CS410 Text Information System

Technology Review

NetId: [rbal2]

Name: Rathiv Bal

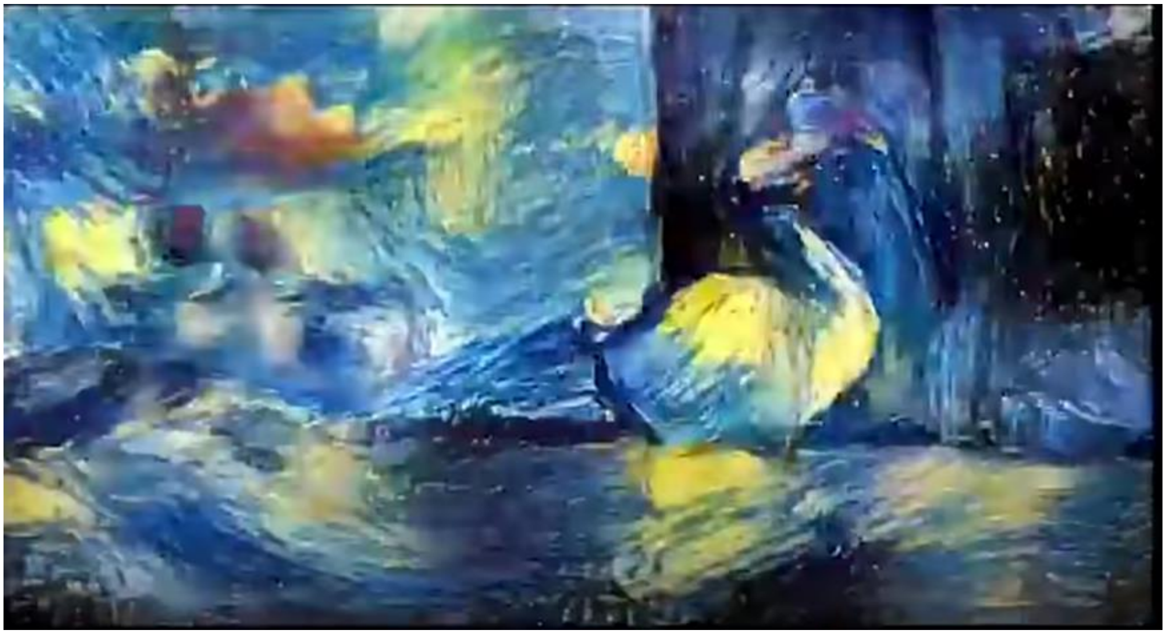
Topic:

This is the review of advancement in “text to video” AI models.

Technology review-

We know that modern AI programs can paint images for us, anything we wish, and these programs include DALL-E , mid journey and stable diffusion in oct 2022, google released new version of imagen that also can also create video just with text input.Videos require a much greater understanding of the world around us, so much more computation, and temporal coherence. Since video is not just a set of images, but a series of images that have

to relate to each other. If the AI does not do a good job at this, we get this.



Flickering, is the side effect of independent per frame processing with no temporal coherence by previous models like stable diffusion, DALL-E and mid journey since all these models generate images and then combine them to form video. But imagen also look for temporal coherence while generating each frame of video to create this

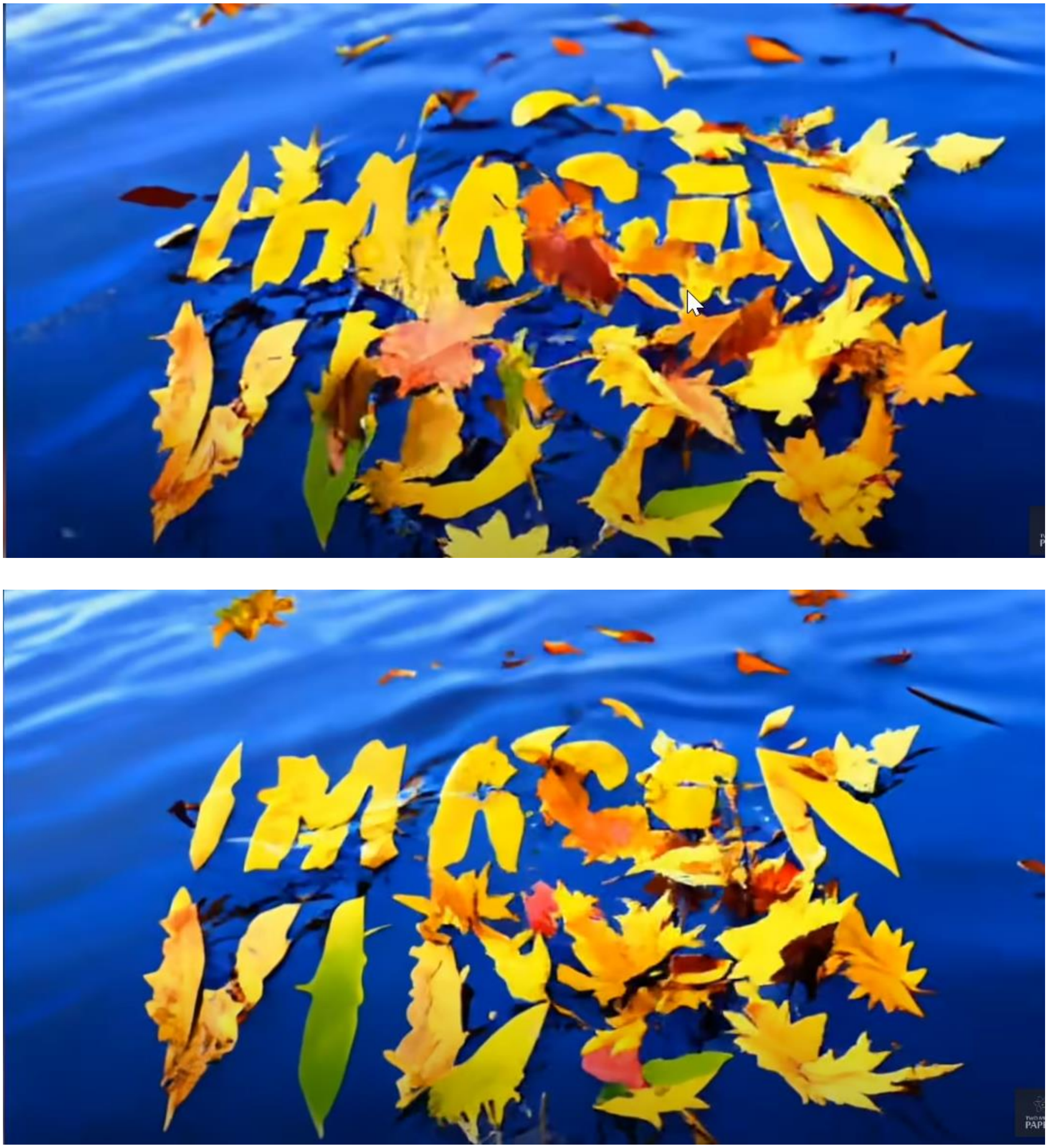
A happy elephant wearing a birthday hat walking under the se

This frame from AI generated video using imagen tool has no flickering.

All of this is so hard, researchers were thinking that they may be able to do this in 5-10 years, or maybe never? Well, scientists at Google did it in few months after text-to-image models became successful.

To set context lets look at 3 of my favorite examples, and then I'll tell you how much time this took. By the way, it is an almost unfathomably short time. Now, one, the concept is the same: one simple text prompt goes in, for instance, a happy elephant wearing a birthday hat walking under the sea, and this comes out. Wow. Look at that! That is exactly what we were asking for in the prompt, plus, look at the waves and the sky through the sea, which is absolutely beautiful, but it doesn't stop there - I also see every light transport researcher's dream there. Water caustics. And gorgeous and realistic patterns in the water. Now, not even this technique is perfect, you see that temporal coherence is still subject to improvement, the video still flickers a tiny bit, the tusk is also changing over time. However, this is incredible progress in so little time. Absolutely amazing.

Now let's ask for a bit of physics, a bunch of autumn leaves falling on a calm lake forming the text "Imagen Video".



Shots from two close by frames in the video

In computer graphics, creating a simulation like this would take quite a bit of 3D modeling knowledge, then, we also have to fire up a fluid simulation. Now, this does not seem to do a great deal of two-way coupling, which means that the water has
an effect on the leaves, you see it advecting leaf in few places, but the leaves do not seem to have a huge effect on the water itself. This is possible with specialized computer graphics algorithms, and I bet it will also be possible with Imagen Video 2.

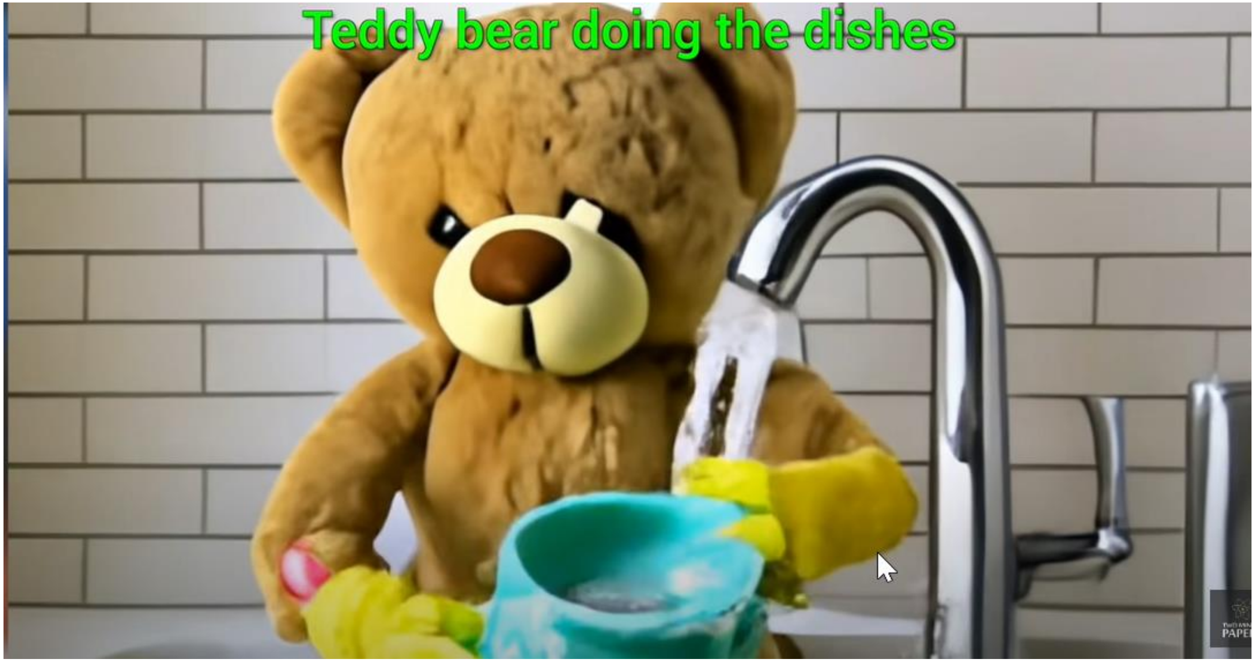
But reflections of the leaves appearing on the water is admirable in the video generated by imagen.

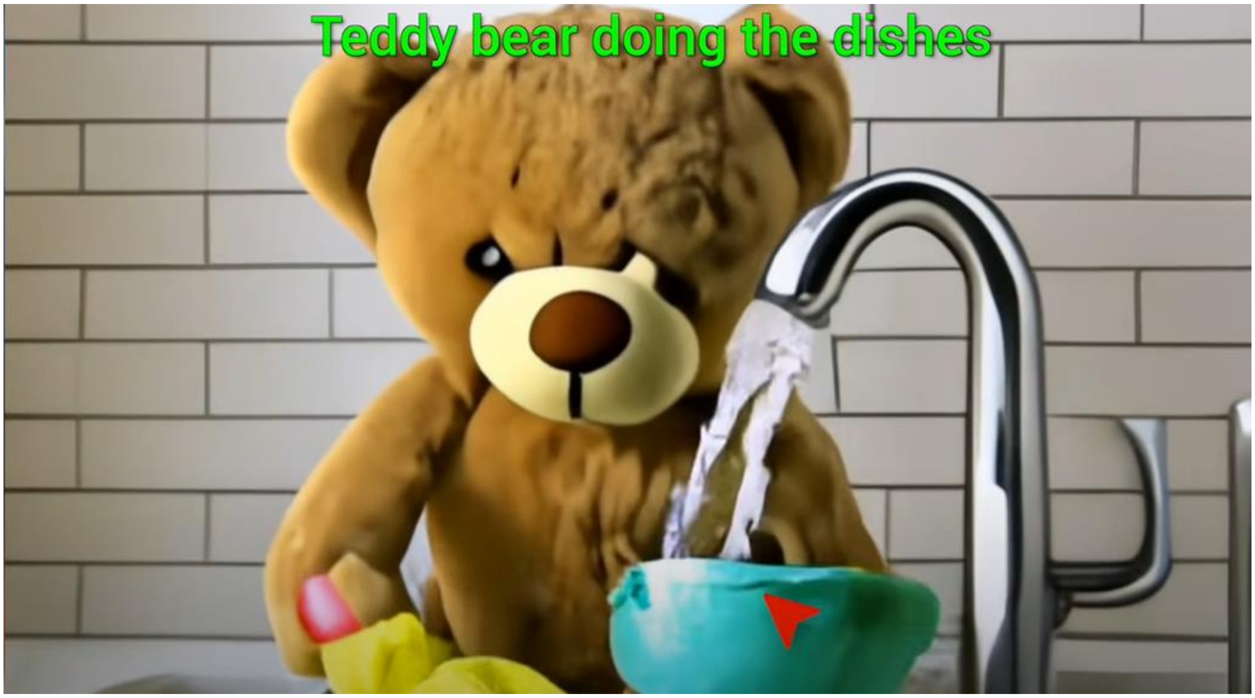
And to think that this is just the first iteration of Imagen Video, it is really a great achievement by AI model – imagen.

However admittedly, the detailed real physics simulation can can be almost as detailed as real life but then it takes long time to create such simulations.

However google imagen model version 2 , released in oct 2022, can create many real physics simulation effects like splash on water etc. And turquoise liquid’s movement in the glass too. Great simulations on version 1. I am so happy!

Now, three, give me a teddy bear doing the dishes. Whoa! Is this real?





Two consecutive frames of bear washing dishes video.

Conclusion:
Google's new model, Imagen released in oct 2022, generates coherent very realistic video just with text inputs. It really feels like we are living inside a science fiction movie. Now, it's not perfect, model is little confused by the interaction of may objects, but even few weeks ago it seemed impossible that AI would be able to videos that are so coherent and realistic. The new AI model don't just have a really good understanding of reality, but it can also combine two previous concepts, a teddy bear and washing the dishes into something new.

We noted that this is incredible progress in so little time. But, how little exactly? Well, OpenAI's DALL-E 2 text to image AI appeared in April 2022, then, Google's Imagen, also text to image appears one month later, May 2022, that is incredible, and now, only 5 months later, by October 2022, we get this, An amazing text to video AI.

Of course, it is not perfect, the hair of pets is typically still a problem, and the complexity of this ship battle is still a little too much for it to shoulder, so version one is not going to make a new Pirates of The Caribbean Movie, but maybe version 3 two more papers down the line it might be possible. The resolution of these videos is not too bad at all, it is in 720p, the literature likes to call it high definition. These are not in 4k like the shows you can watch on your tv, but this quality for a first crack the at the problem is simply stunning. And don't forget that first, it synthesizes a low-resolution video, then upscales it through super resolution, something Google is already really good at, so I would not be surprised for version 2

to easily go to full HD, and maybe even beyond. As you see, the pace of progress in AI research is nothing short of amazing. If like me, you are yearning for some more results, you can check out the paper's website
https://imagen.research.google/video/
you get a random selection of results. Refresh it a couple times and see if you get something new! And if I could somehow get access to this technique, you bet that I'd be generating a ton more of these.