Karlsruhe Institute of Technology

Conference on Robot Learning '24

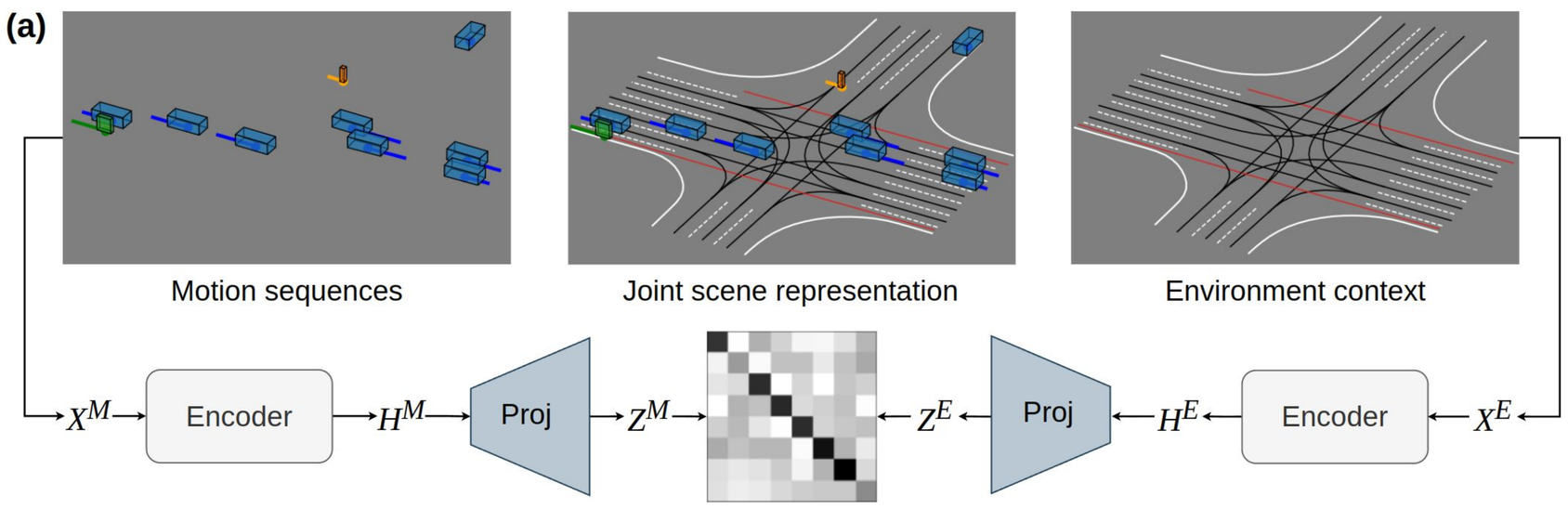
JointMotion: Joint Self-Supervision for Joint Motion Prediction

Authors: Royden Wagner[1,*], Ömer Şahin Taş[2,*], Marvin Klemp[1], Carlos Fernandez[1]

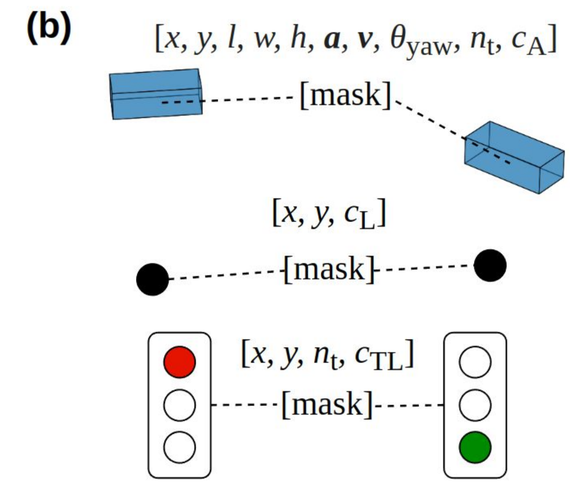
[1]Karlsruhe Institute of Technology [2]FZI Research Center for Information Technology *Joint first authors

Main contributions

We propose two self-supervised pre-training objectives for joint motion prediction in self-driving vehicles:



Connecting motion and environments (CME): Our scene-level objective learns joint scene representations via non-contrastive similarity learning of motion sequences and environment context.

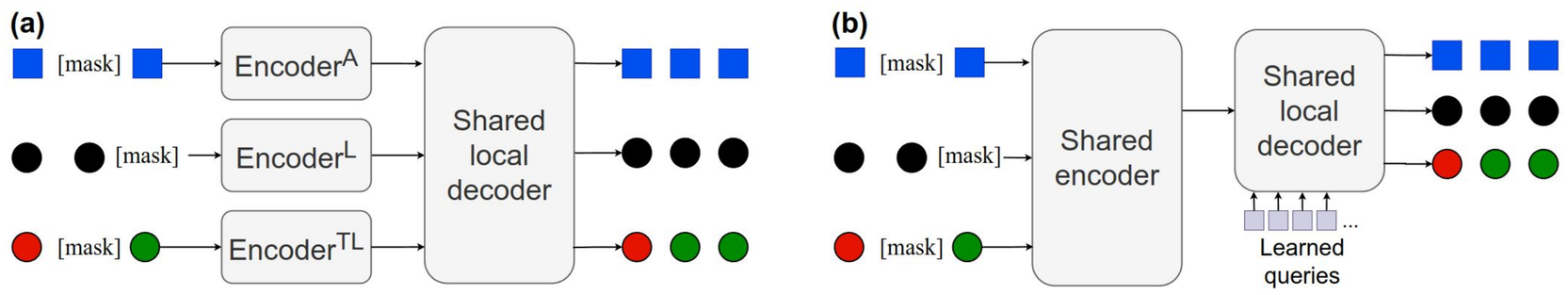


Masked polyline modeling (MPM): Our instance-level objective refines learned representations via masked autoencoding of multimodal polylines.

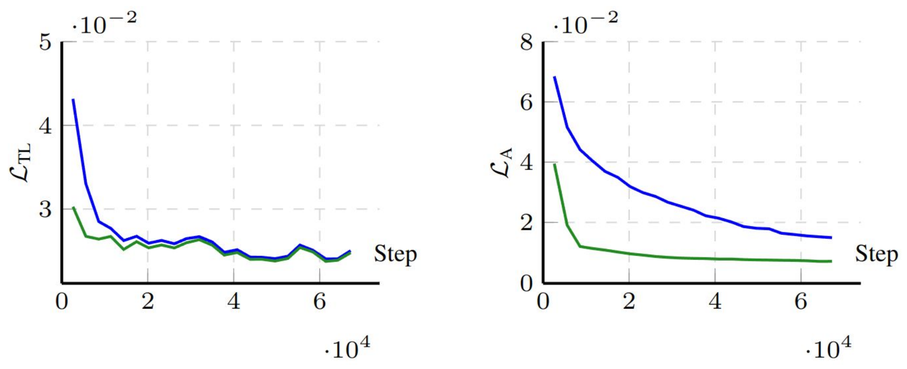
Adaptive pre-training decoder:

(a) Late fusion with modality-specific encoders for agents, lanes, traffic lights.

(b) Early fusion with a shared encoder for all modalities. Compressed features are decoded using learned query tokens.



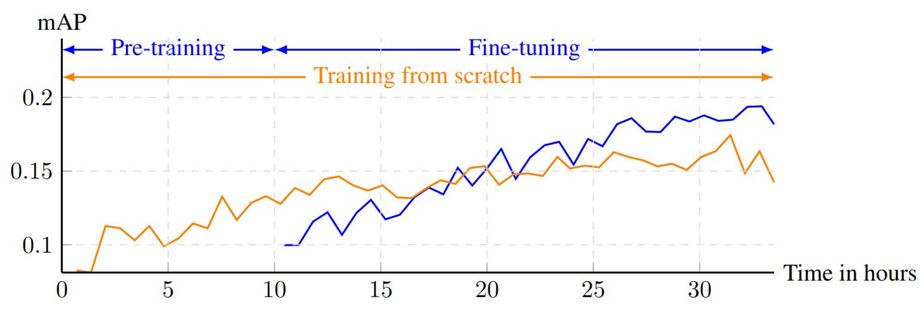
Evaluation on large-scale motion prediction datasets



Loss plots of our complementary pre-training objectives. Blue: JointMotion. Green: JointMotion without CME.

Dataset	Model (config)	Pre-training	minFDE↓		minADE↓		MR↓		OR↓	
WOMD	Scene Transformer	None	3.6715		1.5255		0.7372		0.2868	
		PreTraM [15]	3.6508	-0.57%	1.5415	1.05%	0.7385	0.18%	0.2915	1.64%
		JointMotion	**3.2400**	-11.75%	**1.3830**	-9.36%	**0.7090**	-3.82%	**0.2847**	-0.73%
	HPTR	None	2.6003		1.1682		0.6030		0.2331	
		PreTraM [15]	2.5049	-3.66%	1.0981	-5.99%	0.5863	-2.78%	0.2345	0.60%
		JointMotion	**2.4006**	-7.68%	**1.0564**	-9.58%	**0.5591**	-7.28%	**0.2297**	-1.46%
	Wayformer (joint)	None	2.3529		1.0209		0.5461		0.2273	
		JointMotion	**2.2823**	-3.00%	**0.9939**	-2.64%	**0.5270**	-3.50%	**0.2143**	-5.72%
AV2	HPTR	None	2.2550		1.1380		-		**0.0988**	
		JointMotion WOMD	**2.1530**	-4.53%	**1.1370**	-0.09%	-		0.1025	3.75%

Comparing scene-level self-supervision methods. Metrics for the Waymo Open Motion Dataset (WOMD) interactive and the Argoverse 2 Forecasting (AV2) validation split.



Accelerating and improving training via self-supervised pre-training. Results for Scene Transformer.

Split	Method (config)	Venue	mAP↑	minADE↓	minFDE↓	MR↓	OR↓
Test	Scene Transformer (joint) [20]	ICLR'22	0.1192	0.9774	2.1892	0.4942	0.2067
	GameFormer (joint) [40]	ICCV'23	0.1376	0.9161	**1.9373**	0.4531	0.2112
	MotionDiffuser [24]	CVPR'23	**0.1952**	**0.8642**	1.9482	**0.4300**	**0.2004**
	JointMotion (HPTR)		0.1869	0.9129	2.0507	0.4763	0.2037
Val	GameFormer (joint) [40]	ICCV'23	0.1339	**0.9133**	**1.9251**	**0.4564**	-
	MotionLM (single replica) [23]	ICCV'23	0.1687	1.0345	2.3886	0.4943	-
	JointMotion (HPTR)		**0.1761**	0.9689	2.2031	0.4915	**0.1990**
	MotionLM (ensemble)	ICCV'23	0.2150	0.8831	1.9825	0.4092	-

Comparison with state-of-the-art methods for joint motion prediction. Test metrics are from the leaderboard of the Waymo Open Interaction Prediction Challenge '21.