# Episode Book

EPISODE #5533

# Why AI Characters Drift Between Shots

| | |
|---|---|
| PUBLISHED | 24 September 2026 at 17:54 UTC |
| DURATION | 21:32 |
| LISTEN | myweirdprompts.com/episode/ai-character-consistency-drift/ |
| HOSTS | Corn, Herman, Hilbert |
| SCRIPT MODEL | deepseek-flash |
| VOICES | chatterbox-regular (chatterbox-local) |
| PIPELINE | V5.2 |
| BOOK GENERATED | 19 October 2026 at 17:37 UTC |

## SYNOPSIS

You generate a character, they look perfect in shot one, and by shot four they're someone else's slightly off cousin. Why? This episode digs into the structural reasons image-to-video models drift: stateless sampling, category-level text prompts, and the fundamental tension between identity consistency and scene diversity. We cover when training a LoRA actually makes sense (and when it's a day of setup for nothing), what the ideal single reference image looks like, why matching your reference background to your scene backfires, and the counterintuitive truth about multi-reference conditioning — more angles can make drift worse. If you've ever wondered why the model "fails" at character consistency, this reframes the whole problem.

# TRANSCRIPT



### Corn

Here's what Daniel wrote in this week. He wants to talk about the technical challenges of single-shot character generation across image-to-video and, at the frontier, video-to-video. He points out that the advanced approach for character consistency is training a LoRA, but that it's time-consuming, especially for one-off generations. He's seen examples where the reference image was adhered to perfectly, and examples where the character came out unrecognisable. Sometimes it works, sometimes it doesn't. And when it works, the model is doing something hard, taking one single example and extrapolating that character into a new setting.



### Herman

Right, and then he gets to the practical heart of it.



### Corn

He does. He asks: if we want the most ideal single image for image-to-video, with the objective that the reference is as useful as possible for capturing the character, what characteristics help the model and give it the best chance of a compelling result? Specifically, if we're taking a photograph of a person and putting them in a new scene, do we want the person in as comparable a background as we intend to put them in? What are the attributes you look for in the ideal reference image? And finally, since models now accept several reference images, if we're looking at three-image conditioning, how would you choose those three?



### Herman

That last one has a counterintuitive answer. So let's start with why this is so hard in the first place.



### Corn

Start with the blank slate.



### Herman

The blank slate is the whole thing. These models are stateless. They do not store your character anywhere. They sample from a learned distribution of pixels, frames, and motion, and every new clip is a fresh draw. There's no memory of the person you generated yesterday.



### Corn

So when I give it a photo of a face and a prompt, what does it actually have?



### Herman

It has a conditioning signal and a text prompt, and both of those are weaker than people assume. The text prompt is the more obvious problem. If you write "a woman in her thirties with dark curly hair and a green coat," you have described a class of people. Not one specific individual. A category.



### Corn

And the model complies perfectly with that category.



### Herman

That's exactly why the result drifts. The model is doing what you asked. You asked for a member of a category, and it gave you a plausible member of that category. It just wasn't the same member as last time.



### Corn

So the reference image is supposed to pin down which member of the category.



### Herman

And that's where the inference problem starts. Image-to-video models do not simply move pixels. They infer motion, hidden surfaces, expressions, clothing folds, hands, background depth, camera changes. All of that from a still frame. That inference is powerful, and it is also exactly where character drift begins.



**Corn**

Because every one of those inferences is a guess.



**Herman**

Every one is a guess, and the guesses compound. There are three failure patterns that show up over and over. Cross-shot drift, where the character is correct in shot one and subtly wrong by shot four. Intra-shot morphing, where the face deforms mid-clip, usually around occlusion or fast motion or a hard angle. And wardrobe or prop substitution, where the green coat quietly becomes a different green coat.



**Corn**

Shot one is the person. Shot four is the person's cousin.



**Herman**

The person's slightly off cousin, yes. And there's a real mathematical reason underneath it. Errors in diffusion sampling are recursive. Errors in earlier steps produce iterates that drift away from the training distribution, and each step inherits the last one's mistakes. It's not that step forty is bad. It's that step forty inherited step thirty-nine's problem.



**Corn**

So the drift isn't a bug that gets patched. It's structural.



**Herman**

It's structural, and there's a second structural thing on top. Identity consistency and prompt diversity are fundamentally in tension. The further you push a scene away from the reference, the more identity gives.



**Corn**

Say that again, because I think that's the sentence the whole episode hangs on.



**Herman**

The further a scene sits from your reference, the more the face gives. You want the character in a new setting, new lighting, new pose, new framing. Every one of those moves is a step away from the reference, and identity degrades as you take them.



**Corn**

So the model isn't failing when it drifts. It's paying a price you asked it to pay.



**Herman**

Which reframes "the model failed" as an inherent trade-off rather than a skill issue. And this is why people reach for LoRA.



**Corn**

The advanced approach. Explain what it actually is.



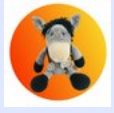
**Herman**

LoRA is low-rank adaptation. You train a small adapter file, usually somewhere between forty and a hundred and fifty megabytes, from a dataset of images. That adapter teaches the base model a specific character. Then you can generate that character in any context without retraining the whole model.



**Corn**

Forty to a hundred and fifty megabytes against what, a six-gigabyte base model?



**Herman**

Against a six-gigabyte base model. So it's tiny, and the identity gains are real. Reported face consistency with a LoRA runs ninety-five to ninety-eight percent. Without one, it's around sixty.



**Corn**

That's a chasm.



**Herman**

It's a chasm. And here's why it's still the wrong answer for most of what Daniel is describing. Training takes thirty to sixty minutes on an RTX 4070, or one to four hours depending on your setup. On cloud GPUs it costs somewhere between fifty cents and a dollar fifty. That sounds fine.



**Corn**

There's a but coming.



**Herman**

The but is the first time. Plan a full day. Half of it goes into setup. Python virtual environment, CUDA-matched PyTorch, the training toolkit and all its dependencies. Most of those "consistent character in thirty minutes" headlines quietly skip the setup part. It's real. Plan for a day.



**Corn**

So for a one-off generation, you spend a day of setup to get one image.



**Herman**

And then the dataset question. Fifteen to thirty images is common. A hundred plus if you want full-body. Two hundred plus for a character you'll use across multiple purposes. For a single campaign, that's absurd. The rule that actually holds is: start with reference images. Train a LoRA only when references stop giving you enough control, or when you need the same character across so many images or videos that supplying and correcting references every time becomes impractical.



**Corn**

So the decision is purely about reuse count.



**Herman**

Purely about reuse count. One-off, use references. Recurring character across dozens of shots, train the adapter. There's no third option worth considering.



**Corn**

Which brings us to the part Daniel actually asked about. The ideal single reference image.



**Herman**

There's a consensus spec, and it's boring. Single subject. Front or three-quarter facing. Evenly lit. Neutral expression. Plain background.



**Corn**

That's it?



**Herman**

That's the spec. Now let me tell you why each one of those is doing work. Start with the face being large and clear. The single most common cause of drift is a face that's too small, blurry, shadowed, or partially covered. If the model can't resolve the features, it invents them.



**Corn**

So a sharp portrait beats a nice full-body shot.



**Herman**

Sharply. A blurry full-body shot contributes almost nothing to facial identity. A sharp, evenly lit portrait contributes enormously. You're not giving the model a picture of a person. You're giving it a specification of a face.



**Corn**

Then the lighting point, which I think is the one people get wrong most often.



**Herman**

Photographs taken in varied or dramatic lighting can embed the lighting conditions as part of the facial features. The model learns the shadows as though they were part of the face, and then it can't relight the face naturally.



**Corn**

So if your reference is lit from the side with a hard shadow down one cheek, the model thinks that shadow is a feature.



**Herman**

It bakes the shadow in. Then you ask for the character in soft daylight and you get a face with a permanent hard shadow on one side. Runway's own guidance asks for high-quality subject images with even lighting and neutral expressions, so the reference behaves as a blank canvas that your prompt then lights.



**Corn**

The reference is a canvas, not a photograph.



**Herman**

That's the mental model. And it explains the background answer too. ByteDance's Seedance guide recommends a head-only close-up, face retained, neutral expression, with shoulders, neck and background minimised as interfering elements. Interfering elements. That's the phrase.



**Corn**

What about resolution? Is there a floor?



**Herman**

There is. Runway recommends references between six forty by six forty and four K. Above ten twenty-four pixels on the short edge is a safe working floor. Below that, you're asking the model to upscale guesswork.



**Corn**

So the ideal reference is a passport photo.



**Herman**

The ideal reference is essentially a very good passport photo. Boring, evenly lit, neutral, plain background, face filling the frame. Everything the model needs and nothing it has to interpret.



**Corn**

Which sets up the background question directly. Should the reference background match the scene you're putting them in?



**Herman**

No. And every source I found says the opposite of what people assume. Use a plain, neutral background and rely on subject masking to prevent bleed.



**Corn**

Why does it bleed?



### Herman

Without subject masking, the background of your references bleeds into the scene. A studio wall in your reference becomes grey haze behind your hero in the output. And the reference palette contaminates the whole grade. If your reference has a warm beige backdrop, your output gets a warm beige cast even if the scene is supposed to be cold blue night.



### Corn

So matching the background doesn't help the model place the person. It just poisons the colour.



### Herman

It poisons the colour. And the reason matching is unnecessary is that identity travels independently of lighting and location. Runway says its image references generate consistent characters across different lighting conditions, locations and treatments from a single reference image.



### Corn

So the model already separates the person from the place.



### Herman

It already separates them. You do not need to help it by matching. You want a neutral canvas, and then you let the prompt and the masking do the placement. The instinct to match is the instinct that causes the problem.



### Corn

That brings us to the multi-reference question, which is where the guidance contradicts itself.



**Herman**

It does. And the contradiction is the most interesting thing in the whole topic. The most repeated tip in creator guides is: upload multiple angles. Front, three-quarter, profile, full body. More angles, better identity.



**Corn**

And that's wrong?



**Herman**

For faces, ByteDance's own Seedance 2.0 prompt guide advises against multi-view character sheets for people. Because the model can read different angles as different subjects, and that worsens ID drift.



**Corn**

The model sees a profile and a front view and thinks it's looking at two people.



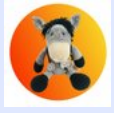
**Herman**

It thinks it's looking at two people, and then it has to average them, and the average is nobody. So the recommendation is: send the hero frame plus at most one or two that clearly agree with it. Resist filling every slot. Two clean frames that agree beat six that argue.



**Corn**

That's a good line.



**Herman**

It's the line. And the mechanism is that too many assets make it harder for the model to judge which features take priority. Every reference is competing for influence. If you give it six, you've given it six arguments and no chairman.



**Corn**

So what do you do with the slot budget?



**Herman**

Put the most precision-critical asset earliest in the prompt. If the face is what matters, the face goes first. And don't fill the slots just because they exist. The existence of a slot is not an instruction to use it.



**Corn**

Now here's where it gets unresolved. There's a competing recommendation.



**Herman**

There is. Domer AI recommends four to six varied angles for narrative video. A neutral front view, a three-quarter view, a profile, a full-body shot for silhouette and proportions, and two expression variations. Their argument is that variety in angle is worth far more than variety in style.



**Corn**

Which is the direct opposite of ByteDance.



**Herman**

Directly opposite. And both are credible. The difference may be faces versus full characters. ByteDance is talking about human faces specifically. Domer is talking about characters in narrative video, where silhouette and proportion matter as much as the face. But neither source resolves it cleanly, and I'm not going to pretend they do.



**Corn**

So the honest answer to Daniel's three-image question is: it depends, and here's the shape of the depends.



### Herman

The shape is this. If the face is the identity, two or three frames that agree. If the character is the identity, meaning the silhouette and the costume and the proportions carry as much as the face, then more angles can help. But you're trading one kind of consistency for another.



### Corn

And there's a weighting point underneath that.



### Herman

There is. A sharp, evenly lit portrait contributes far more to facial identity than a blurry full-body shot. A profile view contributes mostly to hairline and jaw structure, not to the face front-on. So the images aren't equal. You're not choosing three images. You're choosing three contributions.



### Corn

And the one people never think about.



### Herman

Colour temperature. Keep it consistent within the pack. Mixed lighting confuses the encoder about skin tone. If one reference is tungsten and one is daylight, the model gets two different skin tones and has to pick.



### Corn

So the three images should agree with each other on lighting as well as on the face.



### Herman

They should agree on everything except the angle you deliberately varied. Every disagreement is a decision the model has to make, and it will sometimes make it wrong.



**Corn**

Let's talk about the caps, because the numbers are all over the place.



**Herman**

They are. Veo 3.1 takes three images of one person. Seedance 2.0 takes nine, but real human faces are restricted. Seedance 2.5 takes up to four. Kling 3.0 takes three elements. Nano Banana Pro takes five character images. Seedream 5.0 Pro takes ten references. GPT Image 2 takes sixteen input images. FLUX.2 flex takes eight through the API.



**Corn**

Sixteen.



**Herman**

Sixteen. And then there's a stability cliff. ByteDance documents output stability dropping when reference people exceed four. So the model that accepts nine doesn't necessarily reward nine.



**Corn**

The cap is a capability, not a recommendation.



**Herman**

The cap is what the plumbing accepts. It is not what the model wants. Those are two different numbers, and the gap between them is where people get burned.



**Corn**

And then the policy side, which is tightening fast.



**Herman**

It is. Sora 2 rejects real human faces entirely, and its API was removed on the twenty-fourth of September. That's today. Seedance 2.0 requires consent and a face check. The "put a photographed person into a new scene" use case is increasingly gated by policy, not just capability.



**Corn**

So the capability is improving and the permission is narrowing at the same time.



**Herman**

At the same time, and in the same direction. The technical question Daniel is asking may be partly overtaken by the consent question. You can get very good at choosing reference images for a use case that's increasingly restricted.



**Corn**

Which is worth sitting with. But let's land the practical part first, because Daniel asked a direct question and we owe him a direct answer.



**Herman**

The direct answer. One reference image: single subject, front or three-quarter, evenly lit, neutral expression, plain background, face large and sharp, above ten twenty-four on the short edge. Do not match the background to the target scene. Do not use dramatic lighting.



**Corn**

Three reference images: the hero frame plus one or two that clearly agree. Put the face first. Keep the colour temperature consistent. And if the model accepts more than three, don't use more than three unless the silhouette is carrying the identity.



**Herman**

That's the answer. And the honest caveat is that there is no canonical study on this. I looked. The guidance is empirical and vendor-derived. Nobody has run a controlled experiment isolating reference-image attributes as the independent variable.



**Corn**

So the whole field is running on folklore that happens to work.



**Hilbert**

Folklore's a strong word. It's a trade.



**Corn**

Go on.



**Hilbert**

I did portraits for a mall studio chain. Blue gradient backdrop, a stool, and a photographer telling you to say cheese. That's the ideal reference image, and it's ideal because it's boring. Same backdrop every time. Same two lights. Same distance. You could put a thousand people through that setup and the only variable was the person.



**Corn**

The studio accidentally solved the problem.



**Hilbert**

The studio solved it by accident, because the studio's problem was throughput. You can't spend twenty minutes on a lighting setup when there's a queue. So you build one setup and you never touch it. Every portrait comes out looking the same except for the face. Which is exactly what these models want.



**Corn**

And the patterned shirts.



**Hilbert**

People wore patterned shirts. Busy prints, stripes, logos. I'd tell them to change, and they'd say it's their favourite shirt, and I'd say it'll look like noise in the picture, and they'd wear it anyway. Then they'd see the proof sheet and ask why the shirt looked like that.



**Corn**

What was the answer?



**Hilbert**

The answer was that the camera was fine and the shirt was the problem. Same as these models. The reference isn't bad because the model is bad. The reference is bad because nobody told the person how to be photographed.



**Herman**

That's the masking argument in a different costume. The interfering element isn't the background, it's whatever the subject brought with them.



**Hilbert**

I tried to train one of those adapters on my own headshots from that era. The ones from the studio.



**Corn**

How did it go?



**Hilbert**

It kept generating me with a mullet. I never had a mullet.



**Corn**

Are you sure?



**Hilbert**

I'm sure. Short on the sides, short on top. The model disagreed. Every generation, mullet. I don't know where it got it from. Maybe the lighting. Maybe the era. It decided that's what I looked like and it wouldn't be talked out of it.



**Herman**

The model knew you better than you knew you.



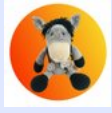
**Hilbert**

The model was wrong. Anyway, I sold the lights. There's an appointment I'm already late for.



**Corn**

The model was wrong about the mullet, and right about everything else. There's something in that.



**Herman**

There's something in the whole segment. The studio solved the problem by removing every variable except the face. That's the spec we've been describing all episode, arrived at by a mall photographer with a queue.



### Corn

The misconception I want to kill is the one people repeat most: that more reference angles are always better. ByteDance's own documentation says the opposite for human faces, because the model can read different angles as different subjects. Two clean frames that agree beat six that argue.



### Herman

The second one, which is the reason people blame themselves: if the model fails, it's user error. It isn't. OpenAI documents that its own image models may occasionally struggle to maintain visual consistency for recurring characters. This is an acknowledged limitation, not a skill issue.



### Corn

The unresolved thing I'm left with is the contradiction. ByteDance says don't use multi-view sheets for faces. Domer says four to six varied angles for narrative video. Both are credible and neither resolves it. And there's no controlled study on what makes an ideal reference image. The guidance is empirical and vendor-derived.



### Herman

Which is the open question. As these models improve, does the reference image matter less, or does the tension between identity and diversity stay fundamental? My guess is it stays. You're asking for two things that pull against each other. Better models narrow the gap, they don't remove it.



### Corn

The policy side may get there first. The capability is improving while the permission narrows. The use case Daniel's describing may be gated by consent and verification long before it's gated by quality.



### Herman

Which is worth watching. Thanks to our producer, Hilbert Flumingtop.



**Corn**

This has been My Weird Prompts. If you enjoyed this episode, please leave a review on your podcast platform of choice. It helps other listeners find the show.



**Herman**

We'll be back soon.