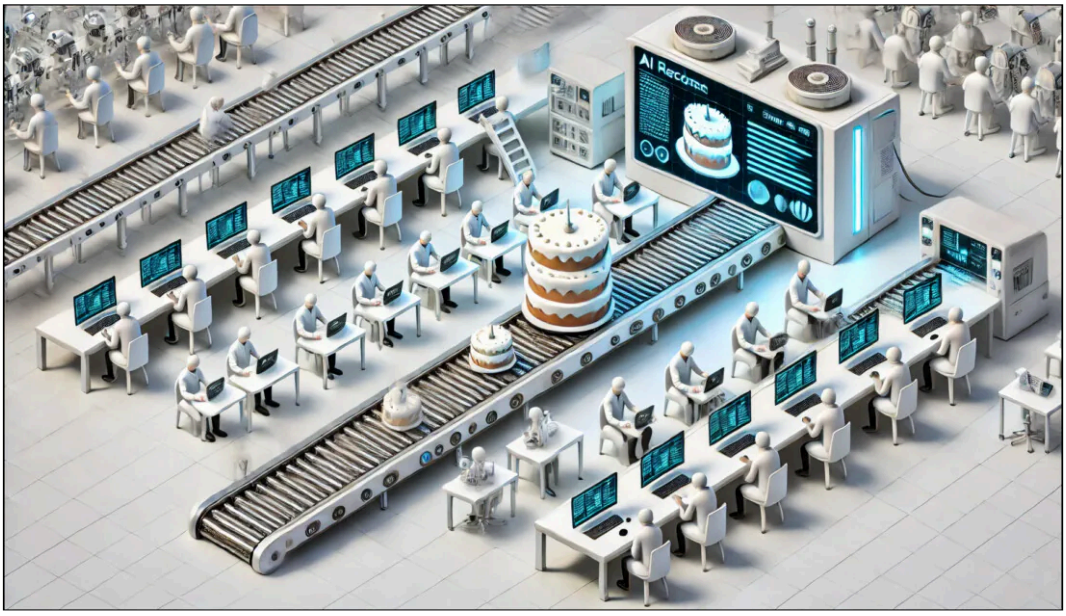
# Cooking Up Narrative Consistency for Long Video Generation

Originally published at https://www.unite.ai/cooking-up-narrative-consistency-for-long-video-generation/




The recent public release of the Hunyuan Video generative AI model has intensified ongoing discussions about the potential of large multimodal vision-language models to one day create entire movies.

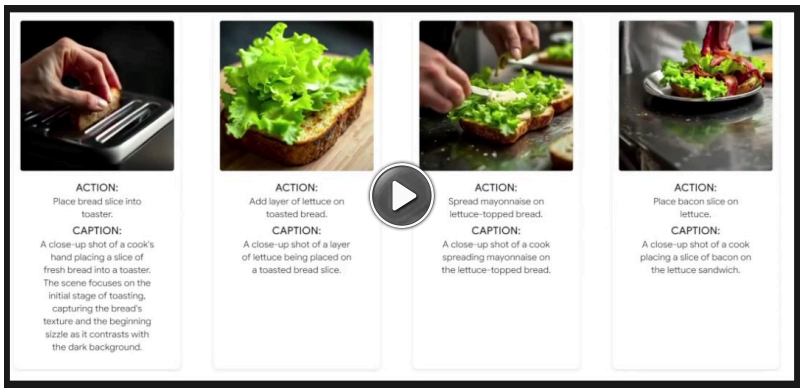
However, as we have observed, this is a very distant prospect at the moment, for a number of reasons. One is the very short attention window of most AI video generators, which struggle to maintain consistency even in a short single shot, let alone a series of shots.

Another is that consistent references to video content (such as explorable environments, which should not change randomly if you retrace your steps through them) can only be achieved in diffusion models by customization techniques such as low-rank adaptation (LoRA), which limits the out-of-the-box capabilities of foundation models.

Therefore the evolution of generative video seems set to stall unless new approaches to narrative continuity are developed.

## Recipe for Continuity

With this in mind, a new collaboration between the US and China has proposed the use of *instructional cooking videos* as a possible template for future narrative continuity systems.

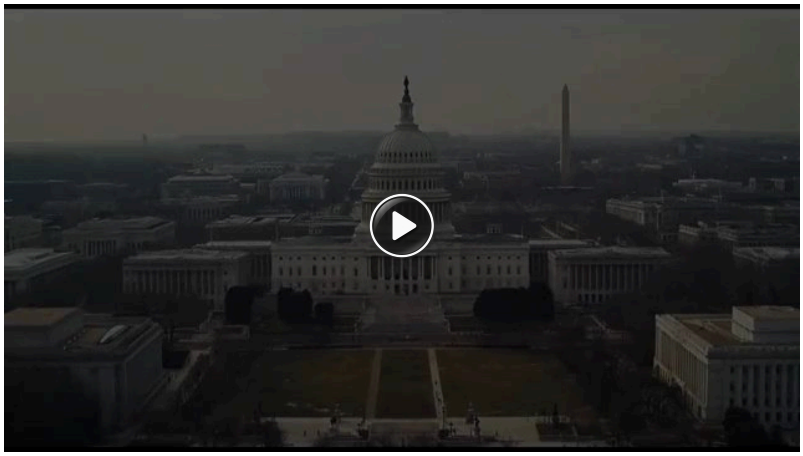



CLICK TO PLAY VIDEO ON SITE

***Click to play.*** *The VideoAuteur project systematizes the analysis of parts of a cooking process, to produce a finely-captioned new dataset and an orchestration method for the generation of cooking videos. Refer to source site for better resolution.* Source: https://videoauteur.github.io/

Titled *VideoAuteur*, the work proposes a two-stage pipeline to generate instructional cooking videos using cohered states combining keyframes and captions, achieving state-of-the-art results in – admittedly – an under-subscribed space.

VideoAuteur's project page also includes a number of rather more attention-grabbing videos that use the same technique, such as a proposed trailer for a (non-existent) Marvel/DC crossover:



CLICK TO PLAY VIDEO ON SITE

***Click to play.*** *Two superheroes from alternate universes come face to face in a fake trailer from VideoAuteur. Refer to source site for better resolution.*

The page also features similarly-styled promo videos for an equally non-existent Netflix animal series and a Tesla TSLA 2.60% car ad.

In developing VideoAuteur, the authors experimented with diverse loss functions, and other novel approaches. To develop a recipe how-to generation workflow, they also curated *CookGen*, the largest dataset focused on the cooking domain, featuring 200, 000 video clips with an average duration of 9.5 seconds.

At an average of 768.3 words per video, CookGen is comfortably the most extensively-annotated dataset of its kind. Diverse vision/language models were used, among other approaches, to ensure that descriptions were as detailed, relevant and accurate as possible.

Cooking videos were chosen because cooking instruction walk-throughs have a structured and unambiguous narrative, making annotation and evaluation an easier task. Except for pornographic videos (likely to enter this particular space sooner rather than later), it is difficult to think of any other genre quite as visually and narratively 'formulaic'.

The authors state:

> *'Our proposed two-stage auto-regressive pipeline, which includes a long narrative director and visual-conditioned video generation, demonstrates promising improvements in semantic consistency and visual fidelity in generated long narrative videos.*
>
> *Through experiments on our dataset, we observe enhancements in spatial and temporal coherence across video sequences.*
>
> *'We hope our work can facilitate further research in long narrative video generation.'*

The new work is titled *VideoAuteur: Towards Long Narrative Video Generation*, and comes from eight authors across Johns Hopkins University, ByteDance, and ByteDance Seed.

## Dataset Curation

To develop CookGen, which powers a two-stage generative system for producing AI cooking videos, the authors used material from the YouCook and HowTo100M collections. The authors compare the scale of CookGen to previous datasets focused on narrative development in generative video, such as the Flintstones dataset, the Pororo cartoon dataset, StoryGen, Tencent's StoryStream, and VIST.

| Datasets | Modality | Type | # Images | Text Length |
|---|---|---|---|---|
| Flintstones | Image | Comic | 122k | 86 |
| Pororo | Image | Comic | 74k | 74 |
| StorySalon | Image | Comic | 160k | 106 |
| StoryStream | Image | Comic | 258k | 146 |
| VIST | Image | Real world | 210K | ~70 |
| **CookGen** | Video | Real world | 39M | 763.8 |

*Comparison of images and text length between CookGen and the nearest-most populous similar datasets.*
Source: https://arxiv.org/pdf/2501.06173

CookGen focuses on real-world narratives, particularly procedural activities like cooking, offering clearer and easier-to-annotate stories compared to image-based comic datasets. It exceeds the largest existing dataset, StoryStream, with 150x more frames and 5x denser textual descriptions.

The researchers fine-tuned a captioning model using the methodology of LLaVA-NeXT as a base. The automatic speech recognition (ASR) pseudo-labels obtained for HowTo100M were used as 'actions' for each video, and then refined further by large language models (LLMs).

For instance, ChatGPT-4o was used to produce a caption dataset, and was asked to focus on subject-object interactions (such as hands handling utensils and food), object attributes, and temporal dynamics.

Since ASR scripts are likely to contain inaccuracies and to be generally 'noisy', Intersection-over-Union (IoU) was used as a metric to measure how closely the captions conformed to the section of the video they were addressing. The authors note that this was crucial for the creation of narrative consistency.

The curated clips were evaluated using Fréchet Video Distance (FVD), which measures the disparity between ground truth (real world) examples and generated examples, both with and without ground truth keyframes, arriving at a performative result:

| Validation Set | w/. GT keyframe | W/o. GT keyframe |
|---|---|---|
| # Clips | FVD | FVD |
| 5504 | **116.3** | 561.1 |

*Using FVD to evaluate the distance between videos generated with the new captions, both with and without the use of keyframes captured from the sample videos.*

Additionally, the clips were rated both by GPT-4o, and six human annotators, following LLaVA-Hound's definition of 'hallucination' (i.e., the capacity of a model to invent spurious content).

The researchers compared the quality of the captions to the Qwen2-VL-72B collection, obtaining a slightly improved score.

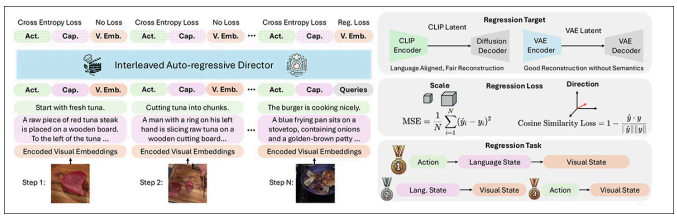
| | GPT-4o Evaluation | | Human Evaluation | |
|---|---|---|---|---|
| Score (0-100) | Qwen2-VL-72B | **Ours** | Qwen2-VL-72B | **Ours** |
| | **98.0** | 95.2 | 79.3 | **82.0** |

*Comparison of FVD and human evaluation scores between Qwen2-VL-72B and the authors' collection.*

## Method

VideoAuteur's generative phase is divided between the *Long Narrative Director* (LND) and the *visual-conditioned video generation model* (VCVGM).

LND generates a sequence of visual embeddings or keyframes that characterize the narrative flow, similar to 'essential highlights'. The VCVGM generates video clips based on these choices.




*Schema for the VideoAuteur processing pipeline. The Long Narrative Video Director makes apposite selections to feed to the Seed-X-powered generative module.*

The authors extensively discuss the differing merits of an *interleaved image-text director* and a language-centric keyframe director, and conclude that the former is the more effective approach.

The interleaved image-text director generates a sequence by interleaving text tokens and visual embeddings, using an auto-regressive model to predict the next token, based on the combined context of both text and images. This ensures a tight alignment between visuals and text.

By contrast, the language-centric keyframe director synthesizes keyframes using a text-conditioned diffusion model based solely on captions, without incorporating visual embeddings into the generation process.

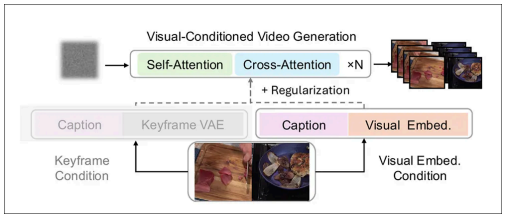
The researchers found that while the language-centric method generates visually appealing keyframes, it lacks consistency across frames, arguing that the interleaved method achieves higher scores in realism and visual consistency. They also found that this method was better able to learn a realistic visual style through training, though sometimes with some repetitive or noisy elements.

Unusually, in a research strand dominated by the co-opting of Stable Diffusion and Flux into workflows, the authors used Tencent's SEED-X 7B-parameter multi-modal LLM foundation model for their generative pipeline (though this model does leverage Stability.ai's SDXL release of Stable Diffusion for a limited part of its architecture).

The authors state:

> *'Unlike the classic Image-to-Video (I2V) pipeline that uses an image as the starting frame, our approach leverages [regressed visual latents] as continuous conditions throughout the [sequence].*
>
> *'Furthermore, we improve the robustness and quality of the generated videos by adapting the model to handle noisy visual embeddings, since the regressed visual latents may not be perfect due to regression errors.'*

Though typical visual-conditioned generative pipelines of this kind often use initial keyframes as a starting point for model guidance, VideoAuteur expands on this paradigm by generating multi-part visual states in a semantically coherent latent space, avoiding the potential bias of basing further generation solely on 'starting frames'.



*Schema for the use of visual state embeddings as a superior conditioning method.*

## Tests

In line with the methods of SeedStory, the researchers use SEED-X to apply LoRA fine-tuning on their narrative dataset, enigmatically describing the result as a 'Sora-like model', pre-trained on large-scale video/text couplings, and capable of accepting both visual and text prompts and conditions.

32,000 narrative videos were used for model development, with 1,000 held aside as validation samples. The videos were cropped to 448 pixels on the short side and then center-cropped to 448x448px.

For training, the narrative generation was evaluated primarily on the YouCook2 validation set. The Howto100M set was used for data quality evaluation and also for image-to-video generation.

For visual conditioning loss, the authors used diffusion loss from DiT and a 2024 work based around Stable Diffusion.

To prove their contention that interleaving is a superior approach, the authors pitted VideoAuteur against several methods that rely solely on text-based input: EMU-2, SEED-X, SDXL and FLUX.1-schnell (FLUX.1-s).


*Given a global prompt, 'Step-by-step guide to cooking mapo tofu', the interleaved director generates actions, captions, and image embeddings sequentially to narrate the process. The first two rows show keyframes decoded from EMU-2 and SEED-X latent spaces. These images are realistic and consistent but less polished than those from advanced models like SDXL and FLUX.*

The authors state:

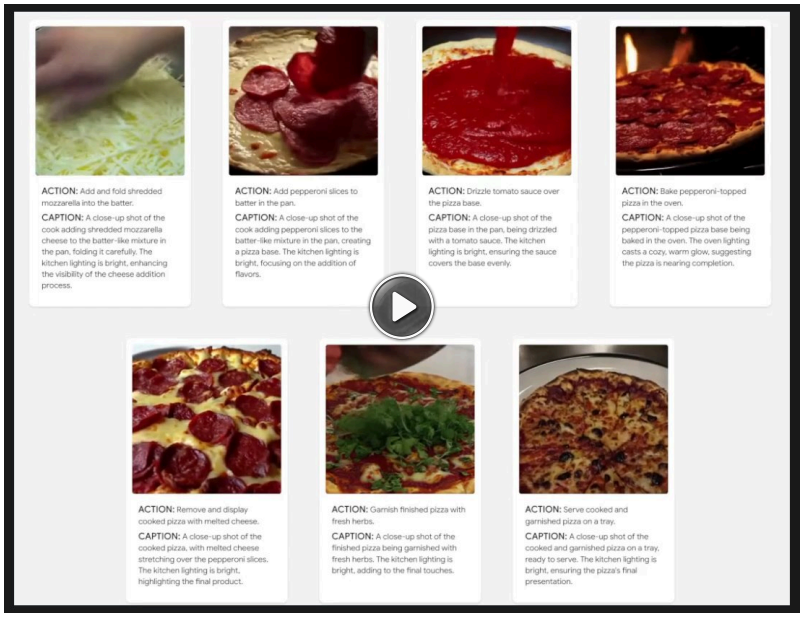
> *'The language-centric approach using text-to-image models produces visually appealing keyframes but suffers from a lack of consistency across frames due to limited mutual information. In contrast, the interleaved generation method leverages language-aligned visual latents, achieving a realistic visual style through training.*
>
> *'However, it occasionally generates images with repetitive or noisy elements, as the auto-regressive model struggles to create accurate embeddings in a single pass.'*

Human evaluation further confirms the authors' contention about the improved performance of the interleaved approach, with interleaved methods achieving the highest scores in a survey.

| Latent condition | Gen. strategy | Aesthetic | Realistic | Visual consist. | Narrative |
|---|---|---|---|---|---|
| EMU-2 Latent | Interleaved | 0.7 | 1.2 | 2.9 | 2.2 |
| SEED-X Latent | Interleaved | 2.1 | **4.3** | **4.5** | **4.4** |
| Text (SDXL) | Language-centric | 4.0 | 2.9 | 3.3 | 4.0 |
| Text (FLUX.1-s) | Language-centric | **4.8** | 3.1 | 3.4 | **4.4** |

*Comparison of approaches from a human study conducted for the paper.*

However we note that language-centric approaches achieve the best *aesthetic* scores. The authors contend, however, that this is not the central issue in the generation of long narrative videos.




CLICK TO PLAY VIDEO ON SITE

***Click to play.*** *Segments generated for a pizza-building video, by VideoAuteur.*

## Conclusion

The most popular strand of research in regard to this challenge, i.e., narrative consistency in long-form video generation, is concerned with single images. Projects of this kind include DreamStory, StoryDiffusion, TheaterGen and NVIDIA's ConsiStory.

In a sense, VideoAuteur also falls into this 'static' category, since it makes use of seed images from which clip-sections are generated. However, the interleaving of video and semantic content brings the process a step nearer to a practical pipeline.

*First published Thursday, January 16, 2025*

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**