The Consistency Problem in Generative AI

Why the tools that dazzle in a demo fail in production, what that costs the people who actually make things, and what closing the gap really requires.

> Generative models are built to surprise you. Professional work needs the opposite. That single mismatch is the consistency problem, and it is the line between an image that impresses and one you can ship.

DEPIX
Design intelligence for people who make real things
depix.ai · 2026

00 Executive summary

One striking image is easy now. Almost anyone can type a sentence and get something beautiful back. The hard part, the part every real workflow depends on, is the tenth image that still matches the first, and the ten-thousandth after that. Making the same product, the same character, the same brand look hold steady while everything around it changes is a different task from making one arresting picture, and today's general-purpose tools are built for the second, not the first.

This is not a rough edge that a better prompt smooths over. It is a distinct, measurable, still-open problem that the research community has a name for: consistency, and its failure mode, drift. Peer-reviewed work through 2025 and 2026 treats subject identity in images and temporal coherence in video as unsolved challenges with their own dedicated benchmarks. The most advanced labs say so in their own launch notes. The cause is structural, which is why prompt tricks, fixed seeds, and per-project fine-tunes reduce drift but never remove it and never scale: these systems draw a fresh sample from a vast range of possibilities every time, and a text prompt is a lossy way to describe one exact thing.

For anyone selling a real product, iterating a real design, or building a brand, drift is not a curiosity. A generated picture of a product is a promise a customer will hold the physical object up against. A campaign is not one image but fifty that have to look like one voice. A concept is not one render but the same object seen across angles, materials, and light. When the subject will not hold, the output is not an asset. It is a draft you cannot use.

Depix exists to close that gap. Depix Design Lab and the Depix MCP are built for one job that general models treat as a hard side-effect: keep the subject accurate and consistent while you change the world around it on purpose, across images and video, at the scale real work demands. This paper sets out the problem, the evidence, and the standard any honest solution has to meet.

> **A note on sources.** Every quantitative claim in this paper is tied to a named primary study or a dated first-party statement, listed in the references. Where reliable public data does not yet exist, we say so rather than borrow a number from a marketing blog. Several widely circulated figures about AI imagery did not survive that check and were left out on purpose.

01 The gap between the demo and the deliverable

The demo is genuinely dazzling, and that is the trap. A single prompt returns a single beautiful frame, and it is easy to conclude the problem of visual production is solved. It is not, because almost no professional deliverable is a single frame. A product listing is a set. A campaign is a system. A design review is the same object seen a dozen ways. A film is a sequence in which every frame has to agree with the last about what it is looking at.

The moment the job stops being "make one good image" and becomes "make this exact thing again, and again, unchanged, while the scene changes," general-purpose generators start to wobble. The shoe grows a stripe it never had. The character's face shifts between panels. The car gains a vent in one frame and loses it in the next. Each output is individually plausible and collectively incoherent. The tool did not fail to make something good. It failed to make the same good thing twice.

> Novelty is the demo. Consistency is the product.

That distinction is the whole subject of this paper. It is not a matter of taste or polish. It is a specific technical property, it is measured with specific instruments, and it is the property that decides whether generative AI can carry real commercial and design work or only pitch it.

02 What "consistency" actually means

Consistency is not one thing. In practice it breaks into four distinct demands, and a professional tool has to satisfy all of them at once. Most general models are strong on the first impression and weak on every one of these.

1. Subject and identity consistency

The same subject stays the same subject across many images: the same product, the same character, the same face, with the same proportions and details, no matter how the pose, angle, or setting changes. This is the axis authors, studios, and product teams feel first, because a mascot that changes face or a SKU that changes shape is not a variation, it is a different thing.

2. Colour and material accuracy

The exact colour, finish, and material hold true and repeat: a specific paint, a specific leather, a specific hue that matches a real reference and does not drift a shade between renders. This is deceptively hard, because colour is continuous and language is not. The words "warm red" point at a whole neighbourhood of reds, and the model picks a different one each time.

3. Temporal consistency

In video, every frame has to agree with the frames around it. A short clip is well over a hundred frames, and each one must hold the subject, the background, and the light steady unless the shot means to change them. When it fails you see it instantly: flicker, textures that crawl, a subject that subtly morphs, a background that slides.

4. Brand and style coherence

Across a whole set, the palette, the light, and the feel read as one intentional system rather than a pile of one-offs. A brand is defined by its consistency. Fifty assets that each look like a different artist made them are not a campaign, however good each one is in isolation.

> **FINDING**
>
> The research community now treats consistency as its own measurable capability, separate from image quality and from prompt-following. In other words, "looks good" and "did what I asked" are not the same as "stayed the same." Benchmarks score them apart.
>
> DreamBench++ (ICLR 2025) scores personalized generation on two separate axes, concept preservation (holding a subject's shape, colour, texture, and features) and prompt following, and reports them independently. VBench (CVPR 2024) decomposes video quality into sixteen dimensions that explicitly include Subject Consistency and Temporal Flickering as distinct, separately scored failure modes.

03 Why generative models drift by default

The instinct is to blame the prompt. Write a better description, the thinking goes, and the model will hold the line. It will not, and the reason is worth understanding, because it explains why this is a problem to be engineered around rather than a habit to be corrected.

Generation is a sample, not a specification

A modern image or video model does not execute a plan. It draws a sample from an enormous space of plausible outputs. A recent survey frames the task precisely, as "a sequential sampling process from a high-dimensional spatiotemporal distribution." Ask twice and you are drawing twice from that space, and the space is designed to be rich. The same machinery that makes the output fresh and surprising is what makes it different every time. Novelty is not a side effect. It is the objective the system was optimised for, and consistency is the opposite of it.

A prompt is a lossy description of an exact thing

Language cannot pin a specific object down to the pixel. This is clearest with colour. As the authors of a 2024 study on colour in diffusion models put it, a name like "beige" or "light green" "may not precisely match the intended color," because "language represents color in a discrete manner, whereas color is a continuous concept." The same looseness applies to a jaw line, a stitch, a wheel, a logo. The model knows products, faces, and cars in general extremely well, and your specific one not at all, so when it has to show a detail it was never given, it invents a plausible one. Plausible is worthless when a customer is comparing the render to the object in their hands.

> The model was never given your product. It was given a paragraph about it, and it fills the gaps with something that merely looks right.

Drift compounds in video

Video makes the problem worse in a specific, well-documented way. When a model generates a clip frame by frame from its own previous output, small errors do not stay small. They feed forward and accumulate. Researchers describe this as models being trained on clean reference frames but forced at generation time to build on their own imperfect ones, so "errors compound and quality drifts over time." Over a longer sequence this shows up as "attribute drift," the slow mutation of the very things that were supposed to stay fixed. This is why temporal coherence is widely described in the literature as one of the hardest open problems in AI video, and why an entire subfield is still working on it in 2026.

Why the usual workarounds do not close it

Fixed seeds, elaborate prompt engineering, reference tricks, and per-subject fine-tuning all help, and none of them solve it. They reduce the variance, they do not remove it, and crucially they do not scale: a technique that gets one character roughly consistent across five images by hand is not a pipeline that keeps a hundred products accurate across a catalogue and a video library. The gap between "I coaxed it once" and "it is reliable at volume" is exactly where professional work lives.

In fairness. None of this means general models "cannot" hold a subject or change one attribute cleanly. For broad, global style changes they often can, and the field is improving fast. The honest and important claim is narrower and it is the one that matters commercially: holding an exact subject while deliberately changing one thing about the world around it is *hard and unreliable by default*. Getting it dependable, repeatable, and accurate at scale is a distinct engineering problem, and that is the problem Depix is built to solve.

04 This is not our opinion. The field measures it, and the top labs admit it

It would be easy to dismiss all of this as a vendor talking its own book. So it is worth being clear that the case rests on the field's own peer-reviewed instruments and the leading labs' own words, not on ours.

Consistency has its own benchmarks

When a research community builds dedicated measuring tools for a property, that property is real and it is not yet solved. VBench, presented at CVPR 2024, breaks video quality into sixteen fine-grained dimensions and scores Subject Consistency and Temporal Flickering as their own line items. VBench-2.0, from 2025, goes further and separates "intrinsic faithfulness" from mere "superficial faithfulness," and finds that models which score increasingly well on surface metrics "still struggle to generate videos that are not just visually plausible but fundamentally realistic." For still images, DreamBench++ (ICLR 2025) scores identity preservation and prompt-following as two separate numbers, precisely because a model can obey your instruction and still fail to keep the subject the same.

Subject consistency "remains a challenge," in the literature's own words

A paper accepted to ICLR 2026 opens by stating that keeping "a consistent subject identity across diverse scenes remains a challenge for text-to-image models." Independent 2025 and 2026 surveys covering hundreds of methods reach the same conclusion for video, naming consistency across subjects and long-sequence fidelity as open problems rather than settled features.

> FIRST-PARTY ADMISSION
>
> The company behind one of the most advanced image models on the market lists character consistency among the things it is still working to make more reliable, inside its own launch announcement.
>
> Google Developers Blog, introducing its state-of-the-art image model (August 2025): the team says it is "actively working to improve long-form text rendering, even more reliable character consistency, and factual representation like fine details in images." A capability being made "even more reliable" is, by definition, not yet fully reliable.

Put those together and the picture is unambiguous. The best available tools make beautiful images. Their own makers and the independent benchmarks agree that keeping a subject consistent, in stills and especially in motion, is still an open problem. That is not a knock on the technology. It is the exact seam this paper is about.

05 Where it breaks: five places drift stops being cosmetic

Drift is a nuisance in a hobby project and a liability in a business. Here is where it becomes concrete, in the places our customers work.

Product and commerce: the picture is a promise

Most commercial imagery is not inventing anything. It is representing something that already exists, a specific item in a specific finish, and the customer will receive the physical object and hold it against the image. Ask a general model to re-render that product from a new angle and it improvises the details it was never given: the stitching, the proportion, the exact colour. Subtle enough to pass a quick review, wrong enough to matter when the box arrives. For an invented mascot that is a quirk. For a real SKU it is a claim about the product that is not true.

Brand and campaign: fifty images, one voice, or no campaign

A brand is its consistency. Generate a week of assets from prompts and you can end up with fifty images that each look like a different artist made them, one warm, one clinical, one painterly, the rest generic. Speed without coherence is not a saving. It is a fast way to produce a hundred assets you then cannot ship because no two of them belong together.

Character and IP: the face that will not hold

Storytelling of every kind, books, comics, games, film, depends on a character being recognisably themselves from one frame to the next. This is the single axis the research community most often names as unsolved, and it is why creators reach for AI to accelerate a project and then abandon it when the hero will not stay on-model. The tool that promised speed becomes the reason the schedule slipped.

Video: the hardest case

Everything hard about a still multiplies in motion. A clip is well over a hundred chances to disagree with the last frame, and the accumulation problem described earlier means the drift builds as the shot runs. For commercial use this is disqualifying: a product spot where the product shimmers, a brand film where the hero restyles itself between shots, an explainer whose character will not hold, none of it ships. And the two things a brand most wants at once, hold the subject rock-solid while changing the light, colour, or setting on purpose, are precisely the two the models are least reliable at combining.

Industrial and automotive design: the same object, or it is not iteration

This is where the stakes are sharpest, and it is where Depix started. Design is iteration: the same concept seen across angles, materials, colours, and environments so a team can judge it. If the object changes shape every time the scene changes, that is not a design cycle, it is noise. A concept car has to be the same car in the studio and on the road, in silver and in matte green, from the front and from above, or the images are not comparable. Holding an exact form steady while deliberately varying its finish and its world is the daily job of concept-phase design, and it is exactly the job general models do least reliably.

06 The cost of drift

The technical certainty that these tools drift becomes, downstream, a set of real costs: assets that cannot ship, work redone, campaigns that cannot go out because the set will not cohere, and time lost to reviewing every frame for the small mutation that slipped through. Those costs are easy to observe in any team that has tried to scale generative imagery, and hard, so far, to quantify cleanly in public research. We would rather be honest about that than reach for a borrowed statistic.

What can be said with named evidence is that consumer trust in AI within the buying journey is fragile and easily lost, which raises the price of getting the image wrong.

38% / 30%

of US shoppers say they somewhat or completely trust AI while shopping; nearly as many, 30%, openly distrust it, with the rest undecided. Trust is split and shallow, not assumed.

Bizrate Insights, nationally representative survey of 1,006 US shoppers, 2025. Measures trust in AI shopping assistance broadly, not AI imagery specifically.

~4×

Consumers are roughly four times more likely to say that *visible* AI in a brand's marketing costs that brand their trust than to say it builds it.

eMarketer, reporting a Klaviyo and Datalily survey of about 8,000 consumers, 2025 (approximately 31% "trust less" versus 7% "trust more").

The lesson is not that customers can always tell AI from real; often they cannot. It is that the instant a detail looks wrong, a button, a texture, a colour, a proportion, confidence drops and does not easily come back. An idealised image that does not match what arrives is a return waiting to happen and a brand promise quietly broken. When the underlying tool drifts by construction, every generated asset is a small bet that the drift stayed invisible this time.

What we deliberately left out. Several figures that circulate widely, that a specific percentage of shoppers read inconsistent AI imagery as a sign of an unreliable seller, that a specific share of self-publishing authors rank character consistency their top problem, that inconsistent characters cause a specific audience drop-off, and the claim that a named consultancy calls visual consistency the number-one barrier to scaling generative AI, could not be traced to a credible primary source. They may well be directionally true. We do not print numbers we cannot stand behind.

07 What a real solution has to do

If drift is structural, the answer is not a better prompt. It is a system engineered so that consistency is the default rather than the exception. That system has to clear four bars at once. Three out of four is not enough, because professional work needs all of them together.

1 **Hold the subject accurate and identical**

The product, character, or brand look stays the same, and true to a real reference, across an entire set. Not roughly similar. The same thing, with its real proportions, details, and colour intact.

2 **Change the world on purpose**

Background, scene, lighting, and colour move deliberately and cleanly when you want them to, as an intended edit, not as drift. Changing one thing must not disturb everything else.

3 **Work for images and for video**

The same subject holds not only across a set of stills but across frames and across shots, so a clip is something you can put in front of a customer rather than a glitch reel.

4 **Do it at production scale**

Reliable on the hundredth asset and the ten-thousandth, not just in a lucky one-off. Consistency that does not scale is a demo, not a pipeline.

Notice what these requirements are not. They are not "make prettier images." The tools already make pretty images. The unmet need is dependable sameness with deliberate, controlled change, and that is a different capability from raw generation.

08 How Depix closes the gap

Depix built its tools around the requirement general models treat as an afterthought. **Depix Design Lab**, the concept-phase tool made for industrial and product designers, and the **Depix MCP**, the engine that drives generation, are designed so that the subject stays accurate and consistent while the world around it changes on purpose, for images and for video.

THE PRINCIPLE

The subject holds. The world moves.

You change the scene, the lighting, and the colour. The product stays the product, the character stays the character, the design stays the design. A specific item and its exact colours stay true and repeatable shot after shot while the background, the setting, and the finish change around them, deliberately and cleanly, across stills and across video. That is the difference between a picture that is merely impressive and one you can put on a listing, in a catalogue, in a campaign, and in a film without anyone ever noticing it shifted.

What that unlocks maps directly onto the places drift breaks things:

- **Commerce:** catalogue imagery that matches the real SKU, at catalogue scale, so the picture keeps the promise the product has to keep.
- **Brand:** a whole campaign that reads as one intentional voice instead of fifty strangers, generated at the volume real campaigns need.
- **Design:** concept work you can genuinely iterate, the same form across angles, materials, and environments, so a team can compare like with like and actually decide.
- **Video:** a subject that stays itself across frames and shots while the light and colour change on purpose, so motion becomes something you can ship.

We are not promising prettier pictures, and we are not claiming the rest of the field can do nothing. We are claiming something more specific and more useful: the hard, unreliable, does-not-scale-by-default problem of keeping a subject accurate and consistent while you deliberately change everything around it is the problem Depix is purpose-built to make reliable, repeatable, and true, for the people whose work depends on it.

> Novelty is the demo. Consistency is the product, and consistency is what we sell.

09

Conclusion

The generative tools of this moment were built to surprise, and they are extraordinary at it. Most professional work needs the opposite quality: the same subject, dependable, again and again, while only the world around it changes. A system optimised for novelty is, by construction, the wrong instrument for controlled repetition, which is why consistency and its failure, drift, show up across every serious pipeline and across the field's own peer-reviewed benchmarks as an open problem rather than a solved one.

That is not a reason to retreat from generative AI. It is a reason to be precise about what the next step requires. The winners in professional imaging will not be whoever makes the single most beautiful frame. They will be whoever can make the same true thing, the tenth time and the ten-thousandth time it is needed, for stills and for motion, at scale. That is the gap Depix was built to close, and it is why any of this matters commercially.

About Depix. Depix builds design-intelligence tools for people who make real things, from industrial and automotive designers to brand and commerce teams. Its products, Depix Design Lab and the Depix MCP, are built to keep a subject accurate and consistent while its scene, lighting, and colour change on purpose, across images and video. Learn more at depix.ai.

References

Primary and first-party sources. Peer-reviewed venues are named; two recent items are preprints and are marked as such. Consumer figures are attributed to the named survey and, where relevant, to the outlet that reported them.

1. Huang, Z. et al. **VBench: Comprehensive Benchmark Suite for Video Generative Models.** CVPR 2024. arxiv.org/abs/2311.17982
2. Peng, Y. et al. **DreamBench++: A Human-Aligned Benchmark for Personalized Image Generation.** ICLR 2025. arxiv.org/abs/2406.16855
3. **VBench-2.0: Advancing Video Generation Benchmark Suite for Intrinsic Faithfulness.** 2025. arxiv.org/abs/2503.21755
4. **A Survey: Spatiotemporal Consistency in Video Generation.** ACM Computing Surveys, 2025 (DOI 10.1145/3802588). arxiv.org/abs/2502.17863
5. **Survey of Video Diffusion Models.** Transactions on Machine Learning Research (TMLR), 2025. arxiv.org/abs/2504.16081
6. **Video Diffusion Generation: Comprehensive Review and Open Problems.** Artificial Intelligence Review 58, Springer, 2025 (DOI 10.1007/s10462-025-11331-6). link.springer.com
7. **CoDi: Subject-Consistent and Pose-Diverse Text-to-Image Generation.** ICLR 2026. arxiv.org/abs/2507.08396
8. Butt, M. A. et al. **ColorPeel: Color Prompt Learning with Diffusion Models via Color and Shape Disentanglement.** ECCV 2024 (Springer LNCS 15065, DOI 10.1007/978-3-031-72667-5_26). arxiv.org/abs/2407.07197
9. **Color Alignment in Diffusion.** CVPR 2025. arxiv.org/abs/2503.06746
10. Wu, Q. et al. **Uncovering the Disentanglement Capability in Text-to-Image Diffusion Models.** CVPR 2023. arxiv.org/abs/2212.08698
11. **BAgger: Backwards Aggregation for Mitigating Drift in Autoregressive Video Diffusion Models.** Preprint, December 2025. arxiv.org/abs/2512.12080
12. **Towards Error-Free Long Video Generation.** Preprint, 2026. arxiv.org/abs/2606.22370
13. Google Developers Blog. **Introducing Gemini 2.5 Flash Image.** 26 August 2025. developers.googleblog.com
14. Bizrate Insights. **How Shoppers Navigate AI and Authenticity: State of Consumer Trust in Online Shopping.** 2025. Nationally representative, n = 1,006 US shoppers. bizrateinsights.com
15. eMarketer. **Visible AI marketing is far more likely to cost brands trust than build it.** 2025. Reporting a Klaviyo and Datalily survey of ~8,000 consumers. emarketer.com