Why Concept Entanglement Means You Can’t Have AI Video ‘Your Way’

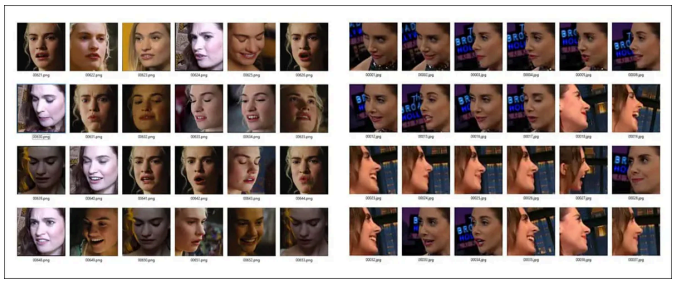
Originally published at https://www.unite.ai/why-concept-entanglement-means-you-cant-have-ai-video-your-way/

AI video tools promise total control, but hidden ‘concept entanglement’ glues identities, expressions and behaviors together, forcing hacks and template tricks that shatter the myth of effortless GenAI magic.

Opinion Since I last got into the subject at length five years ago, the problem of *concept entanglement* in trained AI systems has extended to a far wider range of users, without really being understood any better on its own terms.

Back then, autoencoder deepfake systems (i.e., the now defunct DeepFaceLab and the less porn-centric FaceSwap, both derived from the disgraced and almost immediately-banned 2017 Reddit code release) were the only game in town for creating relatively photorealistic deepfakes of people.

These systems relied on extensive facial training datasets that were intended to provide the AI model with information about A) what the person looked like in repose (a canonical reference embedding) and B) what they looked like under the diverse situations that a face can reflect, from *sleep* through to *laughter*, *horror*, *boredom*, *cynicism*, *sadness*, etc.



Identity comes not singly, but together with facial expressions. Additionally, certain emotions may only have available face data from particular, extreme angles, which will tend to associate the angle with emotion and vice versa.

The trouble was that canonical identity usually had to be inferred from facial captures that were not in themselves ‘neutral’, so that the preponderance of smiles and grins obtained when scraping stock datasets would shift the distribution towards a ‘smiling default’. This was because of the high volume of red-carpet paparazzi shots in the web-scraped training data that typically informs these models, as well as any other equally specious reason why a dataset might be biased towards one kind of image.

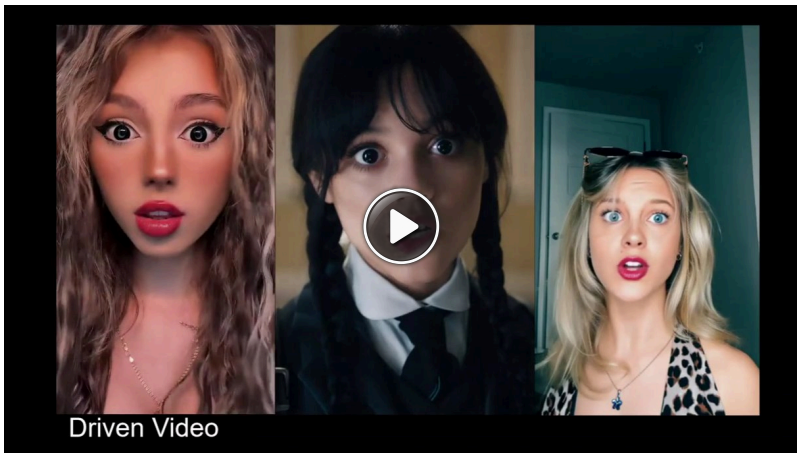
In other words, the autoencoder system would have to attempt to extract a 'neutral' identity concept from thousands of images where the facial features were contorted by *normal facial expressions*.

It also had to try and disentangle semantic facial concepts of different emotions *from the angles that the faces were taken at*. This meant that if the only 'terrified' facial expressions available were taken from a profile view, the trained system would only be able to reproduce that emotion optimally from that view.

Facing Forward

As diffusion-based approaches took over the gen AI image (and later, video) scene from 2022 on, generative systems became far better at extrapolating accurate facial expressions when supplied with limited face data.

Even the severely thorny challenge of creating convincing profile views has been nigh-on overcome, at the current state-of-the-art, while expression data has been quite effectively slipstreamed away from identity – to the extent that the kind of live deepfake puppeteering pioneered by the autoencoder-driven DeepFaceLive streaming system has many effective offline diffusion applications, with real-time enactment a likely future development:



CLICK TO PLAY VIDEO ON SITE

Click to play. *From the 'FlashPortrait' project, diverse examples of driving avatars through source videos. In this case, it does not matter what side the 'realistic' domain sits on, if any.* Source

Yet, as the canvas of genAI has widened and the output has become more sophisticated, the problem of entanglement has simply spread to multiple other areas – and is currently being 'fixed' by some pretty cheap and pretty old tricks. If you don't know what those tricks are, you might have a more positive take on how quickly video and image AI is evolving and overcoming its old bugbears.

Chatty Cats

Hopefully it's clear why identity and emotion proved hard to separate for those old 2017-era autoencoder systems. It was because a) There was too much data of one kind, OR too specific a version of one type of important data, either of which will cause a distributional bias; and/or B) the model architecture wasn't up to the task of separating out these qualities, and tended to 'glue them together' at inference time, unless the user took an extraordinary amount of care to ensure balance in their dataset.

For exactly the same reason, similar problems have emerged in a number of open source and proprietary video models over the last few years, though they have been overshadowed by greater levels of criticism around hallucination, lack of censorship, and diverse other topics.

For instance, in the Wan2.+ system, many users have found that it is very difficult to stop their generated characters from *talking incessantly*, and often also difficult to stop them looking at the camera.

The latter issue (looking to camera, or breaking the fourth wall) predates the advent of video synthesis systems, since it emerged in various image-only diffusion systems, due to the prevalence of 'looking to camera' photos in web-scraped datasets such as LAION.

The issue around 'garrulous' characters comes from the easy abundance of 'influencer' videos on YouTube, which naturally offer thousands of hours of straight-to-lens discourse, often curated into datasets where research scientists can launder the web-scrape by providing an academic context.

But unless the original or subsequent curators take care to limit the number of videos of this type, and balance them against more different types of footage, a severe bias develops in the video model, which will need addressing through prompt-based remedies and diverse third-party adjunct systems.

Faced with Wan's 'chattiness' issue, Reddit user u/Several-Estimate-681 came up with a workaround that leverages a setting in the Wan 2.1 Infinite Talk V2V system – a framework designed to *encourage* influencer-style loquaciousness – that allows the user to silence the rendered character:

CLICK TO PLAY VIDEO ON SITE

Click to play: *Just* listen – *a workaround to achieve character attentiveness in Wan2.+.* Source

Clearly, shortcuts of this kind do not represent low-level architectural solutions, and, absent true solutions being found and implemented by the creators of foundation models (because casual hobbyists don't usually have millions of dollars to recreate or fine-tune such work), this means that the game of entanglement 'whack a mole' is likely to be reset to zero at the next version release.

Cheap and Fragile

There is nothing in diffusion architecture itself that makes these problems inevitable; indeed, if there were some way to apply really effective curation, triage and high-quality captioning and annotation to hyperscale datasets with data points numbering in the millions, nearly all of these problems would likely disappear.

However, that level of attention to detail would be akin to the Manhattan Project in terms of logistics, scope, needed resources, and sheer long-term effort. In a climate where a new architecture, or even a new architecture *version* could undo the sheer extent of such an effort, there is no current will to make this kind of commitment.

Consequently, as far as is concordant with obtaining usable models, the cheapest approaches remain preferred. One such example of 'stinginess' is data augmentation, which, when applied illiberally and to the wrong kinds of dataset videoclips, can have hilarious results:

CLICK TO PLAY VIDEO ON SITE

Because data augmentation often reverses the direction of source videos in the dataset, the AI model can occasionally learn some 'impossible' moves. – Source

However, in the aggregate, rocks rolling uphill and people breaking character by turning on ‘influencer mode’ tend to be considered instances of collateral damage in generative systems that can, in spite of such persistent goofs and Achilles heels, be coaxed into producing impressive results and sufficiently awed headlines.

Boilerplate Solutions

In the current period, hundreds of generative video domains, almost all of which in some way break the new slew of laws and pushback against GenAI, are enjoying their time at the trough before law enforcement, blocklists or other kinds of deplatforming remove these commercial services.

The larger and better-known sites of this nature, such as Kling and Grok, tend to either adhere to some form of self-censorship (eventually), or to respond to criticism by changing the kinds of content their platforms facilitate for users.

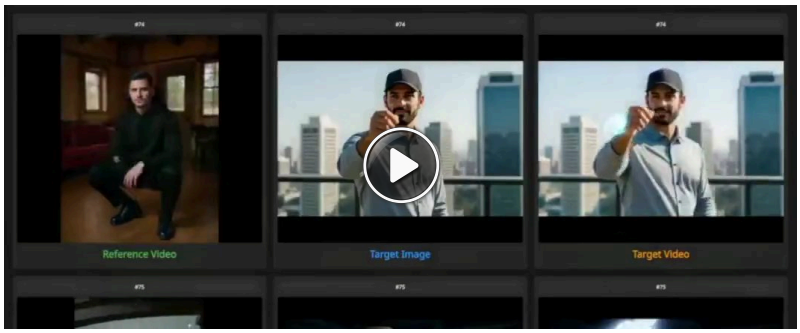
But behind those big names are hundreds of other fly-by-night operations, constantly catering to demand for new (and often more extreme) kinds of content.

This kind of low-effort provisioning precludes the extremely high cost and effort of training foundation models from scratch. Very often, even fine-tuning, which costs considerably less, is precluded.

Therefore these sites offer ‘templates’, which behave 100% identically in practice to custom-trained LoRAs, which have been used by AI hobbyists for more than four years now, to train any desired identity, style, object and (in the case of video LoRAs) motion or action into a dedicated LoRA adjunct.

With the LoRA interposed between the user and the foundation model, the results obtained will be very specific to what the LoRA was trained on, and, usually, the wider performance of the model is undermined by the weight-bending influence of the LoRA, which will reproduce its own subject very well, but would also interpose that material into any request whatsoever (if the fly-by-night GenAI video sites allowed this level of control – they don’t; they just offer an *[ACTION OF YOUR CHOICE]* template, and interpret your input text/images/videos in the way most likely to result in a successful application of the template).

For obvious reasons, I can’t embed website samples in this article; but the research literature has recently offered some analogous examples. Here, for instance, the EffectMaker project shows the principle in action, whereby a specific action is applied to a user-supplied image:



CLICK TO PLAY VIDEO ON SITE

Click to play. *In EffectMaker, fine-tuned specific effects can be applied to custom input.* Source

Even in these highly-curated and targeted circumstances, users often complain that multiple, token-burning attempts need to be made in order to obtain a good result, and we should not perhaps ascribe to provider avarice or sharp practices what is more likely the fault of congenitally ‘hit-and-miss’ DiT GenAI frameworks.

The wider public, it could be argued, gets its impression of the capabilities of GenAI from cherry-picked examples that are not representative of what a casual, neophyte user would be likely to obtain. If a user burns through six attempts at a template (i.e., a LoRA supplied by the AI website), they’ll tend to publish and laud the best of these, conveying the impression that one could obtain such results by querying the base model – and conveying the impression that generative foundation models are far more disentangled than they really are.

Conclusion

The literature continues to examine the problem of entanglement, which first hove seriously into view around 2020, in the Max Planck/Google collaboration *A Sober Look at the Unsupervised Learning of Disentangled Representations and their Evaluation*.

Additionally various successors to *Disentanglement via Contrast* (DisCo) periodically emerge, and the scene remains lively with an awareness of the problem that far exceeds public awareness of what AI *cannot* do, in this regards.

One Chinese study from 2024 suggests that a resolution to entanglement may not be necessary at all, in order to solve the problems that it brings. Historically, this rings true, since many intractable issues in computer vision were overcome not by being solved, but by being surpassed through entirely new techniques and approaches.

Until such a discrete contender emerges, it seems we will continue to need to apply hot-fixes and band-aids to GenAI's shortcomings and limitations, and endure public over-estimation of the flexibility and ductility of foundation models.

First published Monday, March 23, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More