Episode Book

EPISODE #6303

Blind Testing AI Video Models for Architects

PUBLISHED	27 October 2026 at 18:22 UTC
DURATION	23:31
LISTEN	myweirdprompts.com/episode/evaluating-video-models-architecture-02204/
HOSTS	Corn, Herman, Dorothy, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.22
BOOK GENERATED	28 October 2026 at 09:53 UTC

SYNOPSIS

Evaluating AI video models requires more than skimming glossy vendor demos, especially when an architectural project cannot tolerate shifting walls or warping window mullions. Evaluating tools like GenMedia, media-gen-arena, and GenMediaBenchmark reveals the practical mechanics of running internal blind pairwise comparisons. Rigorous testing demands tracking settings, managing per-clip generation costs, inspecting frame-by-frame structural defects with tools like DeltaFrame, and borrowing standardized benchmarking protocols from Artificial Analysis without mistaking strangers' aesthetic tastes for a firm's concrete technical requirements.

TRANSCRIPT



Corn

The usual advice is to pick the video model with the most cinematic demo. An architecture firm needs the one that leaves the building standing. Daniel wants to know how a small practice can test that before paying for a whole batch of walkthroughs. Same renders, same prompts, perhaps ten or twenty models, then blind ratings from a handful of partners.



Herman

A gorgeous camera move can hide a wall changing shape. The clip feels expensive while quietly redesigning the project.



Corn

Unrequested design services.



Corn

His question is whether there's actual tooling to fan those inputs out, collect the clips, and let people judge without knowing which model made what. Not a giant benchmark. A controlled comparison where somebody can point at a window frame and say, that was straight until the camera moved.



Herman

There are purpose-built options now. The choice depends quite heavily on whether the firm generates through fal or Replicate. The more surprising part, to me, is how much of the evaluation procedure is already published and copyable.



Corn

Before you install anything in our flat, Herman, distinguish this from an LLM evaluation.



Herman

No installation required for that distinction. Although I appreciate the preventative maintenance.



Herman

The unit of judgment is a partner watching a particular shot. Did the parapet remain rigid? Did the stone stay stone? A numerical score can record that judgment, but it doesn't supply it. You're choosing a production tool for a specific practice, with its own buildings and tolerance for defects. We went through the labelling tools for text benchmarks a while back, Argilla, Label Studio and the rest. The video twist is that every item on the sheet costs real money to make.



Corn

I'd also distinguish choosing a model from choosing its best accident. If somebody generates six clips and brings the nicest one to the meeting, we've evaluated their patience.



Herman

That matters enormously with video cost. The media-gen-arena README uses sixteen prompts and two seeds across three models. That produces ninety-six generations.



Corn

Ninety-six opportunities to purchase a new window.



Herman

In video, those are ninety-six clips to pay for, download, verify, and review. Scale that same design to twenty models and you're buying six hundred forty clips. At an assumed dollar per clip, that's six hundred forty dollars before retries. That's arithmetic, not a provider quote. Actual pricing depends on the endpoint and settings.



Corn

I'd spend some of that budget deciding which renders are informative. A blank plaster corridor lets nearly anything look competent. Railings, repeated windows, a corner where near and distant surfaces separate... those give the camera something it can get wrong.



Herman

Would you include a deliberately difficult render even if it isn't typical of their work?



Corn

One, perhaps. Mostly representative projects. Otherwise you select the model that wins an obstacle course and then ask it to film a bungalow. I want the difficult example to expose a weakness, not dominate the purchase.



Herman

Then the generation record matters as much as the gallery. Manual uploads leave settings scattered across tabs. Prompt enhancement might be enabled on one endpoint, the duration different on another. A failed generation quietly disappears. Next month, nobody can explain why the winner changed.



Corn

The browser history becomes the laboratory notebook.



Herman

Blind presentation, recorded configurations, and an explicit judgment criterion solve different problems. None substitutes for the others. A beautifully blinded test of mismatched settings still compares the wrong things.



Corn

My guess is that you can buy the voting screen, but somebody still has to stitch the generation together.



Herman

GenMedia gets closer than that. It generates natively on fal, resolves the provider's input contracts, and explicitly maps each prompt and reference image into the run. It preserves the lineage behind each output. You can also import finished outputs from elsewhere, so the evaluation isn't restricted to clips it generated itself.



Corn

Importing matters. Otherwise choosing an evaluation product quietly chooses your generation provider.



Herman

Its review workflows support blind pairwise comparisons or ratings. Candidate order is shuffled deterministically, which means the presentation can be reproduced. Model identities and aggregate results stay hidden until a task completion gate is satisfied.



Corn

No watching the scoreboard before deciding what you saw.



Herman

It also reports uncertainty intervals, pair coverage, weighted rankings, and checks for position effects. Multiplicity-aware tests help when you're comparing a lot of candidates and could otherwise mistake a chance difference for a discovery.



Corn

That dashboard will tempt you, Herman. You'll want twenty models because the chart has room.



Herman

I'll want twenty models because they're interesting. Entirely different weakness.



Corn

Same invoice.



Herman

Their example has twenty-four prompts and eighteen reviewers, with ninety-two percent coverage. That's a useful sense of the review effort behind a polished result, rather than an implied minimum for every team.



Corn

Coverage meaning whether the comparisons actually got reviewed, not whether the model covered the terrace in ivy.



Herman

Yes. The example preference split is thirty-eight percent against sixty-two percent. You need the underlying coverage and uncertainty to interpret that. A large-looking gap on a decorative dashboard is no guarantee that the firm's own tiny panel will reproduce it.



Corn

For this job, I'd be more interested in locating the defect than announcing the winner. Partners need something better than dragging a video scrubber and saying, somewhere around here.



Herman

GenMedia has DeltaFrame for that. Synchronized frame stepping, split views, difference views, and jumping to the moment motion or detail breaks. That's directly useful when a vertical edge starts bending halfway through a dolly.



Corn

Finally, a product feature specifically for finding the bad mullion.



Herman

A difference view still requires interpretation. Moving the camera changes pixels even in a perfectly correct scene. It helps you inspect; it doesn't automatically establish geometric error.



Corn

Replicate users don't get the same ready-made path?



Herman

The closest fit is media-gen-arena, which is open source and Replicate-native. It handles generation, verifies stored media, builds blind comparison pairs, serves the voting interface, runs data-quality reports, and fits ratings. You're operating a package and a web interface rather than buying the whole hosted service.



Corn

That sounds like the part where you say installation is straightforward and I lose the dining table.



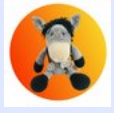
Herman

You would get the dining table back before the bootstrap finished.



Corn

Not an assurance.



Herman

The useful detail is that the voting page never receives model names. It gets opaque pair identifiers and media paths. Reading the page source won't reveal which model produced the left clip.



Corn

Good. Hiding a label with a bit of styling would only blind the partners who don't inspect things.



Herman

Each reviewer gets an individual identifier. That lets the system connect repeated judgments, including repeat pairs shown with sides swapped. It can catch someone consistently favouring the left side or reversing their own preferences.



Corn

I'm less worried about partners cheating than recognizing a house style. A model's particular camera glide could give it away.



Herman

Blindness can't erase that recognition. It can remove explicit labels and leaked filenames, and stop aggregate results influencing later votes. That's still worth doing. Do you think you'd trust yourself not to favour a model you'd already recommended?



Corn

I'd trust myself to explain the favouritism very persuasively afterward. Hide the names.



Herman

The package fits Bradley–Terry ratings with bootstrap uncertainty intervals. Its planted-bias demo shows repeat agreement of about sixty percent and a left-side win rate a little over fifty-four percent, with a statistically detectable position effect. Those are demo results, not a claim about architects. They show the checking machinery has something concrete to detect.



Corn

Repeat judgments also spend attention. I'd rather have fewer useful comparisons than a partner clicking through a punishment.



Herman

Interrupted generation runs resume, too. Repeating the same configuration skips successful cells. A rate limit halfway through shouldn't require buying the first half again.



Corn

Well, that's the tooling question, but before we get back to it, Mum's on the line. You'll be pleased, Herman.

D

Dorothy

Hello dear, it's your mother, with Dorothy's History of Toast.

D

Dorothy

Did you know that raising a glass is called a toast because people used to put a piece of spiced toast in their wine or ale?

D

Dorothy

And in Shakespeare's The Merry Wives of Windsor, Falstaff asks for a quart of sack with a toast in it. It's true.

D

Dorothy

I suppose it just goes to show that even a nice drink gets something soggy in it eventually. Anyway, I've got a load of washing on, so I'll let you go. Bye bye, sweetheart.



Herman

Lovely to hear from you, Mum. Now, resumable generation is a real advantage, but it doesn't turn that package into a hosted service.



Corn

Meaning a Replicate firm still needs somebody responsible for running it.



Herman

I haven't seen a Replicate-native blind-rating SaaS that closes that gap. The other open option I'd look at is GenMediaBenchmark, a local tool, bring your own keys, that runs one prompt across the fal and Higgsfield video models, logs speed and cost, and has a blind arena mode that turns votes into Elo with confidence intervals. It's built for one person comparing, more than for a partner panel.



Corn

So the honest answer for a Replicate shop is a small script and someone who owns it.



Herman

For a cheaper shortlist, Film Arena's Sandbox runs your own content through the top-ranked models, with best, top-three, and custom modes. And the public arenas, Artificial Analysis and LMArena, already rank image-to-video models from blind pairwise votes. Useful for picking the ten or twenty contenders, but that's strangers' taste on strangers' images, not the firm's own panel on its own renders.



Corn

Whatever they use, I'd want a both-bad button. Otherwise the least deformed building acquires a trophy.



Herman

Tooling gets the clips in front of people. The generation protocol decides whether those comparisons mean anything. Artificial Analysis publishes a useful starting point: output at ten-eighty resolution or the closest available, twenty-four frames per second, and ten seconds long.



Corn

Long enough for a confident first impression to become a structural concern.



Herman

They use a sixteen-by-nine frame and seed forty-two where exposed. The seed records a setting; identical seed numbers across different models do not create identical randomness.



Corn

I'd copy the discipline, not treat forty-two as an architectural requirement. A firm selling shorter clips could choose that duration instead, provided it fixes it before seeing which model benefits.



Herman

Prompt enhancement is enabled only when the model creator recommends it, and guidance and inference settings follow first-party defaults. That represents a recommended-use comparison, rather than forcing every model into settings that happen to suit one.



Corn

There's a judgment call there. Same written prompt doesn't mean the same effective instruction after a provider rewrites it. I'd want both versions recorded where available.



Herman

Reference images are lossless PNG, nineteen-twenty by ten-eighty. Avoiding JPEG compression removes an unnecessary source of input artifacts, especially around fine lines and repeated details.



Corn

Don't ask the model to preserve a railing you've already compressed into porridge.



Herman

If a provider requires a different size or format, record that transformation and upload the transformed bytes. No silent crop that removes the troublesome facade. Keep the endpoint identity, too. A cheaper or distilled variant can behave differently from the native model.



Corn

I would also keep failures visible. A model that produces one perfect clip after several failures could still be worth using, but the firm should know what that perfect clip actually cost.



Herman

Would you judge failures as losses, or keep them outside the visual score?



Corn

Separate availability and cost from visual judgment. An absent clip can't be rated for material fidelity. Record it as absent, rather than quietly dropping that render or pretending the failure never happened. Then the decision can consider both.



Herman

Hm. That also avoids giving unreliable models an easier test set by accident.



Herman

The human rubric is where I'd borrow Film Arena's one-dimension-per-vote design. Same task, anonymous outputs, choose the stronger result on a specified dimension. Asking about geometry separately from visual appeal prevents a beautiful sky from buying forgiveness for a sheared wall.



Corn

I'd make geometry a rejection criterion as well as a score. If the entrance changes width, the firm may have no use for that clip, whatever the average rating says. Averaging can make an unusable output look comfortably above average.



Herman

MechVerse offers a relevant expert-panel template. It compares fifteen image-to-video models across twelve human-rated dimensions. Ratings use a one-to-five scale. Its subject is mechanical assemblies rather than buildings, but rigid parts and permissible motion make it a much closer problem than cinematic attractiveness.



Corn

A hinge either stays attached or it doesn't. The sunset has no vote.



Herman

Its headline finding is that current models can preserve appearance and smoothness while generating motion the mechanism couldn't actually perform. Looking good and being right came apart. I'd borrow the separation of dimensions, not transplant all of its mechanical questions into an architecture review.



Corn

My rubric would separate structure, material and lighting fidelity, and camera execution. Structure includes window spacing, wall planes, edges and depth relationships. Materials means the same stone doesn't become wet plastic. Lighting means plausible changes as the view changes, rather than demanding every pixel remain fixed.



Herman

Glass is especially awkward there. A reflection should change with viewpoint. Treating every change as drift would punish a correct result.



Corn

We'd need a few anchored examples before voting. Not a lecture on taste. Show what counts as an unacceptable geometry change, what counts as a minor defect, and what they're willing to repair. Otherwise one partner is scoring usefulness and another is scoring ambition.



Herman

I suspect you'd also distrust the clip where nothing moves.



Corn

Correct. A still render has excellent structural consistency. It has already won that competition without entering.



Herman

Camera execution needs its own check for exactly that reason. Did it perform the requested motion, and was that motion smooth? A model shouldn't evade a difficult orbit by producing a nearly static zoom.



Corn

Why does an ordinary camera move remain so hard? The source image has the edges. It feels as though keeping them attached should be the easy part.



Herman

The image constrains one viewpoint, not the hidden surfaces. As the camera moves, the model must infer depth, reveal unseen areas, and maintain relationships across frames. A broad explanation in the geometry-correction work is that video training data contains abundant moving objects filmed by static or weakly moving cameras. It offers much



Herman (continued)

sparser supervision for appearance changes caused by viewpoint movement.



Herman

No... that explains part of the difficulty, not every model's behaviour. You can get warped structures, inconsistent parallax, texture drift, or a smooth camera move that effectively turns into a new shot.



Corn

The tour has entered a different property.



Herman

Automated metrics only partly help. VBench-I2V measures input subject and background alignment, consistency, image quality, aesthetics, motion smoothness, and whether the requested camera motion happens. Those can expose useful problems.



Corn

A building can count as the subject in that terminology. The trouble is resemblance. Same-looking windows don't establish that they're staying on the same plane.



Herman

Correct. Background coherence doesn't directly measure a wall's geometric reprojection across viewpoints. VBench even includes architecture in its input categories, but that doesn't make its output score an architectural integrity certificate.



Corn

I'd use those metrics to flag clips for inspection, not let them outvote the partners on a bent column.



Herman

Geometry-oriented systems such as VGGRPO distinguish camera smoothness from reprojection consistency. One penalizes jittery trajectories; the other addresses cross-view geometric coherence. Those are much closer to the firm's actual concern, but the implementations are research code rather than a turnkey building-checking service.



Corn

We also shouldn't pretend we can certify unseen space from one render. Once the camera goes behind a corner, plausible invention may be exactly what the system is doing. If the firm has the underlying scene, a conventional rendered camera path could provide a reference.



Herman

For real geometry, yes. Still, they don't need a huge panel to spot blatant deformation. A handful can find gross defects. Ranking close candidates confidently is another matter, and I don't know a published rater threshold for this particular setup.



Corn

Nor should eighteen reviewers on a product dashboard become a magic number. I'd retain individual ratings and inspect disagreements. A partner objecting to a changing staircase may be identifying the only defect that matters commercially.



Herman

A tight confidence interval also needs interpretation when the same few people rate many clips. More clicks don't automatically mean more independent expert opinions. Personally, I'd watch the opening at normal speed first. Nobody approves a walkthrough one frame at a time.



Hilbert

...and my aunt Vera insisted the camera go through the utility cupboard. Through it. The cupboard was the entire property. She had bought a lease on it and wanted a sales film.



Corn

Limited scope for a sweeping entrance.



Hilbert

Four feet across. I was a rostrum camera operator in the late eighties. We filmed drawings and property photographs. The skill was ending a move without the viewer feeling the stop. You could tell who had it from the last inch. Vera wanted a three-minute tour of the cupboard, so I moved across the same photograph until we reached the left hinge, then dissolved back to the right hinge. She called that the return route.



Herman

A repeated image would at least preserve the dimensions.



Hilbert

It did. Corn, lift your level a touch. That last cupboard line fell away. The little illuminated viewer sits beside my telephone now, plugged into the wall socket. I use it to check transparencies. Vera paid me with it. Its switch needs a firm thumb. Herman, your normal-speed first pass is useful, but we checked individual frames before clients saw the film. Before it. A hair in the gate doesn't improve when you run it faster.



Herman

Fair correction. I put the presentation pass ahead of the technical check.



Hilbert

You took the camera movement apart quite carefully earlier. Same business with the stop. You had to judge the move and inspect the image separately. Vera's agent watched the whole three minutes. He wrote down that the return route took longer than the outward journey. It was the same photograph. I shortened the second dissolve. He approved it after lunch. The agent needed a film he could show without explaining it.



Corn

...A private panel can do both passes, then. Technical inspection before the meeting, normal playback for whether the result works as a presentation. Frame stepping afterward to locate disputed defects, rather than making everyone scrub every clip.



Herman

That gives the record useful evidence beyond the rating. A timestamp and a short note, entrance widens here, are actionable. A low score alone leaves the next person guessing what went wrong.



Corn

I'd keep the partners' first votes independent, too. Discuss afterward. Otherwise the most senior person notices a wobble and suddenly everyone has always hated that balustrade.



Herman

One adjacent detail worth keeping: the public methodology distinguishes standard endpoints from modified ones that may change fidelity to reduce cost or increase speed. Comparing model names alone can conceal that distinction.



Corn

You're buying an endpoint's actual output, not the family reputation.



Herman

For this firm, I'd use GenMedia if fal generation and a hosted workflow fit, or media-gen-arena if Replicate and self-hosting fit. Copy the published control discipline, judge geometry separately from polish, and keep the uncertainty visible. The useful next test is whether the chosen clips can be shown to a client without somebody explaining away the building.



Corn

Thanks to Hilbert Flumingtop for producing the show.



Herman

If this was your kind of episode, go back for episode twenty, Architectural AI; and episode sixteen, On Deepfakes, SynthID, And AI Watermarking. This has been My Weird Prompts, the human-AI collaboration podcast.



Corn

If you listen regularly, leave us a review. It helps other curious people find the show.



Herman

See you tomorrow.