How I Build a Full UGC Ad With AI — From Product to Raw Footage

A behind-the-curtain walkthrough by Phil Grabowski / Creative Cat

You asked about UGC, so here's the whole thing — the actual method I use to build a finished UGC ad without a film crew, a creator shoot, or a week of back-and-forth. Not the hype version. The real five-stage pipeline, the tools, and the principles that decide whether the footage is usable or garbage.

I'm giving the "how" away on purpose. The value isn't the secret — it's the reps, the taste, and knowing which knob to turn when a shot breaks. Read this and you'll understand exactly how a modern AI UGC ad gets made.

The shape of the whole thing

Five stages, in order:

1. **Product** — a clean, ad-ready render of the thing you're selling
2. **Creator / avatar** — a consistent, relatable on-camera person
3. **Storyboard / script** — a 6-beat ad structured to sell, not just to look pretty
4. **Raw footage** — the AI renders every clip
5. **Human edit** — an editor assembles the final cut

The engine is **Claude wired into Higgsfield** (over MCP), driving **GPT Image 2** for stills and **Seedance 2.0** for video. Claude generates *all* the raw footage. A human editor only does the final montage — captions, music, grade, pacing. That division of labor is the whole trick: the machine does the volume, the human does the taste.

Stage 1 — Product

Everything downstream inherits the product shot, so it gets built first and built well.

Tool: GPT Image 2 (via Higgsfield). **Flow:** generate a batch of low-res drafts in one call, pick the strongest, regenerate a few variants on feedback, then lock one clean high-detail final. Shoot it **3:4 portrait** — you're feeding video, not a square catalog grid.

The product-design rules are non-negotiable, because this is where most AI product shots quietly look fake:

- **The label wraps the bottle 360°.** A rectangular sticker slapped on the front reads as a mockup, not a product. Wrap it.
- **Premium-but-mass-market.** Think a clean bottle, a brushed-metal cap, an embossed brand, a serif name — the RITUALS shelf feeling. Looks expensive, sells to everyone.
- **Not sterile ultra-minimalism.** A bare white bottle reads as niche and unfinished, not ad-ready. Give it presence.
- **Slim, rounded, aspirational shape.** Vertical text on slim bottles.

One more, and it matters later: **don't put camera-body specs in product prompts.** Save the lens talk for the video stage.

> **The principle:** a great reference beats a long prompt. The product still is the most-reused asset in the entire ad — it shows up in the cutaway, the demo, the close. Invest here and everything after it gets easier.

Stage 2 — Creator / Avatar

Now the person. Same engine (GPT Image 2), but it starts with a short intake every time: market, gender, age, ethnicity, hair and eyes. Defaults exist, but you fill the blanks deliberately.

Two look rules carry the whole stage:

- **For the US market, say it out loud.** Write "American / US social-media creator vibe, raised in the US" explicitly. Skip it and the model drifts toward country-of-origin looks. The casting note has to be in the prompt.
- **Relatable, not editorial.** You want the normal, natural, approachable girl- or guy-next-door — the person whose video you'd actually stop on in the feed. A supermodel kills UGC. The whole format runs on "this looks like a real person who bought the thing."

Then you build a **4-angle character sheet** (front, left profile, right profile, full profile) so the face is locked from every direction. The technical move that keeps identity stable: **cut the sheet into four separate frames and upload those as the custom avatar — never the grid.** Feeding the model a grid causes identity drift; separate frames keep the face consistent shot to shot. Check one frame to confirm the face isn't cropped before you commit.

(Practical aside: image models hard-block underwear even "tasteful." Use activewear — sports bra and leggings — and you get the same energy without fighting the safety filter.)

Stage 3 — Storyboard / Script

This is where the ad gets its spine. The structure is a validated **6-beat "Disguise / Sell"** sequence — distilled from tearing apart ~146 winning ads:

1. **Hook (~1.5s)** — a negative emotion, a problem, or a curiosity gap. *Never the product itself.* Most great hooks are a to-camera exclamation with a twist.
2. **Relatable scene** — grounds it in a real moment so the viewer leans in.
3. **Aikido reveal** — the product arrives as the inevitable answer, not as "and now a word from our sponsor."
4. **Proof** — a live, uncut demo. The effect happening here and now, with an oddly-specific number or mechanism.
5. **Reviews** — the skepticism shield. "Don't believe me? Look at the reviews."
6. **Close** — value, scarcity, "link below."

Pick the engine before you write a word: - **Fast demo, 30–45s** — TikTok-Shop-style product, demo plus offer. - **Long-form mini-VSL, 60s+** — founder story or education.

Don't make 15-second ads. Too short to hook, prove, and close.

The script runs on **two layers at once**: the top layer is emotion and loose creator-talk (the self-interrupt — "wait, look how it just—"; the reframe — "it's not you, it's the X"); the bottom layer is hard fact (the ingredient, the derm/vet-tested line, the "X 5-star reviews"). Emotion pulls; facts justify.

For consistency across shots, you anchor three reference elements: the **character** (the 4-frame avatar locks the face), the **prop** (the hi-res product with a legible label), and the **environment** (without it, the room quietly changes between clips). Shoot it **mirror-POV** — the camera *is* the mirror, the creator looks into the lens.

Two principles worth stealing

Simple prompt + great reference. A good reference image plus a short prompt with two levers — *what the object does* and *how the camera moves* — beats a long, clever prompt riding on a weak image. **Camera movement is the single biggest lever you have.** Get the reference right and the prompt can be almost boring.

Direct it; don't choreograph it. Don't get in the model's way. Tell it *what is said*, not how every hand moves. Emotion is the main performance dial. And negation works — telling it "no smile" flips a default smiley face into genuine frustration. You're a director giving notes, not an animator keyframing fingers.

The two UGC approaches — choose on purpose

- **(A) Talking-head with native dialogue.** Most authentic, hardest for AI (lip-sync, teeth). Great for volume once you accept some misses.
- **(B) No-dialogue performance + voice-over in post.** This is the default going forward, and here's why it's underrated: no talking means no broken mouth, which means cleaner raw footage and a far higher keeper rate. It's also **language-agnostic** — silent visuals plus any-language VO equals cheap localization, and you reuse one visual set across many ads.

If you take one tactical thing from this whole document: **shoot silent, voice it in post.** Variance drops, reuse goes up.

Stage 4 — Raw Footage

Now the machine renders every clip. Engine: **Seedance 2.0 (Standard tier)** via Higgsfield, **1080p**, vertical **9:16** set in the UI (not in the prompt). Never use the fast mode — it caps lower and introduces jitter.

Tempo is set on generation, not fixed in post. Roughly, **duration ≈ words ÷ 2.3** (natural speech sits around 2.3 words/second), rounded to the nearest allowed length. Hit the cadence when you render rather than generating long and speeding it up later.

Prompt template, in order: `[shot / framing] · [subject + ONE clear action] · [environment] · [camera movement] · [lighting + atmosphere] · [visual style / grade] · [one named camera body]`

Pre-flight checklist before you hit GO: - No "fast" anywhere — use *decisive* / *snap* / *kinetic* / *controlled-quick* - Every shot has motion; every shot names a light source - **One action per shot** (this is the big one) - Subject appears in the first ~20–30 words - One camera body, one grade, one lens family; keep references lean - Camera moves stay simple or phased — not every shot is frenetic - The final beat keeps moving — no freeze-frame ending - **Reference images are hi-res — especially the product.** A low-res product yields a garbled, mirrored label. This is preflight #1.

The UGC difficulty ladder

Risk compounds the more you ask the model to do. Plan your shots around it:

Asking the model to…	Risk
Hold a static product	none
Move a product	easy
Big body / arm motion	medium
Expressive face **without talking**	medium — the sweet spot for approach B
Precise fingers (applying, blending)	hard — the #1 break
Talking face (lip-sync)	hardest

Re-roll only for the unfixable — wrong identity, broken product label, plastic-looking skin, wrong line read, deal-breaker artifacts on hands, face, mouth, or eyes. **Don't re-roll for fixable things** — tempo, cut point, music, captions, clip selection. Those are edits, not regenerations. Burning a generation on something the editor solves in ten seconds is the most common waste in this whole pipeline.

(One generation maxes around 15s, so a 30s ad is two passes joined in the edit — carry the last frame of the first into the second for continuity.)

Stage 5 — The Human Edit

This is where a person takes over, and it's deliberate. The AI delivered clean raw footage; the editor delivers the *ad*.

Two 15s clips get joined into one continuous shot — normalized to matching resolution, frame rate, and audio with ffmpeg — then captions, music, and color grade go on in a post layer (DaVinci Resolve / Premiere), and the captions carry the hook. Default montage tempo runs a touch quick; dense demo beats get a gentle bump, sparse hook beats a slightly bigger one. **Never slow a clip down** — it looks ugly. If a beat needs more room, regenerate it longer instead.

The honest part: **a human assembles the final cut.** The AI gets you most of the way with footage no shoot could match for speed, and a real editor's judgment turns a pile of clips into something that actually converts. Both halves matter.

The short version

- **Reference quality beats prompt length** — fix the still before you animate it.
- **Direct like a media buyer** — one message per clip, captions carry the hook, product cutaway at the doubt beat.
- **One action per shot.** Risk compounds; keep each clip asking for one thing.
- **Hi-res product reference, always** — it's the most common failure and the easiest to prevent.
- **Shoot silent, voice in post** when you can — cleaner footage, instant localization, more reuse.
- **A human assembles the final cut.** The AI does the volume; the editor does the taste.

If you want one of these built

I build these UGC ad systems for a small handful of DTC brands and agencies — product render through raw footage through final cut, the full pipeline above, done for you. I keep the roster short on purpose so the craft stays high.

If your brand or your clients could use a steady supply of UGC ads built this way — fast, on-brand, and actually made to sell — **reply to the message that sent you this, or reach out directly.** Tell me what you're selling and who it's for, and I'll tell you straight whether this is a fit.

— Phil Grabowski, Creative Cat