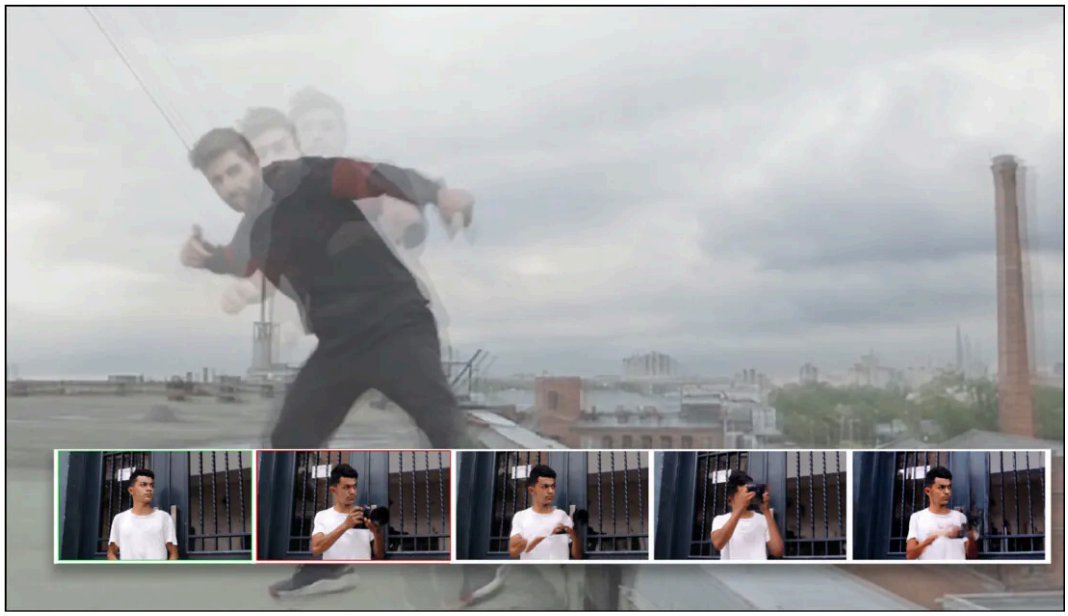
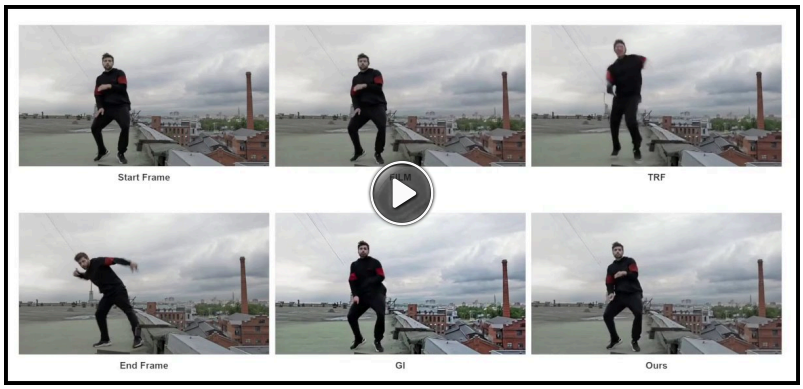
Bridging the ‘Space Between’ in Generative Video

Originally published at https://www.unite.ai/bridging-the-space-between-in-generative-video/



New research from China is offering an improved method of interpolating the gap between two temporally-distanced video frames – one of the most crucial challenges in the current race towards realism for generative AI video, as well as for video codec compression.

In the example video below, we see in the leftmost column a ‘start’ (above left) and ‘end’ (lower left) frame. The task that the competing systems must undertake is to guess how the subject in the two pictures would get from frame A to frame B. In animation, this process is called *tweening*, and harks back to the silent era of movie-making.



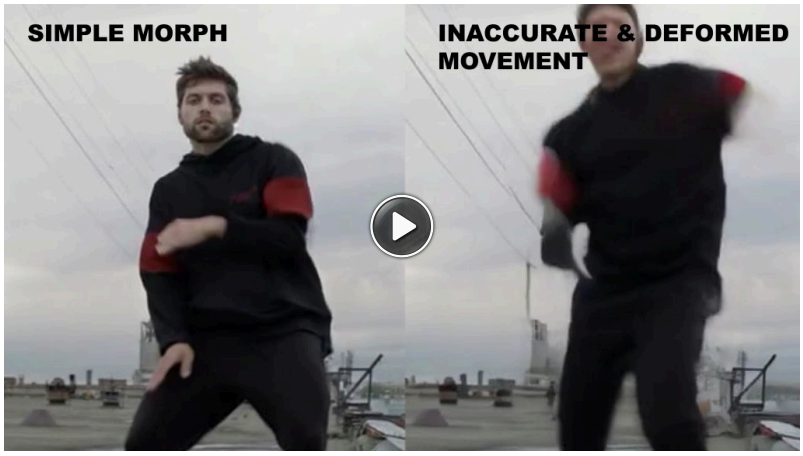
CLICK TO PLAY VIDEO ON SITE

Click to play. *In the first, left-most column, we see the proposed start and end frame. In the middle column, and at the top of the third (rightmost) column, we see three prior approaches to this challenge. Lower right, we see that the new method obtains a far more convincing result in providing the interstitial frames*. Source: https://fcvg-inbetween.github.io/

The new method proposed by the Chinese researchers is called *Frame-wise Conditions-driven Video Generation* (FCVG), and its results can be seen in the lower-right of the video above, providing a smooth and logical transition from one still frame to the next.

By contrast, we can see that one of the most celebrated frameworks for video interpolation, Google’s Frame Interpolation for Large Motion (FILM) project, struggles, as many similar outings struggle, with interpreting large and bold motion.

The other two rival frameworks visualized in the video, Time Reversal Fusion (TRF) and Generative Inbetweening (GI), provide a less skewed interpretation, but have created frenetic and even comic dance moves, neither of which respects the implicit logic of the two supplied frames.



CLICK TO PLAY VIDEO ON SITE

Click to play. *Two imperfect solutions to the tweening problem. Left, FILM treats the two frames as simple morph targets. Right, TRF knows that some form of dancing needs to be inserted, but comes up with an impracticable solution that demonstrates anatomical anomalies.*

Above-left, we can take a closer look at how FILM is approaching the problem. Though FILM was designed to be able to handle large motion, in contrast to prior approaches based on optical flow, it still lacks a semantic understanding of what should be happening between the two supplied keyframes, and simply performs a 1980/90s-style morph between the frames. FILM has no semantic architecture, such as a Latent Diffusion Model like Stable Diffusion, to aid in creating an appropriate bridge between the frames.

To the right, in the video above, we see TRF's effort, where Stable Video Diffusion (SVD) is used to more intelligently 'guess' how a dancing motion apposite to the two user-supplied frames might be – but it has made a bold and implausible approximation.

FCVG, seen below, makes a more credible job of guessing the movement and content between the two frames:

CLICK TO PLAY VIDEO ON SITE

Click to play. *FCVG improves upon former approaches, but is far from perfect.*

There are still artefacts, such as unwanted morphing of hands and facial identity, but this version is superficially the most plausible – and any improvement on the state of the art needs to be considered against the enormous difficulty that the task proposes; and the great obstacle that the challenge presents to the future of AI-generated video.

Why Interpolation Matters

As we have pointed out before, the ability to plausibly fill in video content between two user-supplied frames is one of the best ways to maintain temporal consistency in generative video, since two real and consecutive photos of the same person will naturally contain consistent elements such as clothing, hair and environment.

When only a *single* starting frame is used, the limited attention window of a generative system, which often only takes nearby frames into account, will tend to gradually 'evolve' facets of the subject matter, until (for instance) a man becomes another man (or a woman), or proves to have 'morphing' clothing – among many other distractions that are commonly generated in open source T2V systems, and in most of the paid solutions, such as Kling:

CLICK TO PLAY VIDEO ON SITE

Click to play. *Feeding the new paper's two (real) source frames into Kling, with the prompt 'A man dancing on a roof', did not result in an ideal solution. Though Kling 1.6 was available at the time of creation, V1.5 is the latest to support user-input start and end frames.* Source: https://klingai.com/

Is the Problem Already Solved?

By contrast, some commercial, closed-source and proprietary systems seem to be doing better with the problem – notably RunwayML, which was able to create very plausible inbetweening of the two source frames:

CLICK TO PLAY VIDEO ON SITE

Click to play. *RunwayML's diffusion-based interpolation is very effective.* Source: https://app.runwayml.com/

Repeating the exercise, RunwayML produced a second, equally credible result:

CLICK TO PLAY VIDEO ON SITE

Click to play. *The second run of the RunwayML sequence.*

One problem here is that we can learn nothing about the challenges involved, nor advance the open-source state of the art, from a proprietary system. We cannot know whether this superior rendering has been achieved by unique architectural approaches, by data (or data curation methods such as filtering and annotation), or any combination of these and other possible research innovations.

Secondly, smaller outfits, such as visual effects companies, cannot in the long term depend on B2B API-driven services that could potentially undermine their logistical planning with a single price hike – particularly if one service should come to dominate the market, and therefore be more disposed to increase prices.

When the Rights Are Wrong

Far more importantly, if a well-performing commercial model is trained on unlicensed data, as appears to be the case with RunwayML, any company using such services could risk downstream legal exposure.

Since laws (and some lawsuits) last longer than presidents, and since the crucial US market is among the most litigious in the world, the current trend towards greater legislative oversight for AI training data seems likely to survive the 'light touch' of Donald Trump's next presidential term.

Therefore the computer vision research sector will have to tackle this problem the hard way, in order that any emerging solutions might endure over the long term.

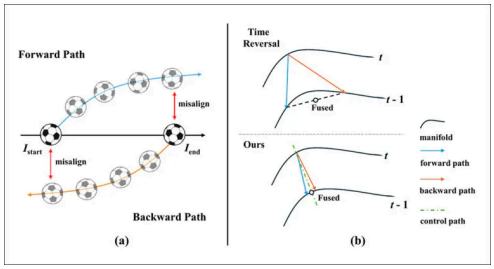
FCVG

The new method from China is presented in a paper titled *Generative Inbetweening through Frame-wise Conditions-Driven Video Generation*, and comes from five researchers across the Harbin Institute of Technology and Tianjin University.

FCVG solves the problem of ambiguity in the interpolation task by utilizing *frame-wise conditions*, together with a framework that delineates *edges* in the user-supplied start and end frames, which helps the process to keep a more consistent track of the transitions between individual frames, and also the overall effect.

Frame-wise conditioning involves breaking down the creation of interstitial frames into sub-tasks, instead of trying to fill in a very large semantic vacuum between two frames (and the longer the requested video output, the larger that semantic distance is).

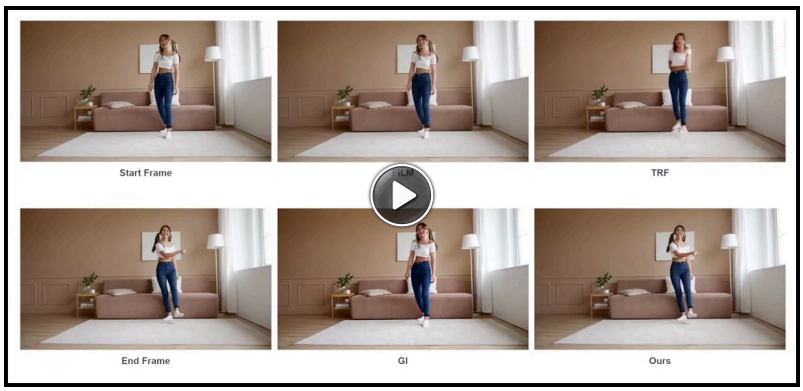
In the graphic below, from the paper, the authors compare the aforementioned time-reversal (TRF) method to theirs. TRF creates two video generation paths using a pre-trained image-to-video model (SVD). One is a 'forward' path conditioned on the start frame, and the other a 'backward' path conditioned on the end frame. Both paths start from the same random noise. This is illustrated to the left of the image below:



Comparison of prior approaches to FCVG. Source: https://arxiv.org/pdf/2412.11755

The authors assert that FCVG is an improvement over time-reversal methods because it reduces ambiguity in video generation, by giving each frame its own explicit condition, leading to more stable and consistent output.

Time-reversal methods such as TRF, the paper asserts, can lead to ambiguity, because the forward and backward generation paths can diverge, causing misalignment or inconsistencies. FCVG addresses this by using frame-wise conditions derived from matched lines between the start and end frames (lower-right in image above), which guide the generation process.



CLICK TO PLAY VIDEO ON SITE

Click to play. *Another comparison from the FCVG project page.*

Time reversal enables the use of pre-trained video generation models for inbetweening but has some drawbacks. The motion generated by I2V models is *diverse* rather than stable. While this is useful for pure image-to-video (I2V) tasks, it creates ambiguity, and leads to misaligned or inconsistent video paths.

Time reversal also requires laborious tuning of hyper-parameters, such as the frame rate for each generated video. Additionally, some of the techniques entailed in time reversal to reduce ambiguity significantly slow down inference, increasing processing times.

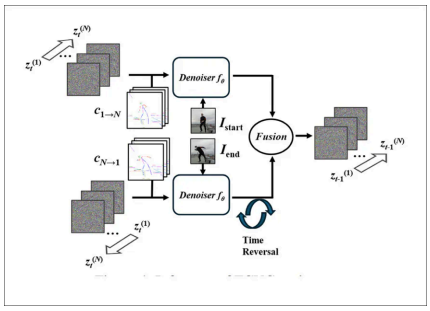
Method

The authors observe that if the first of these problems (diversity vs. stability) can be resolved, all other subsequent problems are likely to resolve themselves. This has been attempted in previous offerings such as the aforementioned GI, and also ViBiDSampler.

The paper states:

> *'Nevertheless [there] still exists considerable stochasticity between these paths, thereby constraining the effectiveness of these methods in handling scenarios involving large motions such as rapid changes in human poses. The ambiguity in the interpolation path primarily arises from insufficient conditions for intermediate frames, since two input images only provide conditions for start and end frames.*
>
> *'Therefore [we] suggest offering an explicit condition for each frame, which significantly alleviates the ambiguity of the interpolation path.'*

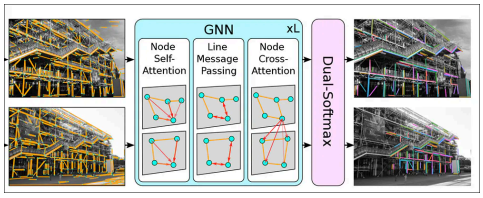
We can see the core concepts of FCVG at work in the schema below. FCVG generates a sequence of video frames that start and end consistently with two input frames. This ensures that frames are temporally stable by providing frame-specific conditions for the video generation process.



Schema for inference of FCVG.

In this rethinking of the time reversal approach, the method combines information from both forward and backward directions, blending them to create smooth transitions. Through an iterative process, the model gradually refines noisy inputs until the final set of inbetweening frames is produced.

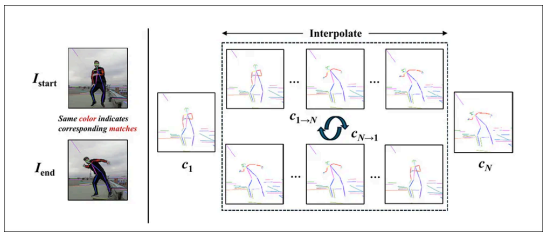
The next stage involves the use of the pretrained GlueStick line-matching model, which creates correspondences between the two calculated start and end frames, with the optional use of skeletal poses to guide the model, via the Stable Video Diffusion model.



GlueStick derives lines from interpreted shapes. These lines provide matching anchors between start and end frames in FCVG.*

The authors note:

> *'We empirically found that linear interpolation is sufficient for most cases to guarantee temporal stability in inbetweening videos, and our method allows users to specify non-linear interpolation paths for generating desired [videos].'*

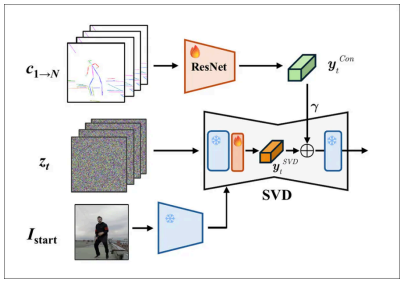


The workflow for establishing forward and backward frame-wise conditions. We can see the matched colors that are keeping the content consistent as the animation develops.

To inject the obtained frame-wise conditions into SVD, FCVG uses the method developed for the 2024 ControlNeXt initiative. In this process, the control conditions are initially encoded by multiple ResNet blocks, before cross-normalization between the condition and SVD branches of the workflow.

A small set of videos are used for fine-tuning the SVD model, with most of the model's parameters frozen.

> *'The [aforementioned limitations] have been largely resolved in FCVG: (i) By explicitly specifying the condition for each frame, the ambiguity between forward and backward paths is significantly alleviated; (ii) Only one tunable [parameter is introduced], while keeping hyperparameters in SVD as default, yields favorable results in most scenarios; (iii) A simple average fusion, without noise re-injection, is adequate in FCVG, and the inference steps can be substantially reduced by 50% compared to [GI].'*



Broad schema for injecting frame-wise conditions into Stable Video Diffusion for FCVG.

Data and Tests

To test the system, the researchers curated a dataset featuring diverse scenes including outdoor environments, human poses, and interior locations, including motions such as camera movement, dance actions, and facial expressions, among others. The 524 clips chosen were taken from the DAVIS and RealEstate10k datasets. This collection was supplemented with high frame-rate videos obtained from Pexels. The curated set was split 4:1 between fine-tuning and testing.

Metrics used were Learned Perceptual Similarity Metrics (LPIPS); Fréchet Inception Distance (FID); Fréchet Video Distance (FVD); VBench; and Fréchet Video Motion Distance.

The authors note that none of these metrics is well-adapted to estimate temporal stability, and refer us to the videos on FCVG's project page.

In addition to the use of GlueStick for line-matching, DWPose was used for estimating human poses.

Fine-tuning tool place for 70,000 iterations under the AdamW optimizer on a NVIDIA A800 GPU, at a learning rate of 1×10^{-6}, with frames cropped to 512×320 patches.

Rival prior frameworks tested were FILM, GI, TRF, and DynamiCrafter.

For quantitative evaluation, frame gaps tackled ranged between 12 and 23.

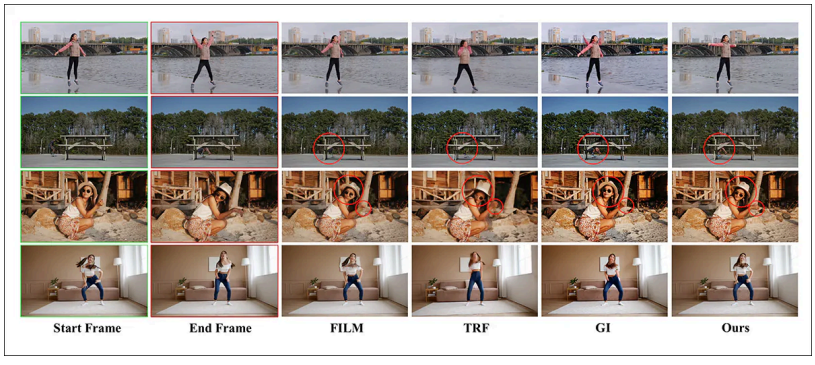
Method	Frame Gap = 23					Frame Gap = 12				
	LPIPS (↓)	FID (↓)	VBench (↑)	FVMD (↓)	FVD (↓)	LPIPS (↓)	FID (↓)	VBench (↑)	FVMD (↓)	FVD (↓)
FILM[32]	**0.1540**	25.00	0.8615	8208.7	543.4	**0.1980**	24.44	0.8667	6975.9	495.4
DynamiCrafter[49]	0.3886	52.66	0.8410	13221.9	978.9	0.3839	37.49	0.8458	11810.7	652.5
TRF[7]	0.3687	42.76	0.8438	10458.0	823.4	0.3742	39.01	0.8478	10076.6	818.4
GI[44]	0.2155	31.39	0.8606	5682.6	524.0	0.2615	32.37	0.8651	4721.0	565.8
Ours	0.1832	**24.05**	**0.8619**	**5607.2**	**437.9**	0.2378	**22.77**	**0.8672**	**4537.4**	**465.6**

Quantitative results against prior frameworks.

Regarding these results, the paper observes:

> *'[Our] method achieves the best performance among four generative approaches across all the metrics. Regarding the LPIPS comparison with FILM, our FCVG is marginally inferior, while demonstrating superior performance in other metrics. Considering the absence of temporal information in LPIPS, it may be more appropriate to prioritize other metrics and visual observation.*
>
> *'Moreover, by comparing the results under different frame gaps, FILM may work well when the gap is small, while generative methods are more suitable for large gap. Among these generative methods, our FCVG exhibits significant superiority owing to its explicit frame-wise conditions.'*

For qualitative testing, the authors produced the videos seen at the project page (some embedded in this article), and static and animated† results in the PDF paper,

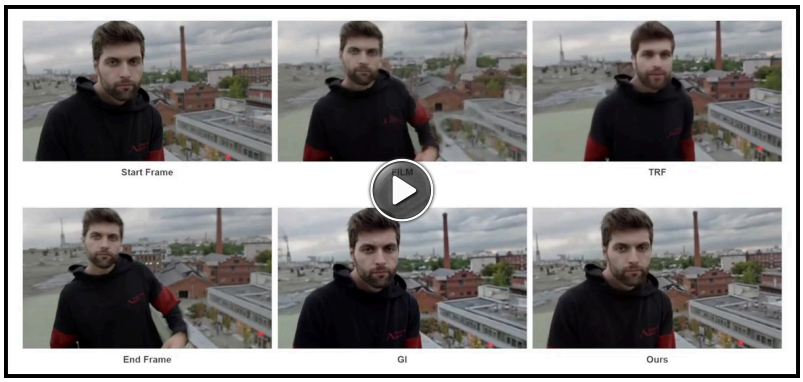


Sample static results from the paper. Please refer to source PDF for better resolution, and be aware that the PDF contains animations which can be played in appl

The authors comment:

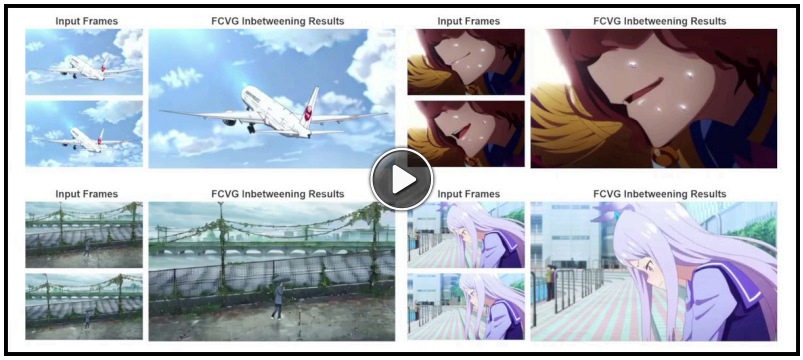
> *'While FILM produces smooth interpolation results for small motion scenarios, it struggles with large scale motion due to inherent limitations of optical flow, resulting in noticeable artifacts such as background and hand movement (in the first case).*
>
> *'Generative models like TRF and GI suffer from ambiguities in fusion paths leading to unstable intermediate motion, particularly evident in complex scenes involving human and object motion.*

'In contrast, our method consistently delivers satisfactory results across various scenarios.'Even when significant occlusion is present (in the second case and sixth case), our method can still capture reasonable motion. Furthermore, our approach exhibits robustness for complex human actions (in the last case).'



CLICK TO PLAY VIDEO ON SITE

The authors also found that FCVG generalizes unusually well to animation-style videos:



CLICK TO PLAY VIDEO ON SITE

Click to play. *FCVG produces very convincing results for cartoon-style animation.*

Conclusion

FCVG represents at least an incremental improvement for the state-of-the-art in frame interpolation in a non-proprietary context. The authors have made the code for the work available on GitHub, though the associated dataset has not been released at the time of writing.

If proprietary commercial solutions are exceeding open-source efforts through the use of web-scraped, unlicensed data, there seems to be limited or no future in such an approach, at least for commercial use; the risks are simply too great.

Therefore, even if the open-source scene lags behind the impressive showcase of the current market leaders, it is, arguably, the tortoise that may beat the hare to the finish line.

* *Source:*
https://openaccess.thecvf.com/content/ICCV2023/papers/Pautrat_GlueStick_Robust_Image_Matching_by_Sticking_Points_and_Lines_Together_ICCV_2023_pap

† *Requires Acrobat Reader, Okular, or any other PDF reader that can reproduce embedded PDF animations.*

First published Friday, December 20, 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