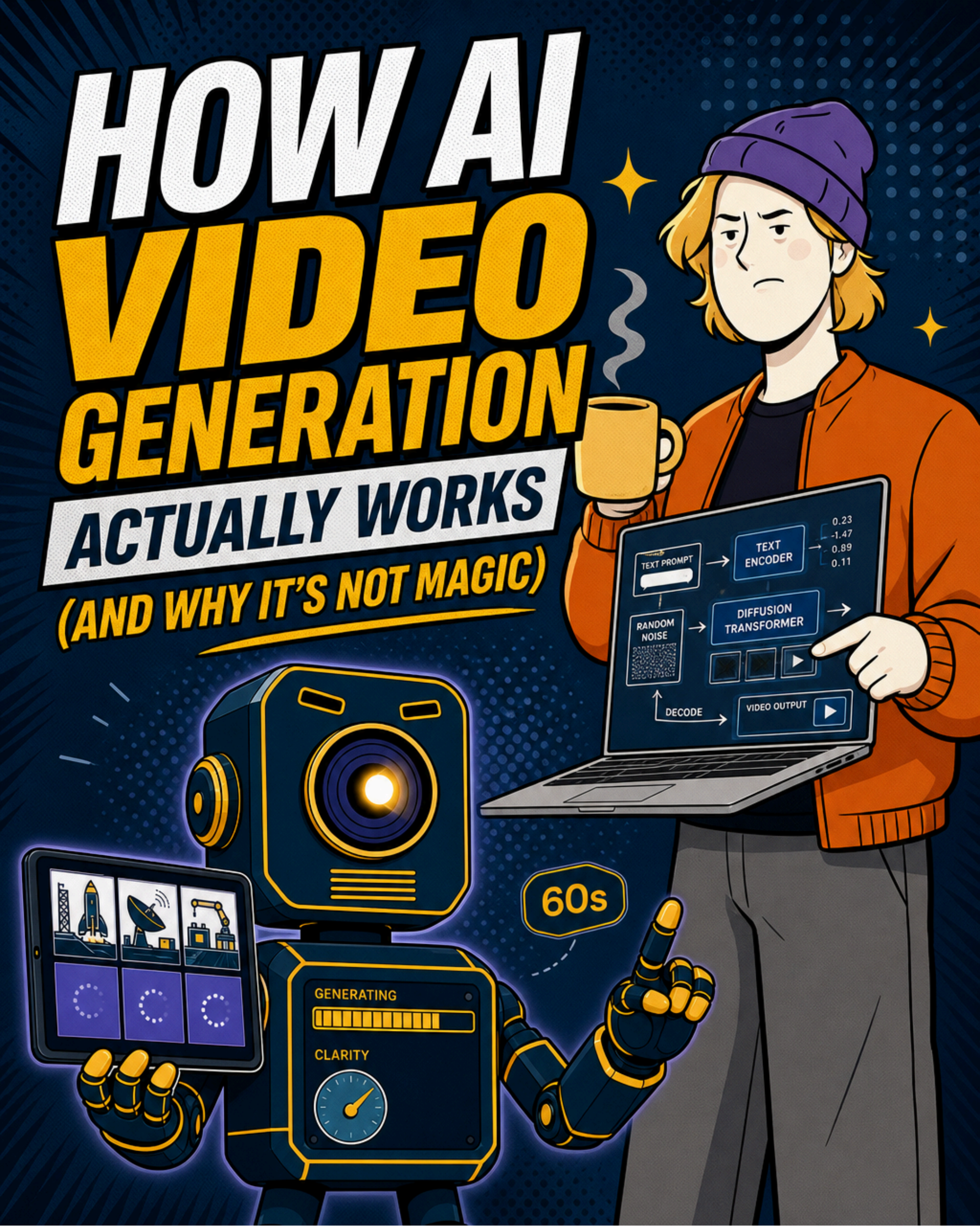
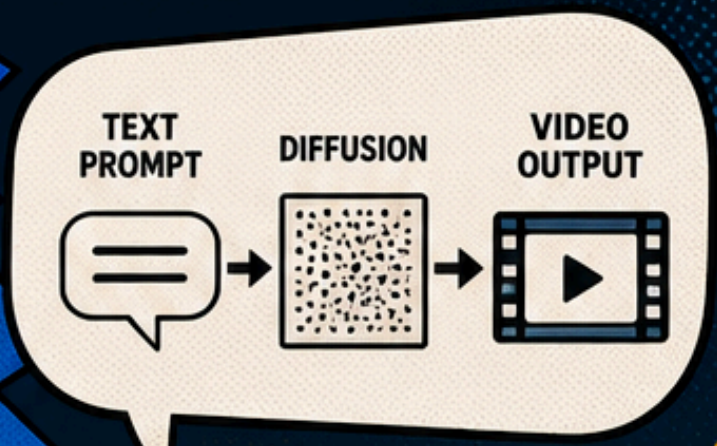
HOW AI VIDEO GENERATION ACTUALLY WORKS

(AND WHY IT'S NOT MAGIC)



HOW AI VIDEO GENERATION ACTUALLY WORKS (AND WHY IT'S NOT MAGIC)



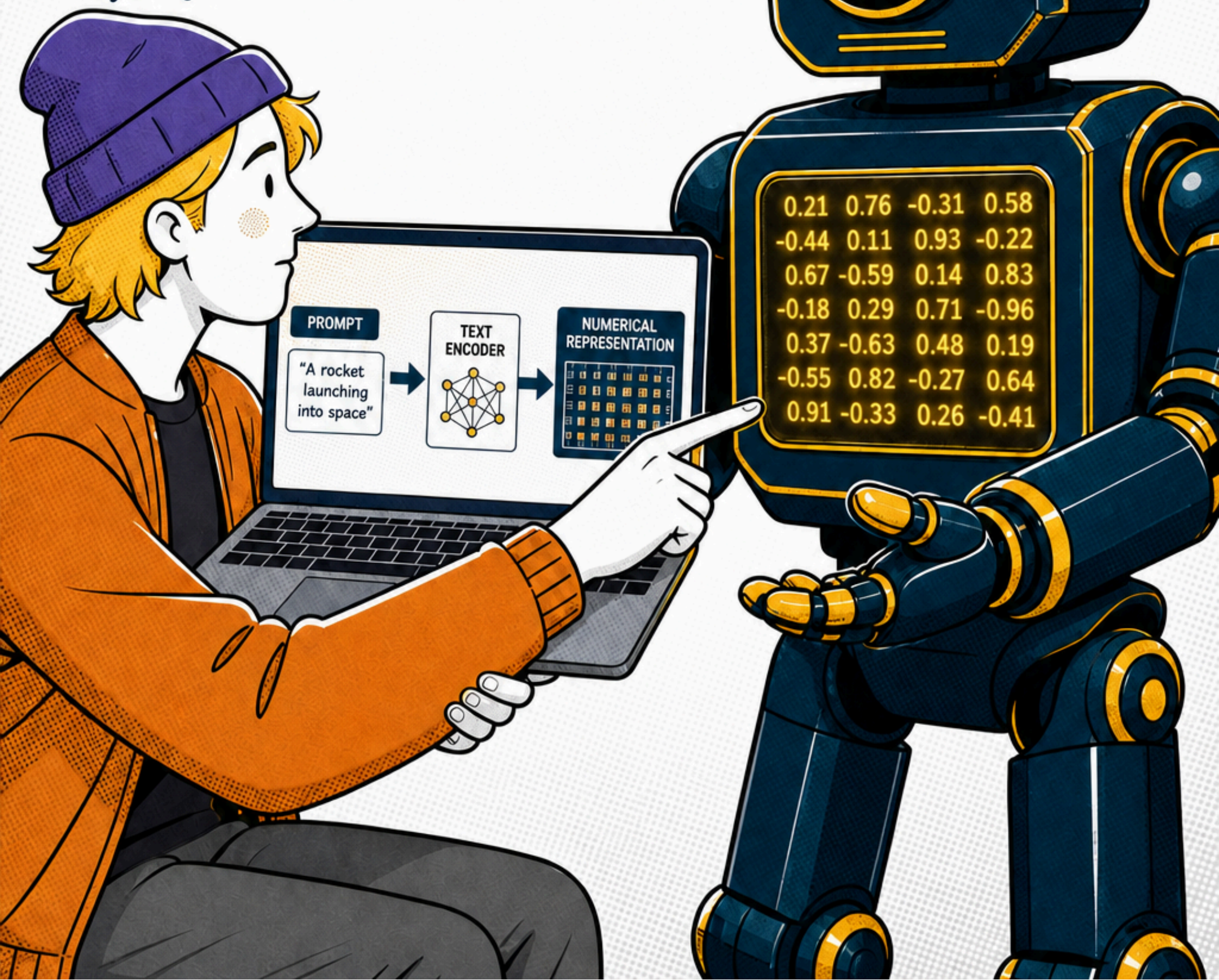
Every week, another CMO watches a Sora demo reel and wonders why their own attempt produced a flickering mess with a character whose hands slowly merged into the background.
The output looked great in someone else's LinkedIn post. Yours looked like a fever dream.

There's a reason for that. And understanding it will save you a lot of wasted afternoons and some very awkward conversations with your sales team.

Here's a plain-language explanation of how AI video generation works, what it's genuinely good at, and where it still falls over.
No PhD required.

YOUR PROMPT DOESN'T TOUCH THE VIDEO DIRECTLY

When you type into Sora, Veo, or Kling, your words don't flow straight into a video renderer. A text encoder first converts your prompt into a structured numerical representation — a compressed map of what you described — before anything visual happens.



FROM RANDOM NOISE TO COHERENT VIDEO

The most common architecture is a **diffusion transformer**.

The model starts with frames of pure random noise and refines them step by step into something coherent — **working backwards from static into signal**,

guided by your **encoded prompt** at every step.

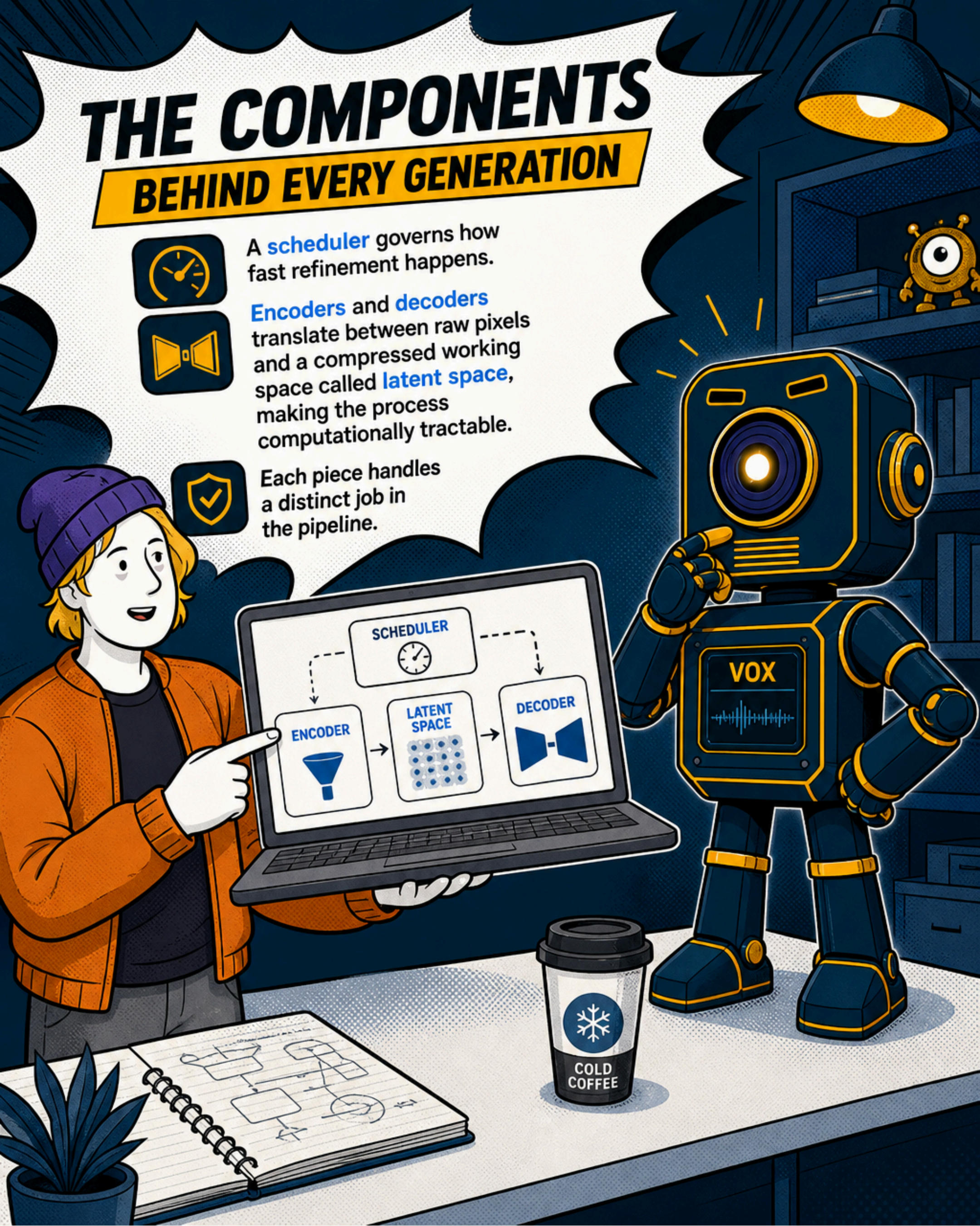
!

THE COMPONENTS
BEHIND EVERY GENERATION

A **scheduler** governs how fast refinement happens.

Encoders and **decoders** translate between raw pixels and a compressed working space called **latent space**, making the process computationally tractable.

Each piece handles a distinct job in the pipeline.

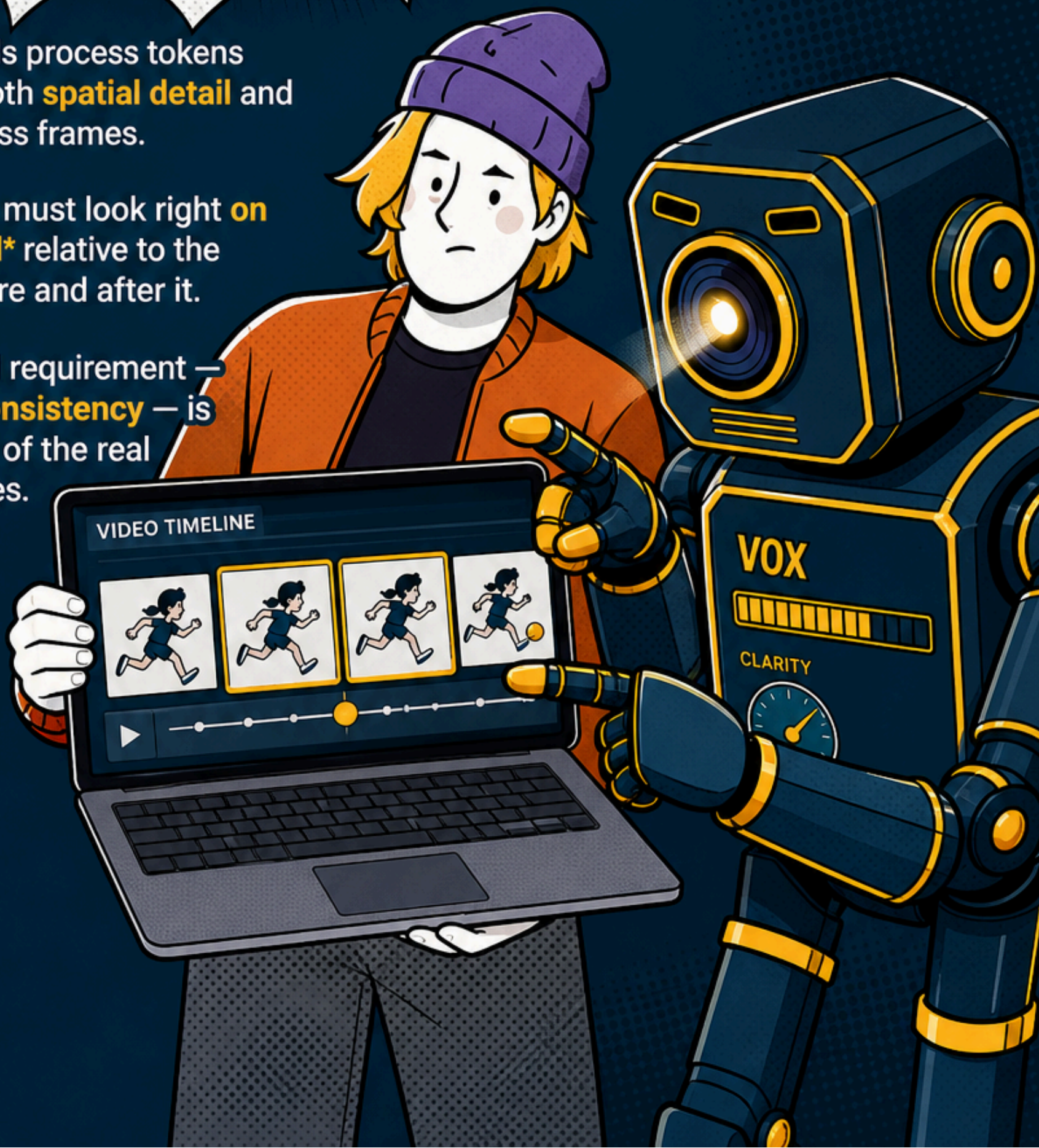


WHY VIDEO IS HARDER THAN IMAGES

Video models process tokens capturing both **spatial detail** and **motion** across frames.

Every frame must look right **on its own *and*** relative to the frames before and after it.

That second requirement — **temporal consistency** — is where most of the real difficulty lives.



WHAT IS TEMPORAL CONSISTENCY?

Temporal consistency means objects, characters, lighting, and textures stay **stable** frame-to-frame.

Without it, video looks glitchy:

- characters morph
- backgrounds flicker
- lighting shifts

for no reason the scene explains.

It is the most common complaint from anyone producing AI video seriously.

WHY FLICKERING HAPPENS: THE FRAME-LEVEL TRAP

A SINGLE FRAME CAN LOOK PERFECT. THE PROBLEM LIVES IN THE **RELATIONSHIP BETWEEN FRAMES**, NOT WITHIN ANY INDIVIDUAL ONE.

OLDER DIFFUSION MODELS PROCESSED FRAMES **NEAR-INDEPENDENTLY**—EACH WAS ITS OWN GENERATION EVENT, ONLY LOOSELY TIED TO ITS NEIGHBOURS—MAKING CROSS-FRAME COHERENCE AN AFTERTHOUGHT.

VOX

THE PROBLEM LIVES IN THE **GAP** BETWEEN FRAMES.

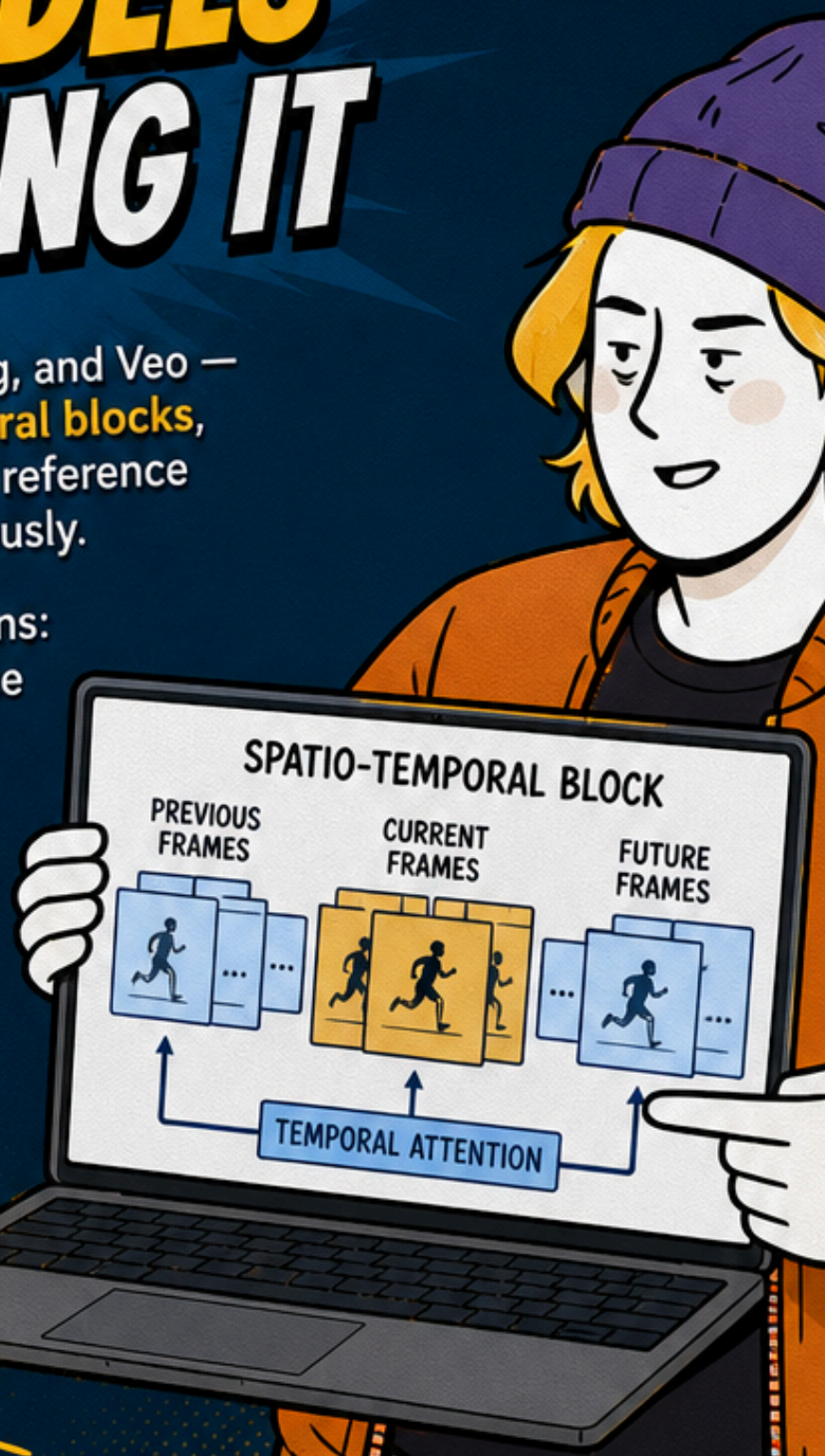
AUDRIUS

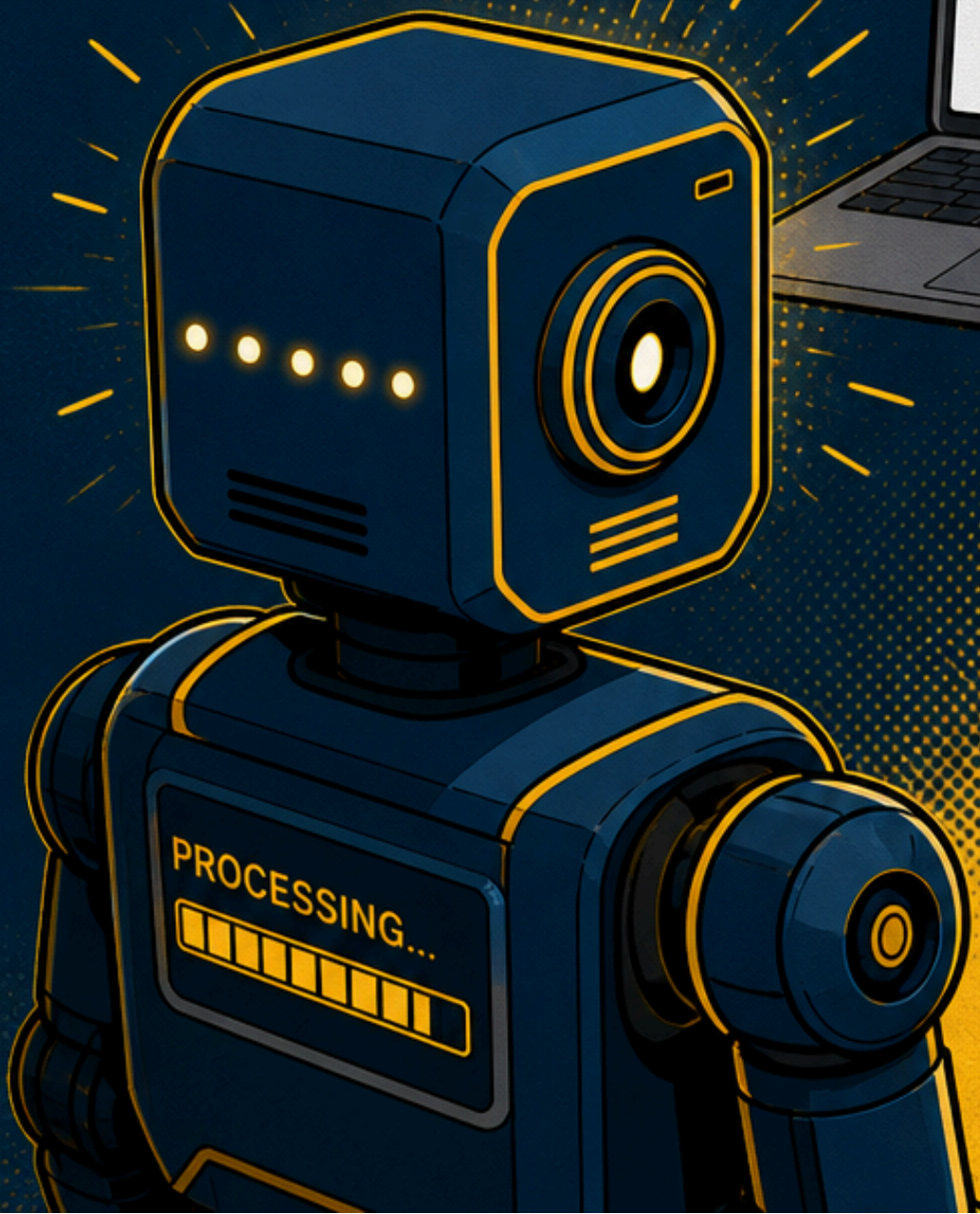
HOW NEWER MODELS ARE TACKLING IT

Leading models — including Sora, Kling, and Veo — now treat entire clips as **spatio-temporal blocks**, adding **temporal attention layers** that reference previous and future frames simultaneously.

Progress is real, but a **trade-off** remains: stronger consistency constraints reduce fine detail, while richer textures increase drift.

NO FREE LUNCH.

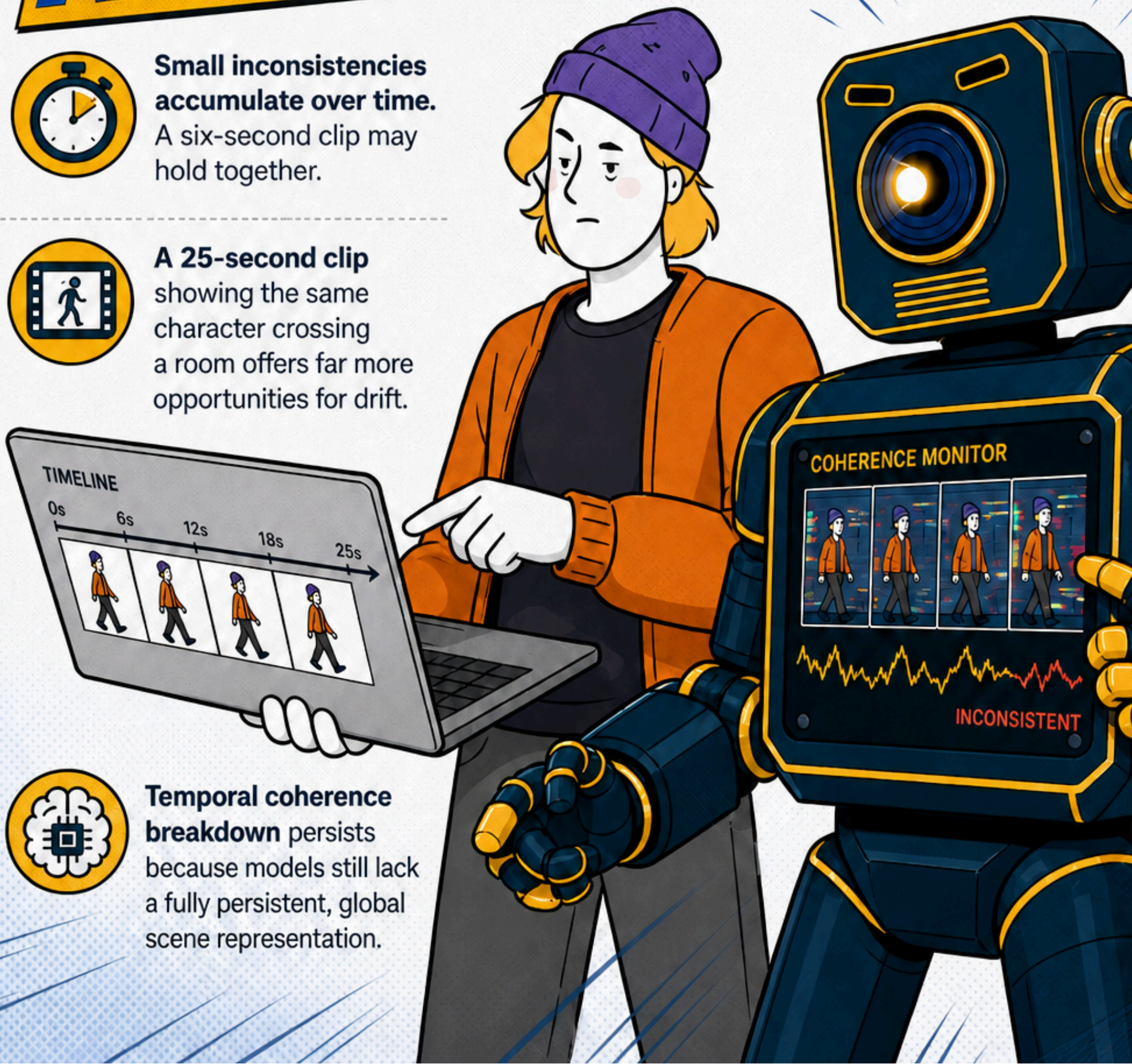




WHY LONGER CLIPS ARE HARDER

Small inconsistencies accumulate over time. A six-second clip may hold together.

A 25-second clip showing the same character crossing a room offers far more opportunities for drift.



Temporal coherence breakdown persists because models still lack a fully persistent, global scene representation.

WHY SEPARATE MODELS MEAN COMPOUNDING ERRORS

Traditional AI video pipelines stitch together separate specialist models: one for visuals, one for audio, one for lip-sync.

Each handoff introduces its own timing errors.

VISUAL MODEL
AUDIO MODEL
LIP-SYNC MODEL

VOX
VISUAL
AUDIO
LIP-SYNC

THOSE ERRORS MULTIPLY.

Even a few frames of lip-sync drift and viewers notice immediately, even if they can't name why.

NATIVE AUDIO-VISUAL CO-GENERATION:

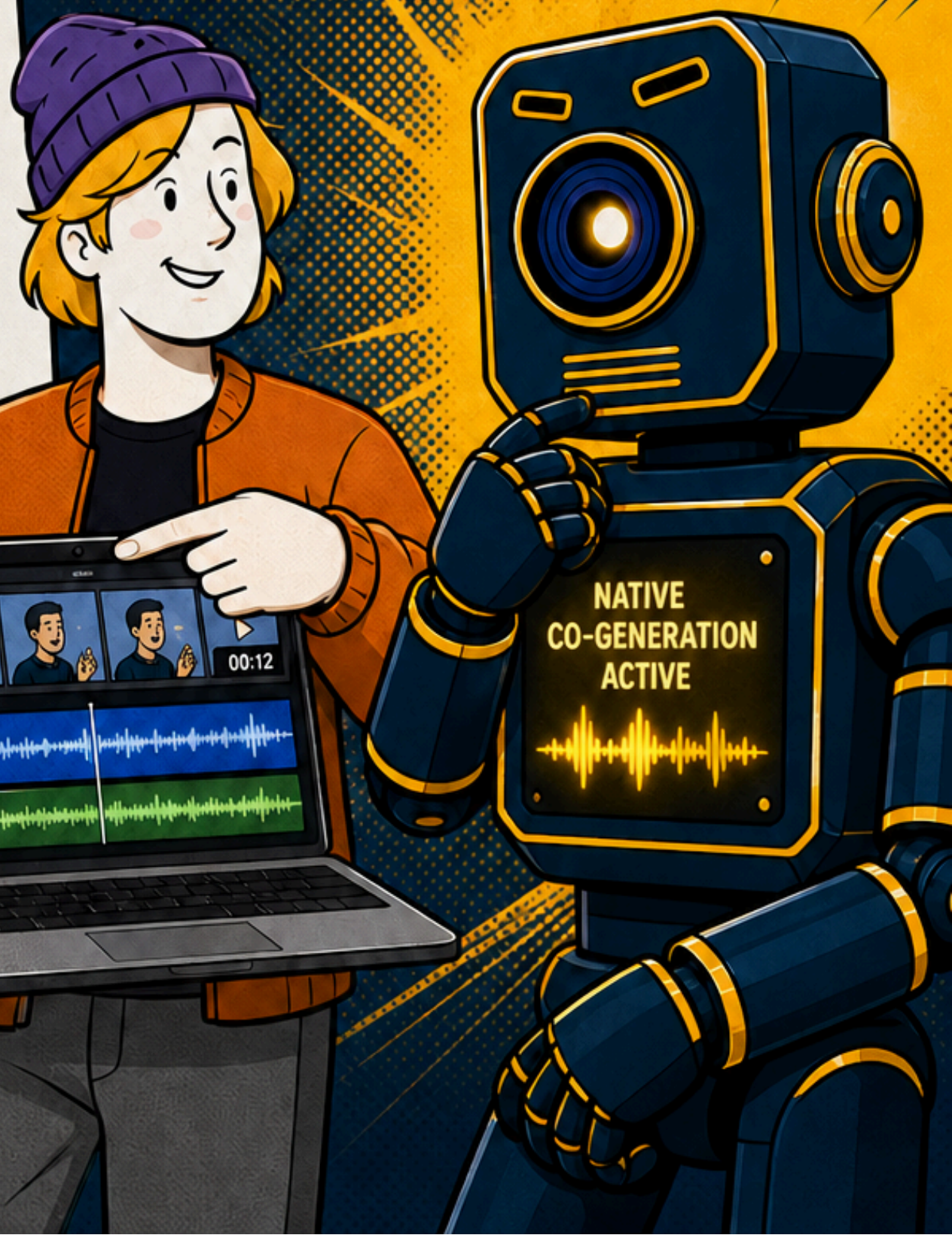
THE 2025–2026 SHIFT

The trend is toward generating **video and synchronised audio** jointly in one pass.

Veo 3 is a confirmed example, producing dialogue, sound effects, and ambient noise alongside visuals simultaneously.

This **reduces compounding errors** substantially — but parity across platforms varies.

Always verify for your specific tool.



LIP-SYNC FOR B2B DIALOGUE IS STILL GENUINELY HARD

Making a real person convincingly mouth technically precise words remains **difficult** even with one-pass architectures.

You still get better results by structuring dialogue into ***short, controlled clips*** and aligning manually in the edit.

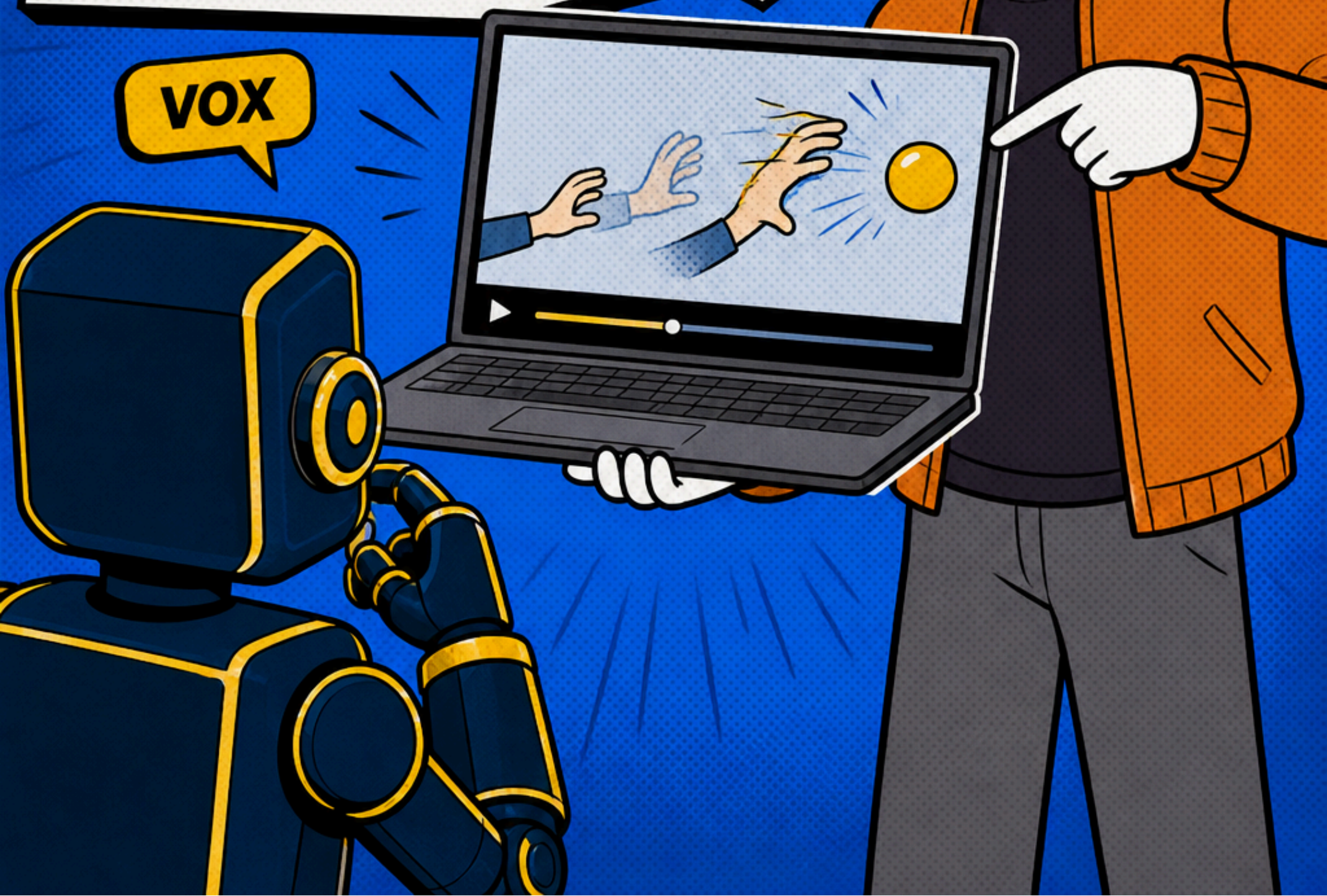
One-pass helps — it doesn't fully solve it.

PHYSICAL MOTION: WHERE CURRENT MODELS STILL FAIL

Current models learn statistical correlations from training data rather than explicit physics. Bodies, fabric, and fluids move inconsistently.

A hand reaching for an object may warp mid-motion.

At normal playback speed, something feels wrong — even if the viewer can't articulate what.



AI VIDEO IS A STATISTICAL AVERAGE OF THE WEB

AI VIDEO MODELS TRAIN ON BILLIONS OF TEXT-VIDEO PAIRS SCRAPED FROM THE INTERNET.

THE OUTPUT REFLECTS WHAT STATISTICALLY APPEARS MOST OFTEN ACROSS THAT CORPUS — NOT WHAT YOU NEED.

THE MODEL PRODUCES THE MOST PROBABLE PIXEL SEQUENCE FOR YOUR PROMPT, NOT THE MOST ACCURATE ONE.

GENERATING

CLARITY

PROBABILITY DISTRIBUTION

PROBABILITY

MOST PROBABLE (AVERAGE)

POSSIBLE OUTPUTS

AVERAGED PIXEL OUTPUTS

GENERIC PROMPTS PRODUCE GENERIC RESULTS

BIOTECH INNOVATION

BIOTECH INNOVATION

VOX
NO FLUFF.
JUST SIGNAL.
ALWAYS.

THE MODEL HAS NEVER SEEN YOUR PRODUCT

For a product demo showing how specific regtech software works, or visualising a novel hardware component, this is a fundamental problem.

The model has seen things that vaguely resemble your product. That is not the same as knowing your product.

Vague resemblance cannot carry accurate technical communication.

A WRONG METAPHOR COSTS YOU CREDIBILITY!

Mullti-shot narrative:
the memory problem

THE LEAST DISCUSSED LIMITATION AND PROBABLY THE MOST IMPORTANT ONE FOR ANYONE TRYING TO MAKE A COHERENT 60-SECOND EXPLAINER.

MISMATCHED CHARACTER FRAMES

00:00–10s | 10–20s | 20–40s | 40–60s

VOX

INCONSISTENT CHARACTER OUTPUT

EACH GENERATION CALL HAS NO MEMORY OF WHAT WAS GENERATED IN THE PREVIOUS CALL.

AUDRIUS

VOX

WHAT THIS MEANS IN PRACTICE

SHORT CLIPS ARE MANAGEABLE. LONG NARRATIVES ARE NOT, YET.

Generation lengths vary by platform and are expanding, but shorter clips tend to maintain better consistency.

GENERIC SCENES WORK WELL. TECHNICALLY SPECIFIC SCENES DON'T.

If your product involves a novel process, hardware, or interface that doesn't exist in the training data, the model will approximate it.

PROMPT QUALITY DETERMINES OUTPUT QUALITY, HEAVILY.

The model amplifies whatever you give it. A vague brief produces vague output. Knowing what story to tell before you touch the tool is the entire strategic layer that AI cannot provide for you.

MANUAL PRODUCTION STILL FILLS THE GAPS.

Precise UI mockups, brand-locked visuals, character continuity, exact text: these still require manual motion graphics on top of AI output.

If you've handed a team member a Sora subscription and called it your explainer video strategy, you now know why the output feels off. The tool is real. The gap between the tool and a finished, credible, technically accurate video is also real, and it's filled by narrative direction, prompt engineering, and production craft, not by prompting harder.

If you're a CMO at an ANZ tech company trying to work out what a production-ready explainer video workflow actually looks like, we've written the process out in detail at Infrairis.

We use the same tools, we've stress-tested them on our own products, and we can tell you exactly where the gaps are and how we bridge them.

startups.infrairis.com

infrairis

https://startups.infrairis.com/