D³Nav: Data-Driven Driving Agents for Autonomous Vehicles in Unstructured Traffic

BMVC 2024

adityang.github.io/D3Nav

Aditya NG, Gowri Srinivasa
adityang5@gmail.com, gsrinivasa@pes.edu

Goals

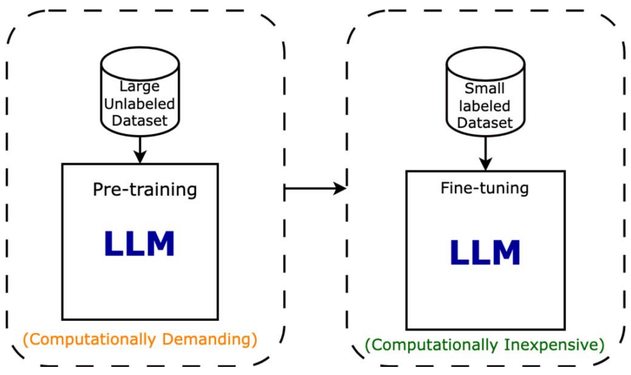
The AV 1.0 stack focuses on modularity and abstraction to build many small models. The AV 2.0 stack focuses more on end to end learning.

We set out with the goal of making use of raw video data to train a model focused on the end to end learning approach and to build a **Data Driven Driving agent** and an end-to-end simulator for autonomy.

Motivations

How do we build a robust world model?

LLMs are pre-trained on large scale text data with the task of "predicting the next word" and then are fine tuned on a task like Chat [1]. The idea is that the next-token-prediction task provides excellent supervision for building a robust world model given sufficient data.



With a similar intuition [2], we pre-train our driving model on the task of predicting the next video frame. We then fine tune the model on producing a driving signal.

Results: Metrics

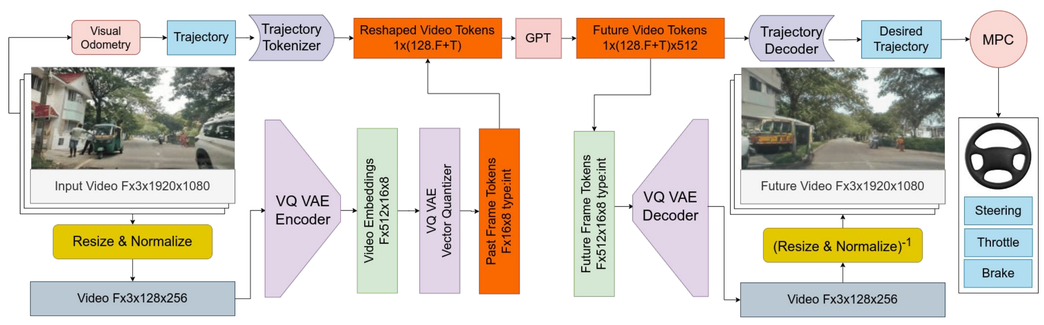
Size	Hyperparameters					Metrics				
	L	D_E	D_R	D_A	LR	$F1$	$Prec$	DTW	CE	FPS
XS	6	0.200	0.300	0.0003	0.00003	0.318	0.318	27.2	3.159	35.021
S	12	0.500	0.500	0.0003	0.0003	0.370	0.370	26.1	2.766	32.683
M	24	0.100	0.500	0.0003	0.0002	0.395	0.395	24.6	2.522	24.570
L	36	0.200	0.200	0.100	0.0003	0.458	0.458	23.2	2.277	20.276
XL	48	0.500	0.300	0.300	0.003	0.462	0.462	18.5	2.230	17.702

Table 1: **Quantitative Results** on our proposed architecture comparing the optimal hyper-parameters and metrics achieved. The table shows the hyper-parameters Number of Layers L, Embeddings Dropout D_E, Residual Dropout D_R, Attention Dropout D_A, and Learning Rate LR. We have evaluated on the metrics $F1$, Precision $Prec$, Dynamic Time Warping Distance [11] DTW, Cross Entropy CE and Frame Rate FPS.

Model	Hyperparameters			Metrics					
	LR	β	BS	$RMSE$	M. $RMSE$	a_1	a_2	a_3	$Comp.$
$V_{16\times8}$	0.0003	0.1	32	0.3920	0.3920	0.7952	0.9549	0.9514	256x
$V_{32\times16}$	0.00003	0.25	32	0.3649	0.3649	0.7979	0.9637	0.9590	64x
$V^{IM}_{16\times8}$	0.0003	0.1	32	0.5396	0.3945	0.7231	0.8892	0.9051	256x
$V^{IM}_{32\times16}$	0.00003	0.25	32	0.4982	0.3587	0.7418	0.8979	0.9134	64x

Table 2: Our Encoder-Decoder pair was trained on our video datasets to learn an efficient embedding space. We optimize for Learning Rate LR, Beta β, Batch Size BS. Beta decides the weight given to the commitment loss[12]. kWe evaluate on the metrics $RMSE$, Masked $RMSE$, $a1$, $a2$, $a3$ and $Compression$. a_i is the fraction of predictions where the threshold maximum between gt/pred or pred/gt is less than 1.25^i. Models with superscript IM were trained with importance masking. We use the $V^{IM}_{16\times8}$ as our primary encoder.

Architecture



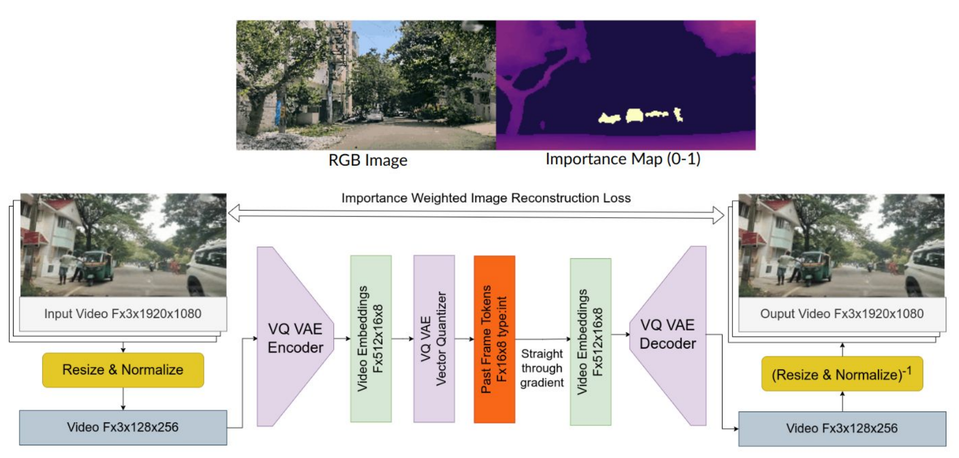
D³Nav Architecture Diagram

Training

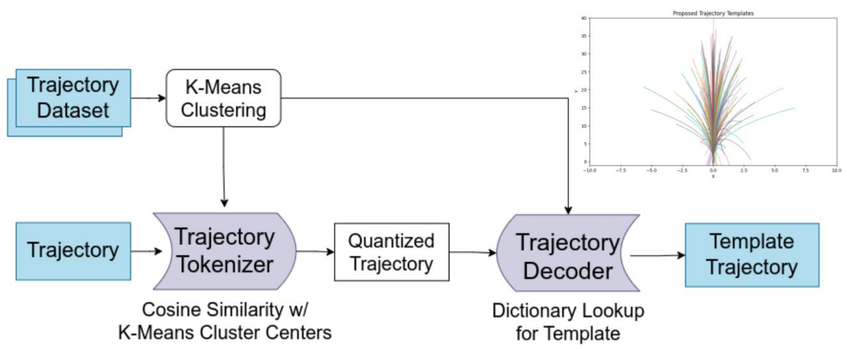
Quantizer Training. We initialize the image and trajectory quantizers by training them on reconstruction loss.

Pre-Training. We first pretrain our GPT world model on the next video frame token prediction task

Fine Tuning. We fine tune our world model on the task of driving the car



VQ-VAE [3] Training with Importance Maps



Trajectory Quantizer

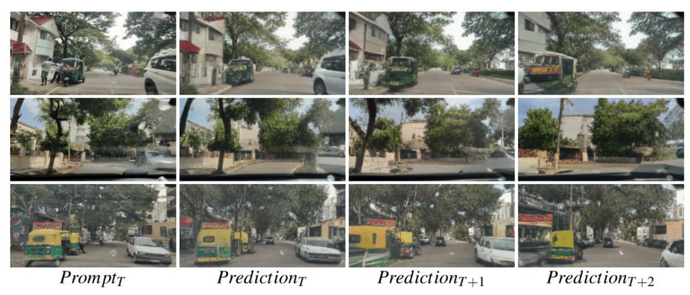
Auto-Labeling Pipeline

Driving Signal Labeling. We make use of Visual-Odometry, GPS, and IMU to extract the traversed ego-trajectory and speed from the video data.

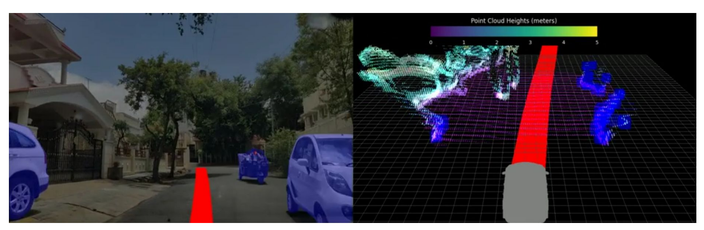
Depth Boosting. Taking inspiration from the depth boosting techniques [4, 5, 6], we merge the depth maps

Semantic Segmentation auto-labeling. A coarse semantic segmentation map is iteratively refined at uncertain regions, typically object boundaries, to produce high-resolution 2D semantic labels that can be projected into 3D using depth information.

Results: Model Outputs



Sampling video frames



D³Nav deciding control signals

Conclusion

By leveraging quantized encodings and automated importance labeling, it efficiently processes high-dimensional data while focusing on critical elements like pedestrians and vehicles. The system's ability to predict future video frames and control signals with minimal human intervention streamlines the training process, making it particularly valuable for developing driving agents and simulations where real-world data is limited. The resulting dataset of embeddings and trajectory labels offers a rich resource for further research in autonomous driving technology.

References

[1] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018.

[2] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers, 2021.

[3] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. NIPS'17

[4] Aditya N Ganesh, Dhruval Pobbathi Badrinath, Harshith Mohan Kumar, Priya S, and Surabhi Narayan. Octran: 3d occupancy convolutional transformer network in unstructured traffic scenarios. Spotlight Presentation at the Transformers for Vision Workshop, CVPR, 2023

[5] S. Mahdi H. Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yagız Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. CVPR 2021

[6] Seyed Mahdi Hosseini Miangoleh. Boosting monocular depth estimation to high resolution. Master's thesis, Simon Fraser University, 2022.