MPHORA × ADOBE USER RESEARCH EDITION

VIVID User Research Test Case.

Moderated User Research study — Photographer editing pain points across two frontier image-editing models, told through the photographer's own voice.

AUDIENCE
Adobe User Research scientists

LENS
Real workflow friction · Verbatim photographer voice

DATE
2026-05-04

STUDY AT A GLANCE

Eighty moderated sessions, in one record.

80	8	5	2	477	445
MODERATED SESSIONS	PHOTOGRAPHER PERSONAS	CANONICAL WORKFLOWS	FRONTIER MODELS	FRICTION EVENTS	WORKAROUNDS CITED

DESIGN

Full-factorial: 8 personas × 5 workflows × 2 models. Five-phase moderator protocol held constant across every run; run order randomized inside the parallel scheduler.

WHAT WAS EXTRACTED

Every event is verbatim-quote-anchored against a 13-category taxonomy. 80 magic-wand feature-request quotes captured at session close-out.

THE QUESTION WE ASKED

“Have photographers walk through their real editing workflow, so we can uncover where the friction is and what slows them down the most.”

User Research FRAMING THAT SEEDED THE STUDY

We ran that interview **80 times** — across 8 distinct photographer personas, 5 canonical workflows, and 2 frontier image-editing models — and kept the full moderated transcript, every generated image, and every extracted friction event in a fully audited record.

METHOD & RIGOR

Five-phase moderator protocol, held constant across all 80 runs.

PHASE 01

Intro

Moderator introduces the session; participant grounds themselves in their real workflow.

PHASE 02

Task brief

Moderator hands a specific scenario — e.g. “2-hour wedding deadline, 20 images.”

PHASE 03 · CORE

Think-aloud edit

Participant narrates while the model executes — 3 to 5 edit turns per session.

PHASE 04

Friction probes

Unscripted follow-ups: “where did that feel fragile?”, “did you almost abandon?”

PHASE 05

Magic wand

“If you had a magic wand, what would this tool just do for you that it doesn't today?”

CAPTURED PER SESSION
Transcript · 20–22 turns

IMAGE EVOLUTION
3–5 PNGs/session

FRICTION EXTRACTION
13-cat · quote-anchored

EXPENSE LEDGER
Per-call JSONL

PARTICIPANTS

Eight photographers, sampled for spread — not balance.

Oliver Martinez
Expert · EN
Age 31 · USA West

Christopher Anderson
Expert · EN
Age 25 · UK

Kevin Taylor
Expert · EN
Age 34 · India South

Mark Garcia
Intermediate · EN
Age 36 · USA

조서준
Intermediate · KO
Age 24 · Korea

Intermediate · ZH
Age 36 · China Tier 1

Mia Schulz
Novice · DE
Age 62 · Germany

Christopher Anderson
Novice · EN
Age 65 · USA

Sampled from the VIVID PSA pool (occupation = photographer; 32 candidates) across age 24–65, four languages, three expertise tiers, and seven culture proxies — sized for failure-mode discovery, not significance testing.

FIVE CANONICAL JOBS

The workflows photographers gave us.

W1 — TIME-PRESSURED

Deadline retouch

Wedding photographer owes the client 20 delivery-ready images within 2 hours of reception end.

DEADLINE · CLIENT TRUST

HOTTEST

W2 — 126 EVENTS

Client revision loop

“Warm the skin tones, deepen the sky, everything else must stay exactly as-is.”

REST-UNCHANGED FIDELITY

W3 — EDITORIAL

Series consistency

A lookbook of 5 images shot in slightly different lighting that needs to share one editorial look.

SERIES UNITY · POLISH

W4 — CAPABILITY TEST

Shot recovery

A portrait with a missed focus — eyelashes slightly soft. Re-shooting is impossible.

CAPABILITY CEILING

W5 — CREATIVE

Mood transform

Same landscape, turned from “happy sunny afternoon” to “dramatic pre-storm.”

CREATIVE REASONING

PART TWO · FINDINGS

Both models fail the workflow. Just in opposite directions.

Hold the workflow constant, vary only the model — and two distinct failure personalities emerge. The product implications differ; you have to pick the one you're optimising for.

FINDING 1

Two failure modes, not one.

GPT IMAGE 2

Over-commits.

Attempts every request — but silently edits regions the photographer pinned as “leave alone,” drifts subject geometry across turns, loses earlier constraints by turn 3.

~150	annoying-severity events (more attempts, more misses)
27	silent unrequested edits across study
80	over-correction events
low	blocker / catastrophic count

NANO BANANA PRO

Under-commits.

Refuses more often (sometimes correctly, sometimes when it should have attempted), abandons mid-edit on hard recovery tasks, and ships more session-killing failures when it does fail.

higher	blocker-severity rate when it does fail
82	subject / geometry drift events (study-wide)
75	instruction-dropped-over-turns events
87	capability-ceiling crossings (study-wide)

FINDING 2

“Rest unchanged” is the most load-bearing constraint — and both models break it.

> *“Warm the skin tones, deepen the sky, everything else must stay exactly as-is.”*
>
> **Client revision brief** · the constraint every photographer in the study expressed

When the brief is local-only, photographers want a **visual region-lock** — masks they can trust — not a longer prompt that the model may or may not respect.

27 SILENTLY-APPLIED UNREQUESTED EDITS	**82** SUBJECT / GEOMETRY DRIFT EVENTS
75 INSTRUCTION-DROPPED-OVER-TURNS	**126** W2 FRICTION EVENTS (MORE THAN THE DEADLINE WORKFLOW)

Multi-turn revision — where the constraint set grows turn by turn — breaks cross-turn consistency faster than a single high-pressure one-shot.

FINDING 3

Capability ceiling without a refusal — the scariest category for trust.

EVENTS ACROSS THE STUDY

87

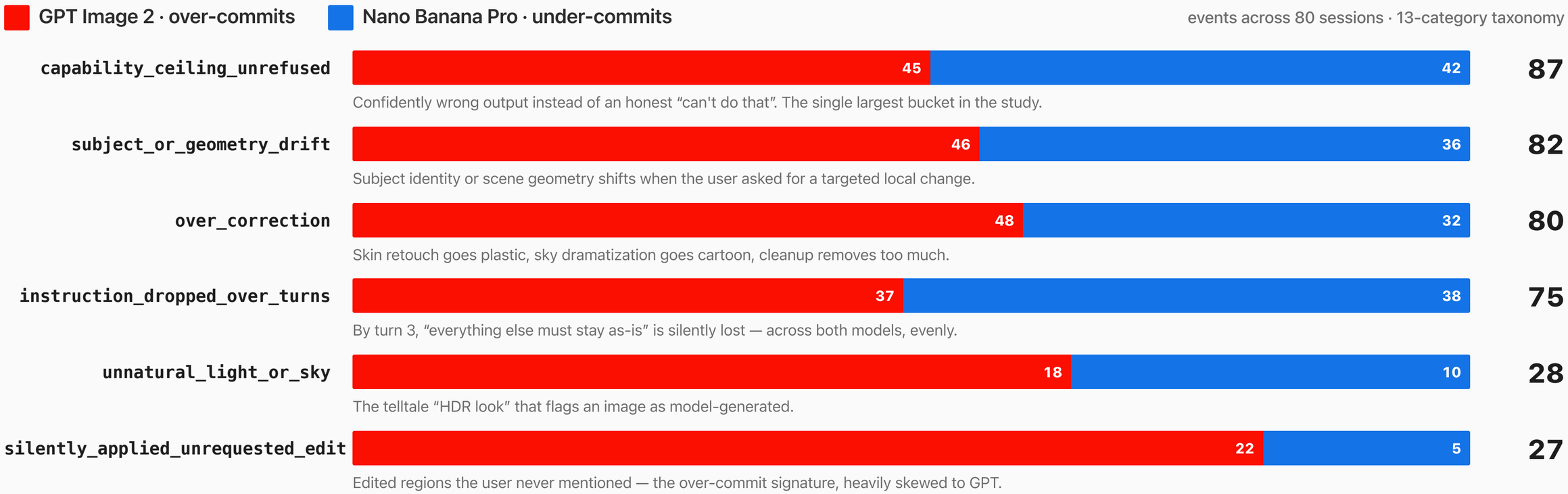
LARGEST FRICTION CATEGORY IN THE STUDY

Both models often produce a confidently wrong output instead of saying “I can't do this.” Photographers describe abandoning sessions within 2–3 turns once they spot the pattern.

> *“I'd rather Aussortieren than push a model to fake what isn't there.”*
>
> **Mia Schulz** · novice · Germany · W4 Shot Recovery

TOP FRICTION CATEGORIES · SPLIT BY MODEL

Where the breakage actually lives.



VERBATIM

The photographer, in their own words.

“I'd make it say, in plain English, that the edit is restricted to two isolated regions only and that no other part of the image may be modified under any circumstance. That's the closest thing to a hard lock the model tends to respect.”

Oliver Martinez · expert · USA · W2 Revision Loop · GPT Image 2

“What I most need is a control that locks things by region. The model tends to adjust everything together, so it would be far more convenient to separate and lock the adjustment range — one skin mask, one sky mask.”

조서준 · intermediate · Korea · W2 Revision Loop · GPT Image 2

“The most troublesome failures aren't the ones you can spot at a glance. They're the kind where the result looks a little better, but it's no longer the original photo.”

· intermediate · China · W2 Revision Loop · Nano Banana Pro

“I stop trying to rescue it, honestly. If the eyes are the whole portrait and the lashes are still mush after a restrained pass, there isn't a clean workaround left.”

Christopher Anderson · expert · UK · W4 Shot Recovery · Nano Banana Pro

FINDING 5 · CONVERGENT MAGIC-WAND REQUEST

Visual region-lock — in three languages, the same ask.

“Let me lock the region the model can't touch.”

The most-frequent magic-wand quote across **80 sessions**, **three languages**, every expertise tier.

Photographers describe wanting masks they can *trust* — not text instructions the model may or may not respect. The W2 cluster converges on: **skin-mask + sky-mask + everything-else-locked.**

THE CONSENSUS DESIGN ASK

01 A mask the photographer authors, not one the AI guesses

02 A visual diff preview before the edit commits

03 Pinned constraints that persist across turns — no re-stating

PART THREE · IMPLICATIONS

Six directional reads for the Adobe roadmap.

Hypothesis-grade — the kind of thing worth taking into a roadmap discussion, not contractual claims. Each maps to specific findings; each names the surface it points to.

SIX PRODUCT IMPLICATIONS

From friction events to roadmap moves.

01

FIREFLY · EXPRESS

Build an “honest refusal” affordance.

Capability ceiling without refusal is the #1 friction category. Surface a confidence band — or an explicit “not supported on this image” — instead of a confident-wrong output.

F3 · 87 EVENTS

02

PHOTOSHOP · FIREFLY

Ship visual region-lock as a primitive.

The consensus magic-wand request across three languages. First-class affordance in generative-fill and edit-image — with a visual diff preview before commit.

F5 · W2 CLUSTER

03

PHOTOSHOP · FIREFLY

Persist pinned constraints across turns.

By turn 3, “everything else must stay as-is” is silently lost. Pinnable instruction chips that the canvas re-asserts on every model call.

F4 · INSTRUCTION-DROP PATTERN

04

PHOTOSHOP · EXPRESS

Different models, different UX surfaces.

Over-commit is friendlier where iteration is cheap (Express); under-commit is safer for professional retouch (Photoshop). Or surface an explicit “safe” vs “creative” mode.

F1 · TWO FAILURE MODES

05

LIGHTROOM · PHOTOSHOP

Series consistency as its own surface.

One anchor frame drives the look; AI propagates targeted skin / black-point / temperature matches. Colorist on a propagation pass — not one image at a time.

W3 · SERIES-CONSISTENCY CLUSTER

06

PHOTOSHOP PRO TIER

Document the capability frontier.

Shot-recovery has known physical limits. Surface them; don't let the model lie. Closer to a creative co-pilot than a magic black box.

W4 · CAPABILITY-CEILING, BLOCKER-HEAVY

HOW VIVID ENABLED THIS

Speed, cost, reproducibility — at production fidelity.

SPEED

~58 min

80 sessions, wall-clock, at 8× parallelism with provider concurrency budgets respected. A real moderated panel of 8 photographers takes 3–6 weeks of recruiting, scheduling, conducting, coding.

COST

$499.11

Total customer price = 3,327 credits @ $0.15, ledger-audited. A real moderated User Research with transcript coding at this scale is typically $40K–$80K. Useful as a pre-flight for protocol design before the real budget commits.

REPRODUCIBILITY

Cell-level cache

Resumable: pass-1 cached sessions survived a crash; pass-2 only executed the unfinished cells. Same persona × scenario × model + same model versions → the same study.

REAL VS. NOT CLAIMED

What this study is — and what it isn't.

WHAT'S REAL

— **The images.** All 221 generated images came from the actual GPT Image 2 and Nano Banana Pro endpoints; preserved on disk.

— **The cost accounting.** Every SDK call's usage metadata + billable charge is JSONL-logged; reconciles to provider billing within rounding.

— **The friction extraction.** No category was pre-seeded; every event is verbatim-quote-anchored against a 13-category taxonomy frozen after pre-pilot.

— **The personas.** Drawn from a pre-existing PSA pool with persistent Big-Five + voice + behavior traits — not invented for this study.

WHAT WE DON'T CLAIM

— **Not a human user study.** Participants are LLM-driven personas — they simulate a photographer's narration, not a photographer's behaviour in real software.

— **Moderator + extractor are LLMs.** Buys protocol consistency at the cost of improvisational follow-ups; IRR vs. trained human coders not validated.

— **Per-cell n is 1 session.** Descriptive design — read deltas as direction, not as significance.

— **No ground-truth correct edit.** We measure participant-reported friction, not pixel-level correctness against a gold image.

VIVID × ADOBE USER RESEARCH EDITION

A high-fidelity preview of your next moderated panel.

Use this study to narrow which three questions to take into a real lab — not to ship features. The friction events, magic-wand quotes, and product implications are a structured artifact for hypothesis generation, with every claim traceable to a specific transcript and a specific image.

EXPERIMENT
`ur_moderated_full`

GENERATED
2026-05-04

SESSIONS
80 · audited end-to-end

Want to take this further — a different participant pool, a deeper protocol, or a custom taxonomy for your team?

TALK TO MPHORA

mphora.ai

EMAIL contact@mphora.ai

WEB mphora.ai

BEST FOR Custom panels · protocol design · transcript-coded findings · pre-flight before moderated lab work