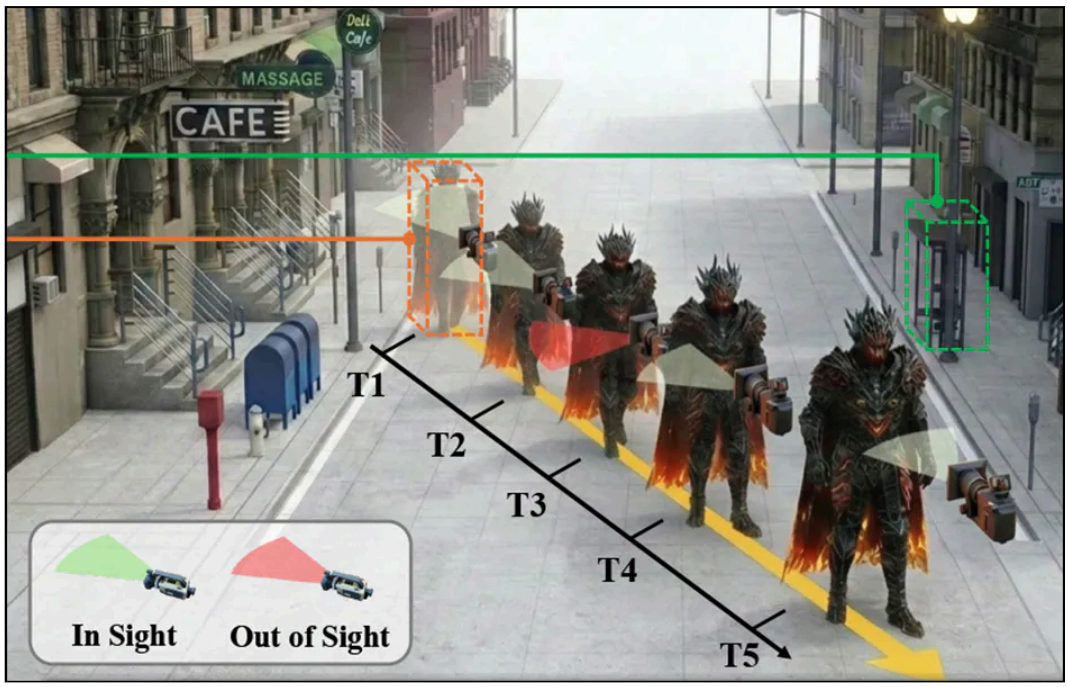
Out of Sight, Out of Mind: Tackling the Biggest Problem in AI Video

Originally published at https://www.unite.ai/out-of-sight-out-of-mind-tackling-the-biggest-problem-in-ai-video/

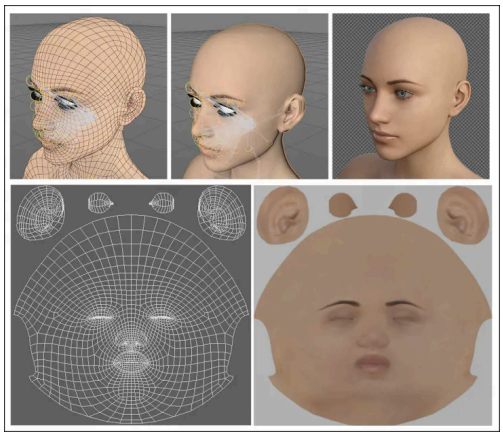


The biggest problem with even the best AI video generators is that they have chronic amnesia – a challenge that new research from China is now tackling.

The biggest problem with even the best and most state-of-the-art AI video generation systems is that they all have *chronic amnesia*: if the camera pans away from what it's focused on and then pans back, it will never find what was there at the start – characters will have disappeared, changed appearance and/or type of movement, and the background will likely have also changed.

This is because the diffusion-based generation system has a limited rolling window of attention, and because it is always dealing with *what it can see in that moment*; in a true enactment of solipsism, what is *outside* of the frame is non-existent for generative AI – it becomes literally dumped from memory.

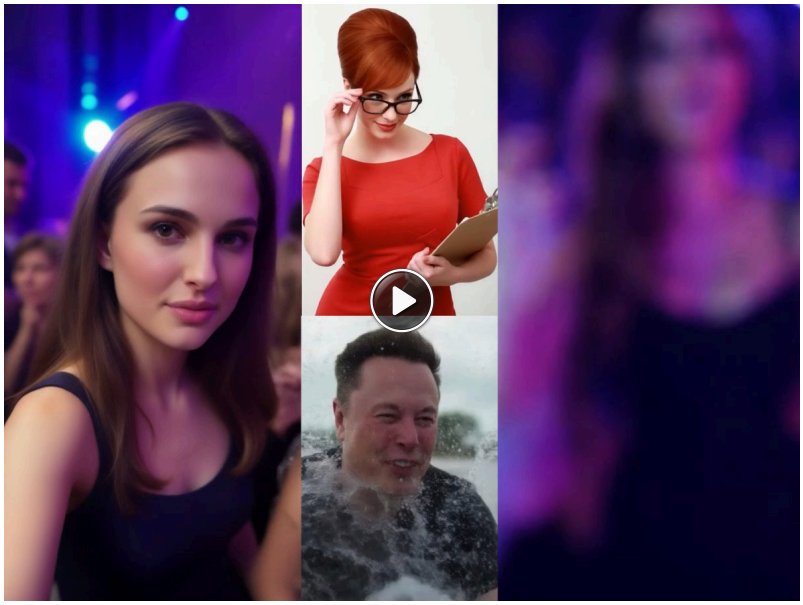
This has never been a problem in traditional CGI, which can always refer to and accurately recreate a subject, including appearance and motion, at any point in a rendered video where they may be needed again:



Traditional CGI meshes and bitmapped textures can always be drawn back into a render, providing consistent appearance – a trick that is much harder to achieve in AI approaches, because there is no equivalent 'flat reference' file, or collection of related files.

This is because CGI's component elements, such as the mesh and textures (see image above), as well as movement files and other dynamic behaviors, can live discretely on disk, and be drawn into a composition any time.

There is no such 'flat repository' in generative video AI; the nearest it can get to this functionality is LoRAs – specially-trained adjunct files which can be trained on consumer equipment, allowing novel characters and specific clothing to be 'forced' into the video:



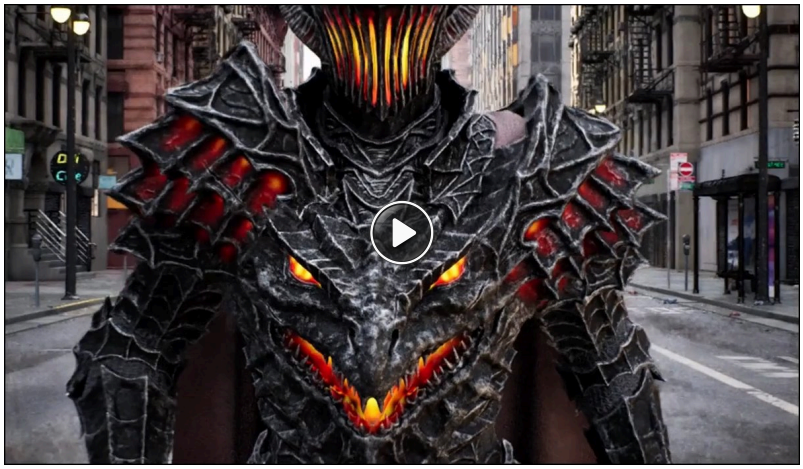
CLICK TO PLAY VIDEO ON SITE

Click to play. *AI video's solipsism problem can be mitigated to a certain extent by using LoRAs – but the results can be overwhelming.*

This is not an ideal solution, though. For one thing, LoRAs are tied to an exact specific version of a foundation model (such as Wan2+, or Hunyuan Video), and need recreating every time the base model changes. For another, LoRAs tend to distort the weights of the foundation model, so that the LoRA's trained identity gets imposed on all characters in a scene. Additionally, fine-tuning methods of this kind are very sensitive to poorly-curated datasets.

Accurate Encores

Now, a new academic/industry collaboration from China is offering the first significant remedy that has come to my attention in over three years of reporting on this issue. The method uses what the researchers dub *hybrid memory* to keep the off-screen character and their direct environs active and accurate in the latent space of the model, so that when our viewpoint returns to them, the effect is consistent:

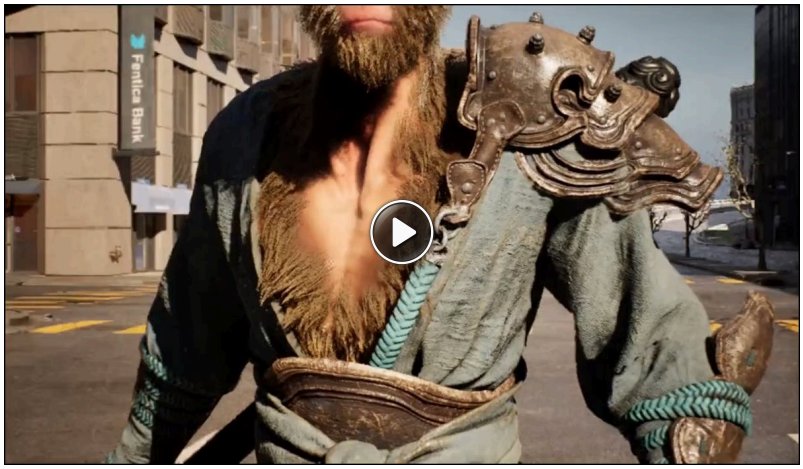


CLICK TO PLAY VIDEO ON SITE

Click to play. *From the project site for the new paper, two examples of AI-generated (WAN) characters exiting frame and accurately re-entering.* Source

It should be emphasized that this is not the same thing as achieving *character consistency* across different shots – something that was claimed to have been achieved a year ago in Runway's Gen 4 release, and which remains an ongoing pursuit in the research literature.

Rather, what's solved here is something that no commercial or experimental framework I've seen has been able to achieve – the *visually-consistent re-appearance* of an off-screen character's former look, motion and environment:



CLICK TO PLAY VIDEO ON SITE

Click to play. *The other two main examples given at the new initiative's project site.*

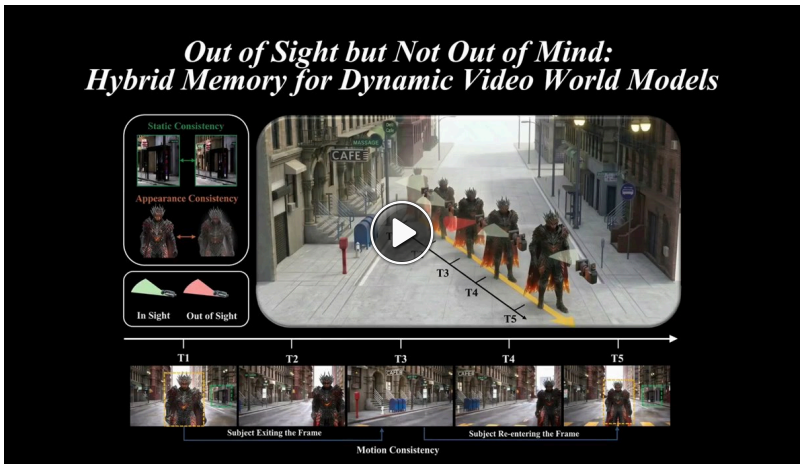
Obviously the principles at work here can be equally applied to other domains, such as urban exploration, POV driving, or other kinds of non-character renderings.

It should be emphasized also that this new approach does not solve or address the issue that Runway Gen4 and other closed-source platforms claim to have addressed, by recreating characters *across different shots*; instead it does what none of them have yet succeeded at– persisting a character and environment in memory, *without needing them to remain visible to the viewer at all times*.

The new work comprises a dedicated dataset generated through Unreal Engine, as well as custom metrics for the solipsism problem*, and a bespoke generative framework built over WAN. In tests against the few analogous systems available, the authors claim state-of-the-art results, and they comment:

> *'[Memory] mechanisms have emerged as a critical frontier in advancing world models, as memory capacity dictates the spatial and temporal consistency of generated content.*
>
> *'Specifically, it is the cognitive anchor that allows the model to retain historical context during viewpoint shifts or long-term extrapolation.*
>
> *'Without robust memory, a simulated world quickly unravels into disconnected, chaotic frames.'*

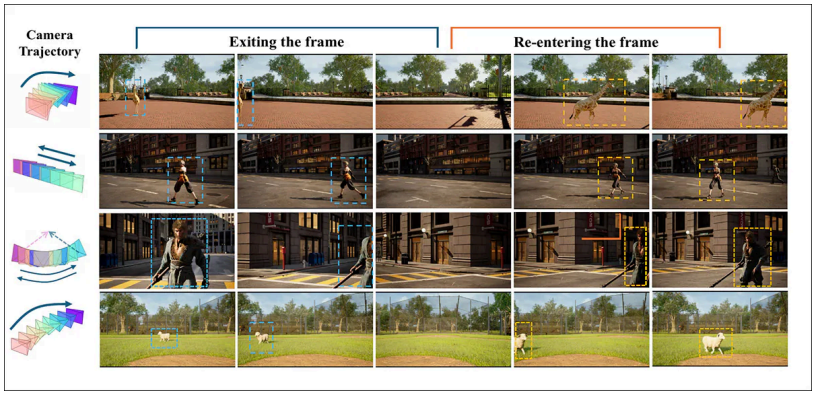
The new paper is titled *Out of Sight but Not Out of Mind: Hybrid Memory for Dynamic Video World Models*, and comes from seven researchers across Huazhong University of Science and Technology, and the Kling Team at Kuaishou Technology.



CLICK TO PLAY VIDEO ON SITE

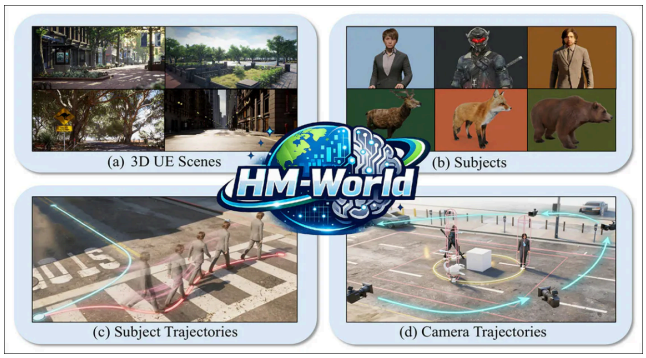
Method

The central plank of the new work is *hybrid memory*, which facilitates 'out of view extrapolation' – the retention of characters and their contexts while the viewer 'looks away' (or while the character themself exits from view). In this scenario, the framework is required to perform *spatiotemporal decoupling*, wherein it is simultaneously focused on the viewer-visible generation, and the off-screen existence of the now out-of-view character.



Examples of entry/exit camera motion. In these instances, it is the camera's movement that causes the character to exit the frame, but in diverse samples we can themselves temporarily propelling themselves offscreen. Source

The authors note that in diffusion latent embeddings, the features that need to be extracted and used are heavily entangled with other features and properties; and that attempting to extract them often causes the subject to 'freeze' into the background. Therefore they devised and curated the *HM-World* dataset**, specifically aimed at training hybrid memory:



From the paper, samples from the four categories contained in the HM-World dataset.

The collection is constructed along four dimensions: *subject trajectories*, *camera trajectories*, *scenes*, and *subjects*.

The synthetic data in HM-World features 17 scenes and 49 subjects, including people of diverse appearance, as well as animals of multiple species. Combinations of these are procedurally placed in a scene via Unreal Engine, each with a distinct motion animation, and then set upon a randomly-selected trajectory.

The authors state that a variegated set of *exit-entry* events are depicted in the dataset, with 28 different camera trajectories included, each with multiple starting-points.

The final collection comes to 59,225 video-clips, each one annotated by the MiniCPM-V Multimodal Large Language Model (MLLM).

The researchers point out the statistical advantages of their collection against prior datasets *WorldScore*; *Context-As-Memory*; *Multi-Cam Video*; and *360° Motion*:

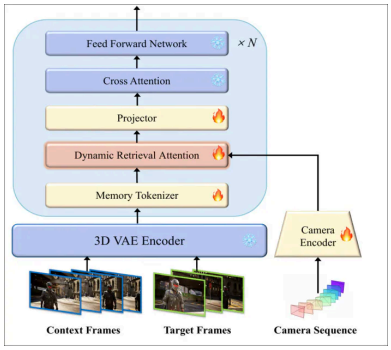
Dataset	Reference	Dynamic Subject	Subject Exit-Enter	Subject Pose	Camera Movable	Total Num.
WorldScore [35]	ICCV 25	✓	✗	✗	✓	3K
Context-As-Memory [15]	SIGGRAPH Asia 25	✗	✗	✗	✓	10K
Multi-Cam Video [22]	ICCV 25	✓	✗	✗	✓	136K
360°-Motion [33]	ICLR 25	✓	✗	✓	✗	5.4K
HM-World (ours)	-	✓	✓	✓	✓	59K

Comparison between existing datasets and the HM-World dataset, where 'Dynamic Subject' indicates the presence of moving entities, 'Subject Exit-Enter' denotes clips containing subjects leaving and re-entering the frame, and 'Subject Pose' refers to the inclusion of annotated 3D poses.

The Path Less Traveled

Given several past frames and a known camera path, the task is to predict future views as the viewer's perspective shifts, while accounting for subjects that move independently and may leave the frame before returning. This requires more than preserving a stable background, since the model must also retain a coherent internal record of how each moving subject looks and behaves, even during periods when it is not visible.

The authors' *Hybrid Dynamic Retrieval Attention* (HyDRA) method addresses this by introducing a dedicated memory pathway that separates dynamic subjects from the static scene representation, allowing them to persist over time, and to reappear with consistent appearance and motion:



Conceptual schema for the HyDRA model.

HyDRA is built over Wan2.1-T2V-1.3B, with the core diffusion pipeline left largely intact, while introducing a modified transformer block that incorporates dynamic retrieval attention. This allows the model to selectively recall motion and appearance cues from past frames, rather than relying on fixed or local context.

This process utilizes an adapted Flow Matching training objective in place of standard diffusion loss.

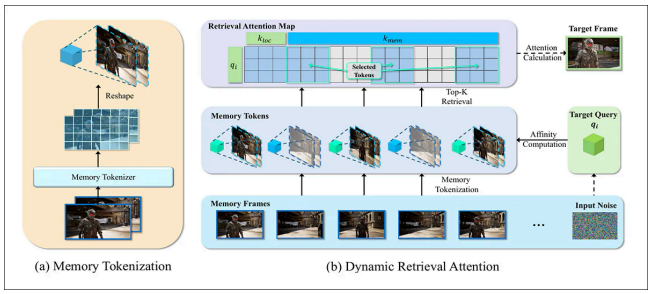
To keep scenes aligned with camera motion, camera trajectories are injected as an explicit conditioning signal, with each frame's pose defined by rotation and translation, and then converted into a compact representation capturing how the viewpoint evolves over time.

In line with the prior (Kling) ReCamMaster initiative, the result is then parsed by camera encoder, implemented as a Multi-Layer Perceptron, then broadcast and added to the Diffusion Transformer features, allowing the model to maintain consistent object-placement as the camera moves.

Tokenization

Raw diffusion latents mix subject motion, appearance, and background into a single entangled representation, and trying to retrieve directly from this space risks to introduce irrelevant context, or cause moving subjects to 'blend into' the scenery.

HyDRA addresses this with a 3D-convolution-based Memory Tokenizer that processes space and time together – rather than forwarding full latent histories, it compresses them into compact, motion-aware memory tokens that preserve how subjects look and move:



Overview of HyDRA. Left, the Memory Tokenizer converts past frames into compact, motion-aware memory tokens; right, Dynamic Retrieval Attention evaluates the current query against these tokens, retrieves the most relevant ones, and uses them to restore consistent appearance and motion in the generated frame.

These tokens form a structured hybrid memory that filters noise while retaining long-range dynamics. Passed to the Dynamic Retrieval Attention module, these allow the model to selectively recall off-screen subjects, so that they reappear with consistent appearance, motion, and context.

Dynamic Retrieval Attention

HyDRA's dual memory mechanism also uses *dynamic retrieval attention* in a distinct but complementary role within the framework.

Memory tokenization compresses past latent representations into structured, motion-aware tokens that separate dynamic subjects from static scene content, reducing the entanglement that often causes subjects to blend into the background. These tokens form a persistent memory bank rather than a full frame history.

Dynamic Retrieval Attention then operates over this bank during generation, evaluating the current query against stored tokens and selectively recalling those most relevant to the evolving frame. This allows off-screen subjects to continue their latent evolution (i.e., to keep walking, running, when you can't see them), and to reappear with consistent appearance and motion when they return to view, instead of resetting or degrading.

Data and Tests

In tests, the Wan-based HyDRA system encoded and downsampled 77 context frames before parsing them with a 3D Variational Autoencoder (VAE), while the aforementioned memory tokenizer used 3D convolution at a kernel size of 2x4x4.

The model was trained on HW-World for 10,000 iterations on 32 (unspecified) GPUs, at a batch size of 32.

An unusually high number of metrics were used in the tests: besides the customary Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Similarity Metrics (LPIPS), the authors also employed *subject consistency* and *background consistency* from the VBench suite, to evaluate frame-level coherence.

Additionally, they devised a custom metric titled *Dynamic Subject Consistency* (DSC), which uses bounding boxes from YOLO V11, to create cropped regions featuring moving subjects, from which semantic features were extracted and their similarities then calculated.

HyDRA was pitched against Diffusion Forcing Transformer (DFoT), and Context-As-Memory, over a baseline Wan2.1-T2V-1.3B model outfitted with a camera encoder (to represent the subjective viewpoint common to all the clips). All models were trained on HW-World, and WorldPlay was also used as a zero-shot, secondary testing collection:

In initial quantitative comparisons, HyDRA outperformed all baselines, raising PSNR from 18.696 to 20.357, and SSIM from 0.517 to 0.606. It also achieved the highest contextual and ground truth Dice scores, 0.827 and 0.849, with Subject and Background Consistency reaching 0.926 and 0.932:

Method	Reference	PSNR	SSIM	LPIPS	DSC_{ctx}	DSC_{GT}	Subj. Cons.	Bg. Cons.
Baseline	-	18.696	0.517	0.356	0.812	0.837	0.903	0.925
DFoT [20]	ICML 25	17.693	0.482	0.410	0.803	0.826	0.893	0.913
Context-as-Memory [15]	SIGGRAPH Asia 25	18.921	0.530	0.342	0.816	0.839	0.911	0.922
HyDRA (ours)	-	**20.357**	**0.606**	**0.289**	**0.827**	**0.849**	**0.926**	**0.932**

Results of the initial quantitative comparison against prior approaches.

DFoT reached 17.693 PSNR and Context as Memory 18.921, with the gains attributed to memory tokenization combined with dynamic retrieval attention:

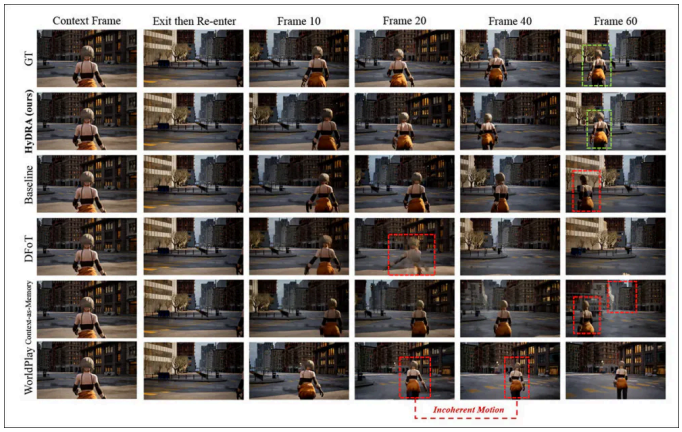
Method	PSNR	SSIM	LPIPS	DSC_{ctx}	DSC_{GT}	Subject Consistency	Background Consistency
WorldPlay [14]	14.855	0.355	0.500	0.822	0.832	0.910	0.925
HyDRA (ours)	**20.357**	**0.606**	**0.289**	**0.827**	**0.849**	**0.926**	**0.932**

Quantitative comparison pitching HyDRA against the current state-of-the-art.

Regarding the tests against WorldPlay, the authors state:

> *'Our method surpasses WorldPlay across all metrics, with a notable PSNR gap of 5.502. Although WorldPlay exhibits lower performance on GT-referenced metrics (e.g., PSNR of 14.855, DSCGT of 0.832) due to domain distribution gap and lack of specific finetuning, it demonstrates remarkable robustness on context-referenced metrics by achieving a DSCctx of 0.822.*
>
> *'This observation not only confirms that extensively trained models possess fair hybrid consistency but also indirectly validates the rationality of our proposed DSC metrics in reflecting dynamic subject consistency.*
>
> *'Ultimately, these impressive results highlight the exceptional capabilities of our model, demonstrating its superiority even over established commercial models.'*

The paper offers static representation of qualitative comparisons undertaken for the tests:

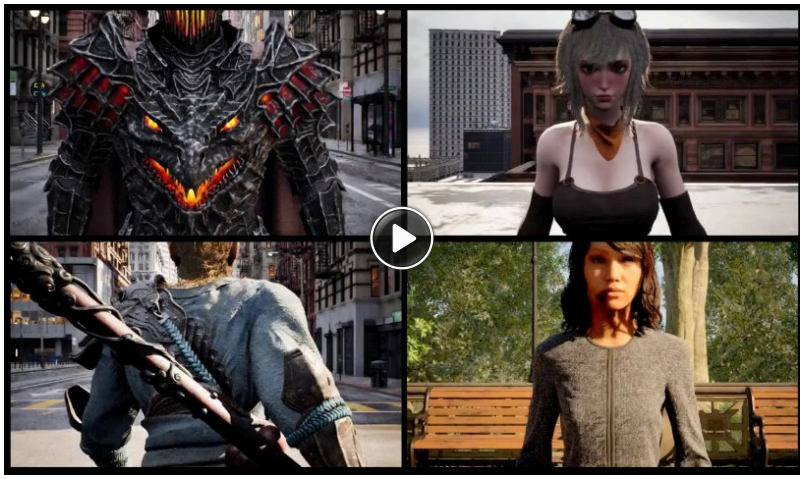


Qualitative comparison of exit and re entry under camera motion. The authors assert that HyDRA preserves subject identity, pose, and motion continuity after leaving and returning to the frame, closely matching ground truth, whereas competing methods exhibit drift, incoherent motion, or subject degradation, highlighted in red (consistent recoveries are marked in green).

Of these results, the authors comment:

> *'In the case of complex exit-and-entry events, the baseline and Context-as-Memory exhibit severe subject distortion and motion incoherence. DFoT fails to maintain subject integrity, leading to complete vanishing. While WorldPlay manages to preserve the subject's appearance consistency, it suffers from stuttering movements and unnatural actions.*
>
> *'In contrast, our method successfully maintains hybrid consistency, preserving both the subject's identity and motion coherence after the subject re-enters the frame.'*

Further results can be seen in video format at the supplementary site, of which the first four examples have been assembled (by us) into the video below:



CLICK TO PLAY VIDEO ON SITE

Click to play. *Four of the six test results featured at the project site.* Source

Conclusion

While any attempt to address one of the biggest bugbears of AI video generation is welcome, it seems inevitable to me that the optimal solution for exit/re-entry issues of this kind will prove to be, as it was with CGI, in the form of distinct reference materials that can be discretely edited and brought into a composer-space.

This business of trying to keep an embedding alive in an *ad hoc* and on-the-fly manner seems exhausting, and also offers no clear way forward to the intra-shot consistency now on offer at various black-box portals such as Runway. If it turns out that a follow-on shot will require access to the latent space of the prior shot, why not have both instances place a discrete and separate character embedding?

* *No-one else has named it, and discussion is difficult without common terms.*

** *Currently reported to be 'coming soon', at the project page.*

First published Friday, March 27, 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