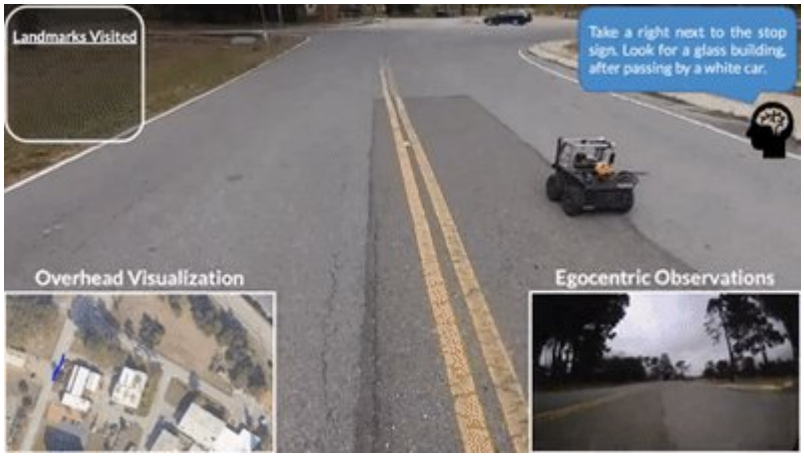
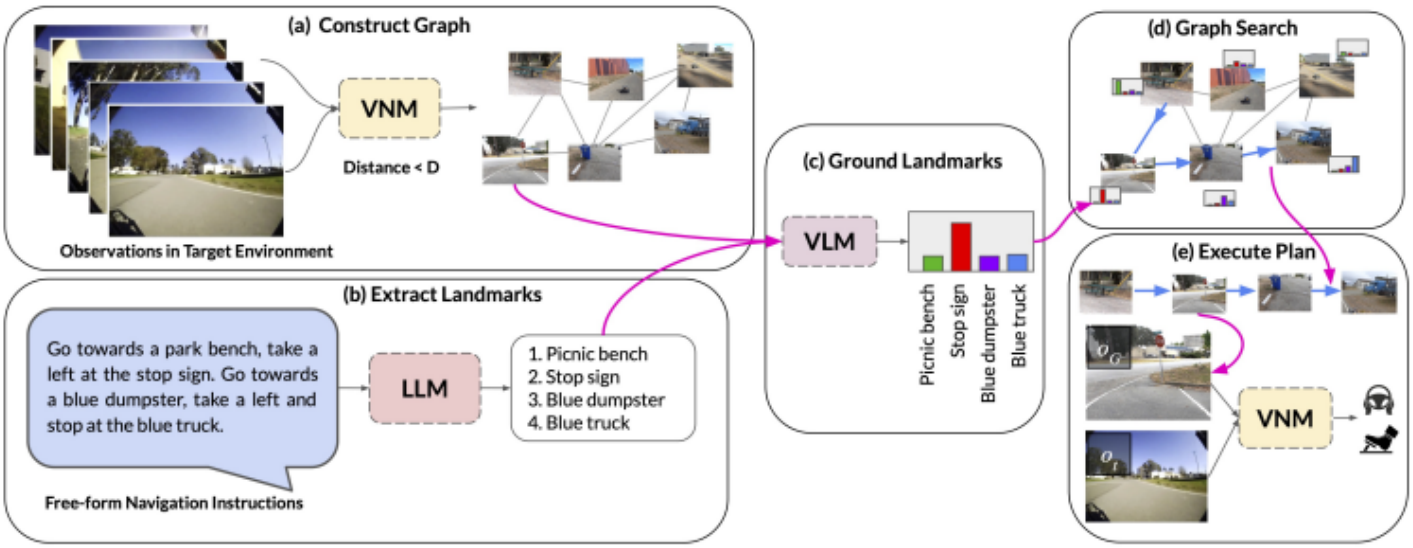
Twitter Thread by Sergey Levine

Sergey Levine
@svlevine

Can we get robots to follow language directions without any data that has both nav trajectories and language? In LM-Nav, we use large pretrained language models, language-vision models, and (non-lang) navigation models to enable this in zero shot!
https://t.co/EVsFOj1JhS
Thread:



LM-Nav first uses a pretrained language model (LLM) to extract navigational landmarks from the directions. It uses a large pretrained navigation model (VNM) to build a graph from previously seen landmarks, describing what can be reached from what. Then...



It uses a vision-language model (VLM, CLIP in our case) to figure out which landmarks extracted from the directions by the LLM correspond to which images in the graph, and then queries the VNM to determine the robot controls to navigate to these landmarks.

The results in fully end-to-end image-based autonomous navigation directly from user language instructions. All components are large pretrained models, without hand-engineered localization or mapping systems. See example paths below.

LM-Nav uses the ViKiNG Vision-Navigation Model (VNM): https://t.co/oR1bpIlvMt
This enables image-based navigation from raw images.

w/ @shahdhruv_, @blazejosinski, @brian_ichter
Paper: https://t.co/ymU9kALQ9G
Web: https://t.co/EVsFOj1JhS
Video: https://t.co/QRhQDAZydM

If you want to play with LM-Nav in the browser, check out the colab demo here: https://t.co/GKrOr24hU6

Of course, our colab won't make a robot materialize and drive around, but you can play with the graph and the LLM+LVM components ■