product? >> First, we want them to try it out. We made sure that's as easy as possible.

You log in, put in put in a credit card, and you can just use the product. A lot of our contracted revenue even comes from inbound like people already tried the product and they have some sort of spend and now we are reaching out to them. All right, you are on track to spend $10,000 this month if we give you I don't know 10% discount, 7% discount, would you be willing to do a yearly or two yearly commitment? your annual enterprise customer started as there's a couple engineers doing pay as you go.

Basically, we have some signals on on Salesforce. If someone's, let's say, I think it's $300 a day or something like that. If they spend more than that, it creates an opportunity on Salesforce. One of the AES take that opportunity, you know, there's an email associated with it. They reach out, they try to get on the phone and convert that to a yearly contract. There are very few high-profile like me on Burkai would would reach out to someone who we know they have a lot of generative media spend. There's like couple of those but majority happened inbound first and then we we converted to a yearly contract.

>> I mean outside all of the work that goes into product are there any examples of how you guys are using AI internally? Yeah, our product team uses cursor or equivalent uh tools a lot like I see the monthly bill and it keeps increasing. Um yeah, I think it's better suited for product engineering type work um as opposed to some of the low-level optimizations we are doing on the on the ML side. Yeah, I would say our our product engineering team and like our API team is loves cursor and similar products. On the sales side, I believe we are using clay to, you know, enrich some of the leads we have and figuring out who to reach out, emails, stuff like that. I think that's pretty much it. So shifting gears a little bit but staying on team. You guys have, you know, some awesome open roles now for marketing.

But I think that you guys have done a pretty incredible job marketing the product in the company without having sort of formal marketing people. You've got this like amazing swag, you know, the GPU rich GPU poor hats. Uh the website is awesome. You've been you sponsored the Padell Courts in San Francisco, you know, nano banana hackathons, generated video awards, and your first developer conference coming up. So, how have you thought about brand building and kind of having taste as as founders and where does that inspiration kind of come from?

>> Yeah, shout out to Adam Ho who worked on our initial branding which to this day we use it and we we had couple other iterations to even expand on it. He did a great job and we just trusted him. We didn't have any designer in house or or a marketer. So we just asked him to go crazy with it and he came up with this amazing

brand that we have today and we love it so much that we want to like show it everywhere. So it's it's been great. In terms of developer marketing, I think that's a unique taste.

Traditional marketing doesn't work for developers. People think it's cringe. People like don't don't want don't want everything to be very obvious on their face. They want this this subtle tasteful marketing that is you know only towards developers. I believe indirectly we are doing it very well. We we all have really active X profiles. We hired couple people just because they had an active X and they ended up being like really active members of the community as well. So that helps a lot. Yeah, we try to do hackathons. We try to go to conferences but at some scale this needs to be more organized. Therefore we are looking for couple not a lot uh of positions to help us you know put put more organization into this. Everything has been like founder le more spontaneous. We just decided to do something one day and one of us goes and executes it but we need to do this little more organized and we want to make sure we are not missing important conferences. We don't have a hackathon and like another event on the same day.

We want to make sure we are like spreading this enough. Did a lot of it just come from the way that you and Burkhai like to be marketed to yourselves? Like a lot of it is fairly intuitive. >> I think so. Like in a way, we didn't hire a marketer early on. This is a really important position. You're representing the the file brand. You're representing the company. You wanted this person to be as good as it can be to represent us basically. So we wanted to do this on our own in the beginning to to know how to do it where where we fall short like maybe fail a couple of times and exactly understand where we need help so so we can hire and and I believe now is the time.

>> How did you come up with the GPU rich GPU poor hats? >> That's like one of the earliest things we we did at the time. Dylan from semi analysis blog wrote an article about how everyone is GPU poor except Google. I think that was the main point of the article and we had a conference coming up maybe maybe in like two weeks 3 weeks like we didn't have that much time and this became a meme on Twitter. We just you know jumped onto the meme and created these these two hats. one of them very basic GPU poor just white on black plain font and then like this GPU rich looks like a country club green on on white and we just showed up to to this conference with with these two types of hats we thought like oh people would like the GPU rich all of a sudden like and we made equal number of them and we ran out of the GPU poor hats like way before the GPU rich ones everyone thought like GPU poor was hilarious. Is there anything about your culture or how you guys operate that would sort of seem I don't know kind of weird to outsiders but is very essential to how foul runs?

Yeah, we don't have engineering managers. We have around like 32 34 engineers. We do have leads obviously we have leaders in the in the team but you know we don't have this engineering manager role. Everyone is always like contributing writing code. There's no one whose job is just to like manage people. >> That reminds me of early Google actually. I think each of there was, you know, some of engineering VPs there and each of them had like 40 or 50 direct reports, something like that.

>> At Amazon, we had engineering managers who their job was to maybe set goals, but also like have one-on-ones with everyone. And that was it. And we don't have that. >> When's that going to break? >> We'll see. I

don't know. One thing I like to do is like we we also have very few one-on- ones. Instead of one-on- ones, we try to do like smaller groups of discussions like one on three or whatever, one on four. And we try to bring like people from within the team, but maybe someone who joined recently, someone who's been there for a while, someone who's remote, someone who's in office, like different types of people together so that, you know, it's a discussion. If they want to complain about something, they can complain. But usually those meetings are a lot more constructive than the old one-on-one style where you're almost forcing someone to complain about. Okay, this is your time to start complaining about it.

And like when you give that that time, people complain, right? That's that's what they do. And so we believe this this group style is a lot more constructive. >> Well, some questions to wrap up. What is the hardest part about building foul that you didn't anticipate when you started? I would say this is the hardest part of I would say any growing company not specific to FAL is hiring executives, trusting people who are very experienced to come in and take over pieces of the company that have been there for a while and trusting them to actually like do a good job. when you ask the question of like where did you make mistakes this is usually what people say and like we are trying to be extra careful with it not to make mistakes and we are we are taking as much time as possible yeah that's that's been the hardest we built a sales team before we hire the head of sales I think this is number one question like series A or series B companies ask themselves what comes first we we decided to hire I think like six AES first everyone reported to either me or Burkai and then we we hired a head of sales. I think that was the right thing to do. But maybe it was the wrong thing to do to do when we look six 6 months from now and like the whole team is fighting with each other. I don't know. Hopefully that doesn't happen. But I'm now much more confident about evaluating this new head of sales we hired and the AES like how what success looks like. I saw it. I experienced it. It was painful to get there, but now I'm much more confident in my ability to actually evaluate the situation than if I had hired the head of sales before and they hired the the whole sales team.

>> What's been the biggest surprise about being a founder versus being a technical leader? >> It's just the wide range of things you have to do any given day. You might be negotiating marketing material on a PAL court also like getting on recruiting calls with ML engineers but also giving investor updates and all of these might be happening backtoback meetings and you are responsible for all these different things. What are the things that you're thinking about personally right now and and maybe Burkai too if you know about leveling up you know as an incredible founder over the next year or two. One difference from having a fivep person team and 45 person team is actually the amount of good information you get from everyone in the team and the amount of market intelligence, the amount of insights that like we get to discuss in our office or in our standups and now I get to go represent those ideas to other people. that that is like insane amount of leverage I think you don't necessarily get if you are an engineering manager in in a small part of a big company or if you are the founder of a smaller company. I think that's been the biggest difference and it's been huge leverage for for my personal development because I get access to all this very smart engineers in my team and I get to like represent their ideas.

>> Girkham, thanks for being here. It was great. >> Of course. [Music]