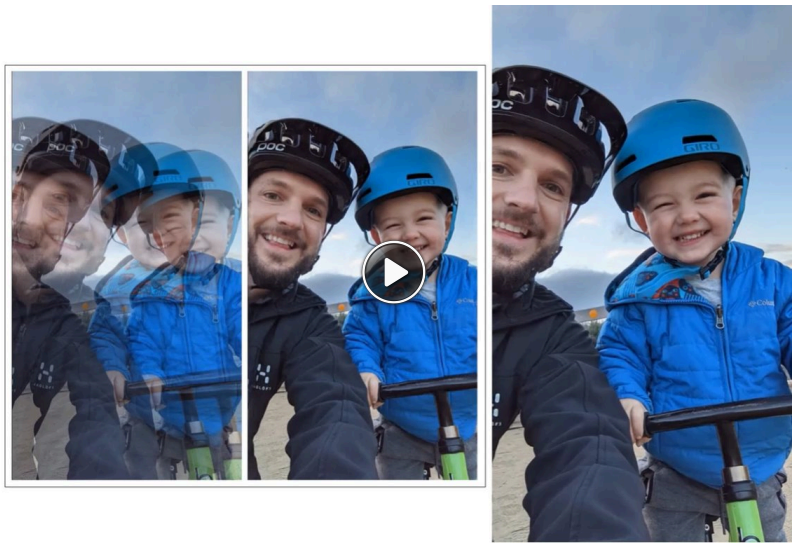
# Generating Better AI Video From Just Two Images

Originally published at https://www.unite.ai/generating-better-ai-video-from-just-two-images/

Video frame interpolation (VFI) is an open problem in generative video research. The challenge is to generate intermediate frames between two existing frames in a video sequence.



CLICK TO PLAY VIDEO ON SITE

***Click to play.*** *The FILM framework, a collaboration between Google and the University of Washington, proposed an effective frame interpolation method that remains popular in hobbyist and professional spheres. On the left, we can see the two separate and distinct frames superimposed; in the middle, the 'end frame'; and on the right, the final synthesis between the frames.* Sources: https://film-net.github.io/ and https://arxiv.org/pdf/2202.04901

Broadly speaking, this technique dates back over a century, and has been used in traditional animation since then. In that context, master 'keyframes' would be generated by a principal animation artist, while the work of 'tweening' intermediate frames would be carried out as by other staffers, as a more menial task.

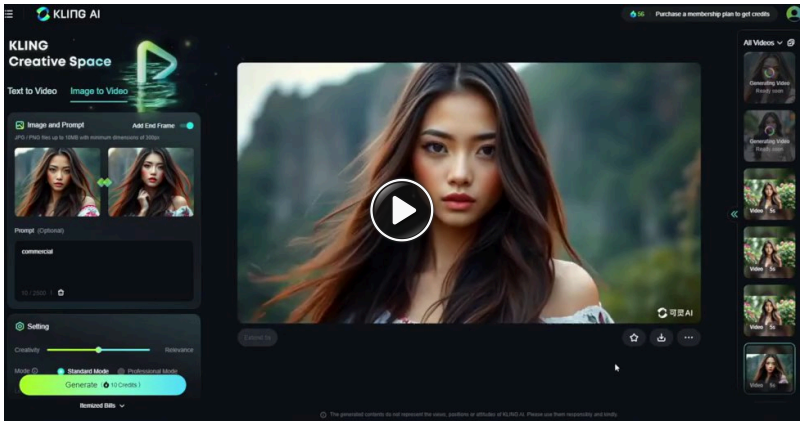
Prior to the rise of generative AI, frame interpolation was used in projects such as Real-Time Intermediate Flow Estimation (RIFE), Depth-Aware Video Frame Interpolation (DAIN), and Google's Frame Interpolation for Large Motion (FILM – see above) for purposes of increasing the frame rate of an existing video, or enabling artificially-generated slow-motion effects. This is accomplished by splitting out the existing frames of a clip and generating estimated intermediate frames.

VFI is also used in the development of better video codecs, and, more generally, in optical flow-based systems (including generative systems), that utilize advance knowledge of coming keyframes to optimize and shape the interstitial content that precedes them.

## End Frames in Generative Video Systems

Modern generative systems such as Luma and Kling allow users to specify a start and an end frame, and can perform this task by analyzing keypoints in the two images and estimating a trajectory between the two images.

As we can see in the examples below, providing a ‘closing’ keyframe better allows the generative video system (in this case, Kling) to maintain aspects such as identity, even if the results are not perfect (particularly with large motions).

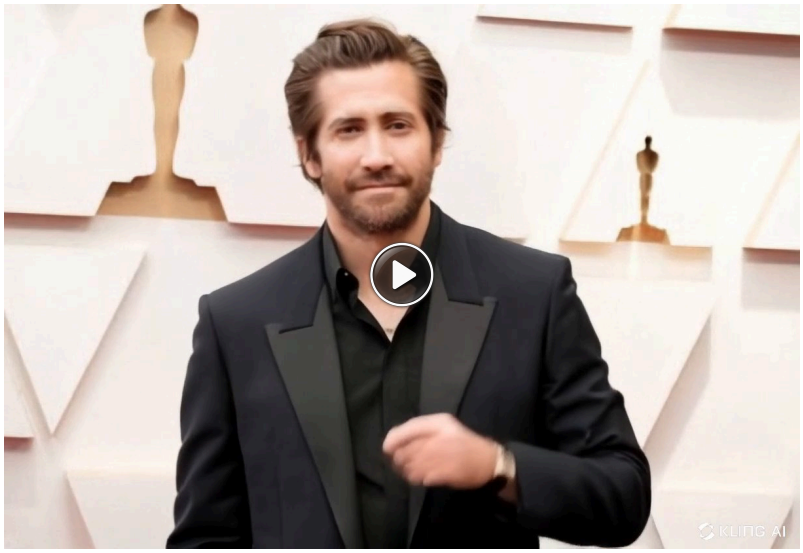



CLICK TO PLAY VIDEO ON SITE

***Click to play.*** *Kling is one of a growing number of video generators, including Runway and Luma, that allow the user to specify an end frame. In most cases, minimal motion will lead to the most realistic and least-flawed results.* Source: https://www.youtube.com/watch?v=8oylqODAaH8

In the above example, the person’s identity is consistent between the two user-provided keyframes, leading to a relatively consistent video generation.

Where only the starting frame is provided, the generative systems window of attention is not usually large enough to ‘remember’ what the person looked like at the start of the video. Rather, the identity is likely to shift a little bit with each frame, until all resemblance is lost. In the example below, a starting image was uploaded, and the person’s movement guided by a text prompt:




CLICK TO PLAY VIDEO ON SITE

***Click to play.*** *With no end frame, Kling only has a small group of immediately prior frames to guide the generation of the next frames. In cases where any significant movement is needed, this atrophy of identity becomes severe.*

We can see that the actor’s resemblance is not resilient to the instructions, since the generative system does not know what he would look like if he was smiling, and he is not smiling in the seed image (the only available reference).

The majority of viral generative clips are carefully curated to de-emphasize these shortcomings. However, the progress of temporally consistent generative video systems may depend on new developments from the research sector in regard to frame interpolation, since the only possible alternative is a dependence on traditional CGI as a driving, ‘guide’ video (and even in this case, consistency of texture and lighting are currently difficult to achieve).

Additionally, the slowly-iterative nature of deriving a new frame from a small group of recent frames makes it very difficult to achieve large and bold motions. This is because an object that is moving rapidly across a frame may transit from one side to the other in the space of a single frame, contrary to the more gradual movements on which the system is likely to have been trained.

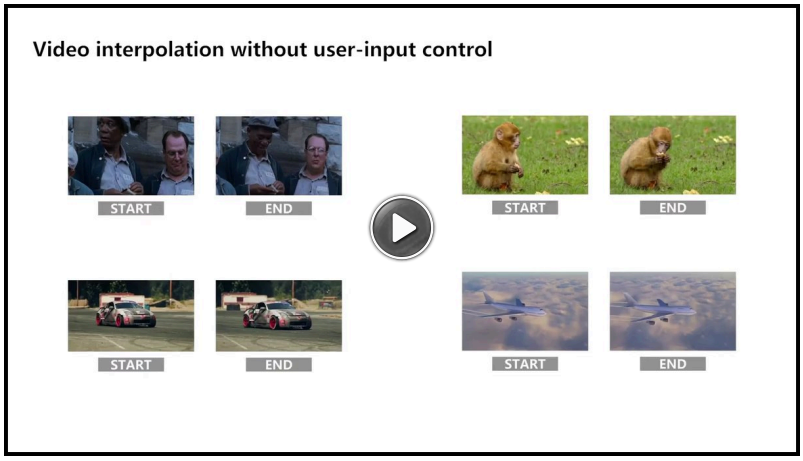
Likewise, a significant and bold change of pose may lead not only to identity shift, but to vivid non-congruities:


CLICK TO PLAY VIDEO ON SITE

***Click to play.*** *In this example from Luma, the requested movement does not appear to be well-represented in the training data.*

## Framer

This brings us to an interesting recent paper from China, which claims to have achieved a new state-of-the-art in authentic-looking frame interpolation – and which is the first of its kind to offer drag-based user interaction.

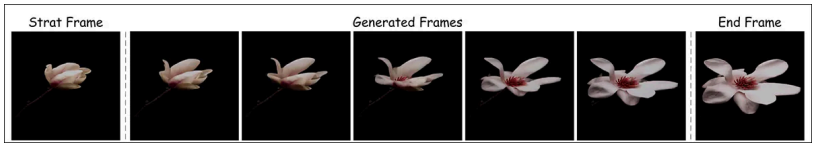



CLICK TO PLAY VIDEO ON SITE

*Framer allows the user to direct motion using an intuitive drag-based interface, though it also has an 'automatic' mode*. Source: https://www.youtube.com/watch?v=4MPGKgn7jRc

Drag-centric applications have become frequent in the literature lately, as the research sector struggles to provide instrumentalities for generative system that are not based on the fairly crude results obtained by text prompts.

The new system, titled *Framer*, can not only follow the user-guided drag, but also has a more conventional 'autopilot' mode. Besides conventional tweening, the system is capable of producing time-lapse simulations, as well as morphing and novel views of the input image.




*Interstitial frames generated for a time-lapse simulation in Framer.* Source: https://arxiv.org/pdf/2410.18978

In regard to the production of novel views, Framer crosses over a little into the territory of Neural Radiance Fields (NeRF) – though requiring only two images, whereas NeRF generally requires six or more image input views.

In tests, Framer, which is founded on Stability.ai's Stable Video Diffusion latent diffusion generative video model, was able to outperform approximated rival approaches, in a user study.

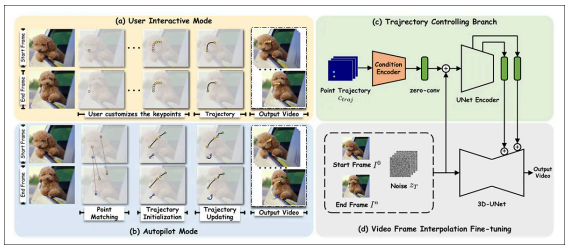
At the time of writing, the code is set to be released at GitHub. Video samples (from which the above images are derived) are available at the project site, and the researchers have also released a YouTube video.

The new paper is titled *Framer: Interactive Frame Interpolation*, and comes from nine researchers across Zhejiang University and the Alibaba-backed Ant Group.

## Method

Framer uses keypoint-based interpolation in either of its two modalities, wherein the input image is evaluated for basic topology, and 'movable' points assigned where necessary. In effect, these points are equivalent to facial landmarks in ID-based systems, but generalize to any surface.

The researchers fine-tuned Stable Video Diffusion (SVD) on the OpenVid-1M dataset, adding an additional last-frame synthesis capability. This facilitates a trajectory-control mechanism (top right in schema image below) that can evaluate a path toward the end-frame (or back from it).




*Schema for Framer.*

Regarding the addition of last-frame conditioning, the authors state:

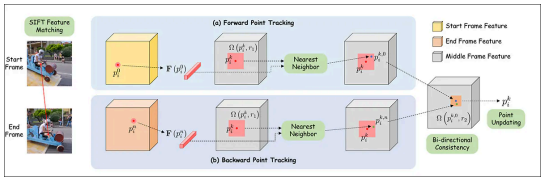
> *'To preserve the visual prior of the pre-trained SVD as much as possible, we follow the conditioning paradigm of SVD and inject end-frame conditions in the latent space and semantic space, respectively.*
>
> *'Specifically, we concatenate the VAE-encoded latent feature of the first [frame] with the noisy latent of the first frame, as did in SVD. Additionally, we concatenate the latent feature of the last frame, zn, with the noisy latent of the end frame, considering that the conditions and the corresponding noisy latents are spatially aligned.*
>
> *'In addition, we extract the CLIP image embedding of the first and last frames separately and concatenate them for cross-attention feature injection.'*

For drag-based functionality, the trajectory module leverages the Meta Ai-led CoTracker framework, which evaluates profuse possible paths ahead. These are slimmed down to between 1-10 possible trajectories.

The obtained point coordinates are then transformed through a methodology inspired by the DragNUWA and DragAnything architectures. This obtains a *Gaussian heatmap*, which individuates the target areas for movement.

Subsequently, the data is fed to the conditioning mechanisms of ControlNet, an ancillary conformity system originally designed for Stable Diffusion, and since adapted to other architectures.

For autopilot mode, feature matching is initially accomplished via SIFT, which interprets a trajectory that can then be passed to an auto-updating mechanism inspired by DragGAN and DragDiffusion.




*Schema for point trajectory estimation in Framer.*

## Data and Tests

For the fine-tuning of Framer, the spatial attention and residual blocks were frozen, and only the temporal attention layers and residual blocks were affected.

The model was trained for 10,000 iterations under AdamW, at a learning rate of 1e-4, and a batch size of 16. Training took place across 16 NVIDIA A100 GPUs.

Since prior approaches to the problem do not offer drag-based editing, the researchers opted to compare Framer's autopilot mode to the standard functionality of older offerings.

The frameworks tested for the category of current diffusion-based video generation systems were LDMVFI; Dynamic Crafter; and SVDKFI. For 'traditional' video systems, the rival frameworks were AMT; RIFE; FLAVR; and the aforementioned FILM.

In addition to the user study, tests were conducted over the DAVIS and UCF101 datasets.

Qualitative tests can only be evaluated by the objective faculties of the research team and by user studies. However, the paper notes, traditional *quantitative* metrics are largely unsuited to the proposition at hand:

> *'[Reconstruction] metrics like PSNR, SSIM, and LPIPS fail to capture the quality of interpolated frames accurately, since they penalize other plausible interpolation results that are not pixel-aligned with the original video.*

> *'While generation metrics such as FID offer some improvement, they still fall short as they do not account for temporal consistency and evaluate frames in isolation.'*

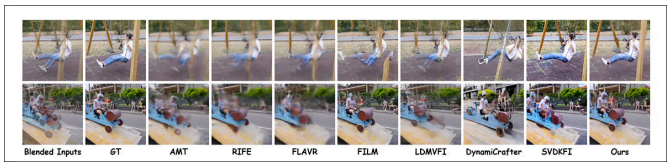
In spite of this, the researchers conducted qualitative tests with several popular metrics:

| | DAVIS-7 | | | | | UCF101-7 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | LPIPS↓ | FID↓ | FVD↓ | PSNR↑ | SSIM↑ | LPIPS↓ | FID↓ | FVD↓ |
| AMT (Li et al., 2023) | 21.66 | **0.7229** | 0.2860 | 39.17 | 245.25 | 26.64 | 0.9000 | 0.1878 | 37.80 | 270.98 |
| RIFE (Huang et al., 2020) | **22.00** | 0.7216 | 0.2663 | 39.16 | 319.79 | **27.04** | **0.9020** | 0.1575 | 27.96 | 300.40 |
| FLAVR Kalluri et al. (2023) | 20.94 | 0.6880 | 0.3305 | 52.23 | 296.37 | 26.50 | 0.8982 | 0.1836 | 37.79 | 279.58 |
| FILM (Reda et al., 2022) | 21.67 | 0.7121 | **0.2191** | **17.20** | 162.86 | 26.74 | 0.8983 | **0.1378** | **16.22** | 239.48 |
| LDMVFI (Danier et al., 2024) | 21.11 | 0.6900 | 0.2535 | 21.96 | 269.72 | 26.68 | 0.8955 | 0.1446 | 17.55 | 270.33 |
| DynamicCrafter (Xing et al., 2023) | 15.48 | 0.4668 | 0.4628 | 35.95 | 468.78 | 17.62 | 0.7082 | 0.3361 | 61.71 | 646.91 |
| SVDKFI (Wang et al., 2024a) | 16.71 | 0.5274 | 0.3440 | 26.59 | 382.19 | 21.04 | 0.7991 | 0.2146 | 44.81 | 301.33 |
| **Framer (Ours)** | 21.23 | 0.7218 | 0.2525 | 27.13 | **115.65** | 25.04 | 0.8806 | 0.1714 | 31.69 | **181.55** |
| Framer with Co-Tracker (Ours) | 22.75 | 0.7931 | 0.2199 | 27.43 | 102.31 | 27.08 | 0.9024 | 0.1714 | 32.37 | 159.87 |

*Quantitative results for Framer vs. rival systems.*

The authors note that in spite of having the odds stacked against them, Framer still achieves the best FVD score among the methods tested.

Below are the paper's sample results for a qualitative comparison:

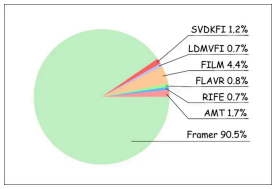



*Qualitative comparison against former approaches. Please refer to the paper for better resolution, as well as video results at https://www.youtube.com/watch?v=4MPGKgn7jRc.*

The authors comment:

> *'[Our] method produces significantly clearer textures and natural motion compared to existing interpolation techniques. It performs especially well in scenarios with substantial differences between the input frames, where traditional methods often fail to interpolate content accurately.*
>
> *'Compared to other diffusion-based methods like LDMVFI and SVDKFI, Framer demonstrates superior adaptability to challenging cases and offers better control.'*

For the user study, the researchers gathered 20 participants, who assessed 100 randomly-ordered video results from the various methods tested. Thus, 1000 ratings were obtained, evaluating the most 'realistic' offerings:




*Results from the user study.*

As can be seen from the graph above, users overwhelmingly favored results from Framer.

The project's accompanying YouTube video outlines some of the potential other uses for framer, including morphing and cartoon in-betweening – where the entire concept began.

## Conclusion

It is hard to over-emphasize how important this challenge currently is for the task of AI-based video generation. To date, older solutions such as FILM and the (non-AI) EbSynth have been used, by both amateur and professional communities, for tweening between frames; but these solutions come with notable limitations.

Because of the disingenuous curation of official example videos for new T2V frameworks, there is a wide public misconception that machine learning systems can accurately infer geometry in motion without recourse to guidance mechanisms such as 3D morphable models (3DMMs), or other ancillary approaches, such as LoRAs.

To be honest, tweening itself, even if it could be perfectly executed, only constitutes a 'hack' or cheat upon this problem. Nonetheless, since it is often easier to produce two well-aligned frame images than to effect guidance via text-prompts or the current range of alternatives, it is good to see iterative progress on an AI-based version of this older method.

*First published Tuesday, October 29, 2024*

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**