OpenDriveLab

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

What are Good (Pre-training) Representations for Robotic Manipulation?

With a few autonomous driving work first ...

Hongyang Li

Research Scientist, Shanghai AI Lab

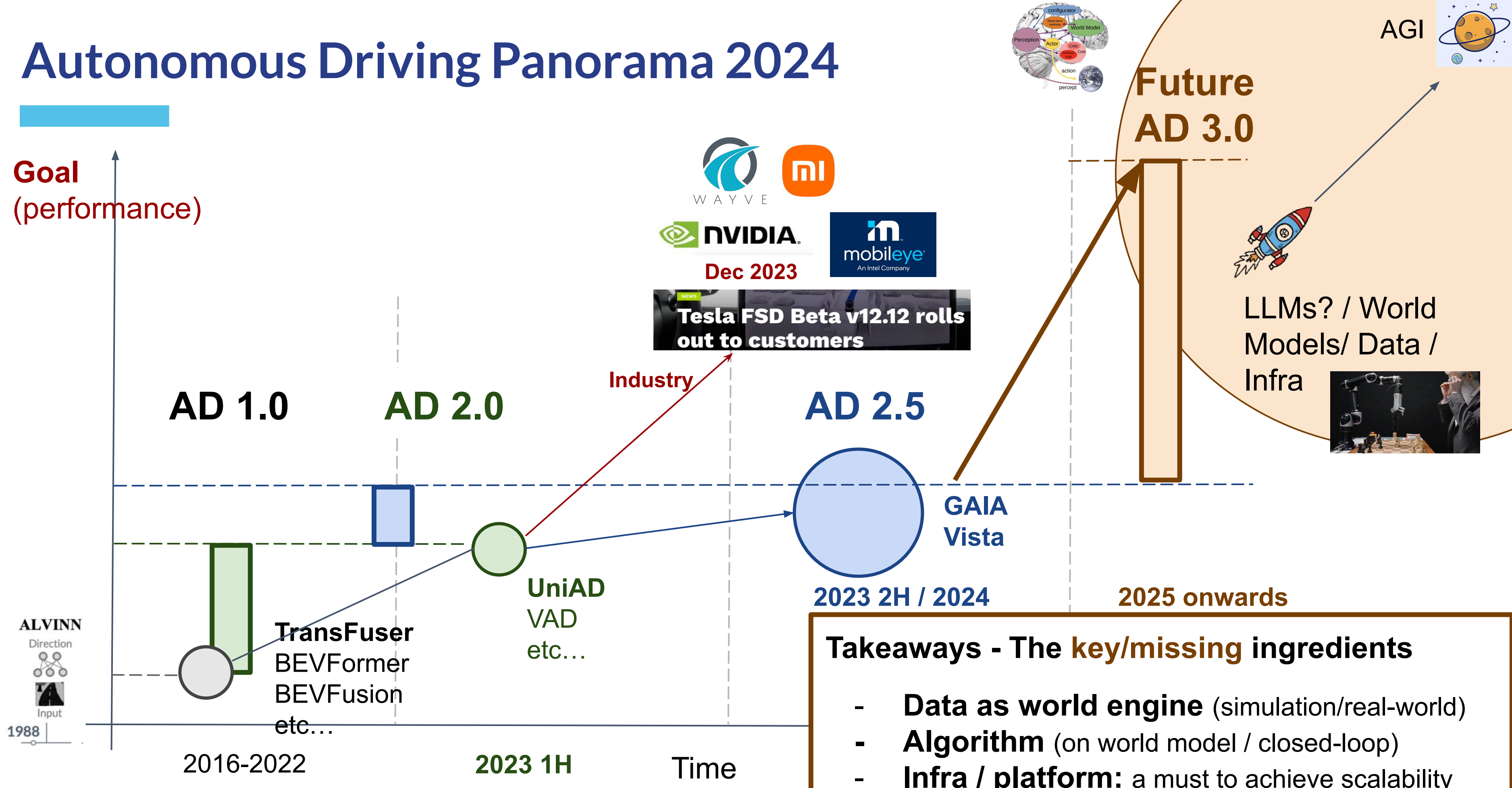
@ TU Delft, July 18 2024

Outline

- Autonomous Driving
- Introduction to Robotic Manipulation
- MPI and ClOVER
- Concluding Remarks

Autonomous Driving

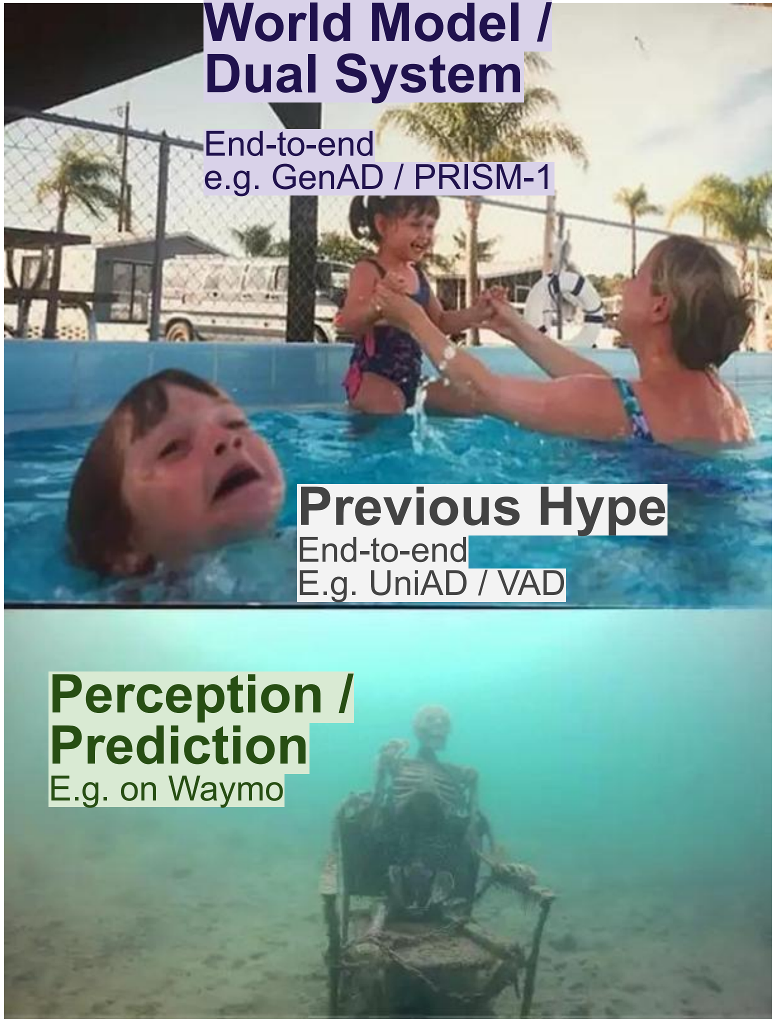
Autonomous Driving Panorama 2024



Takeaways - The key/missing ingredients

- **Data as world engine** (simulation/real-world)
- **Algorithm** (on world model / closed-loop)
- **Infra / platform:** a must to achieve scalability
- **Deployment**: dual system / onboard-chip

Any pizza left?



Autonomous Driving
Research

Academia

Q: I don't have too much resource. Is there any impact research ***left*** for autonomous driving **research** at University?

A: My guess is **Maybe.**
Open-set Perception / Motion Prediction / etc.

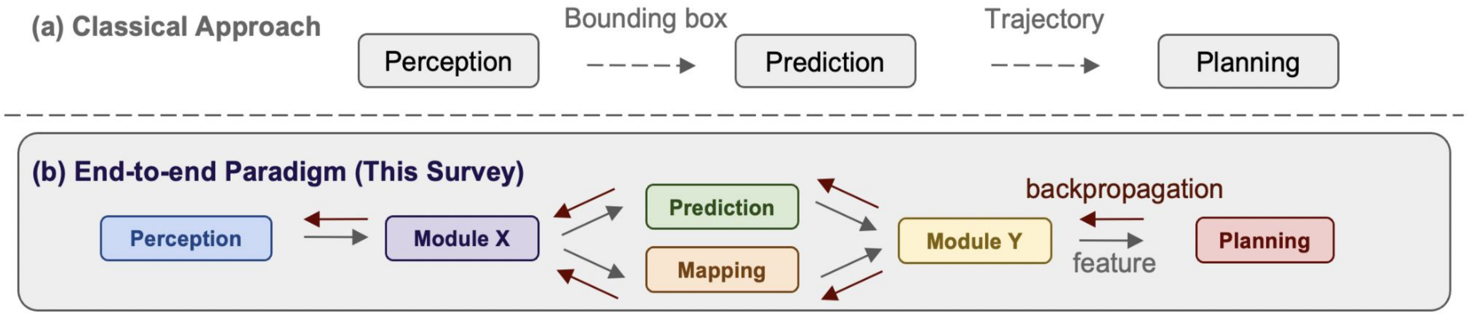
Industry

Q: We have ***lots of*** GPUs, millions of driving data. ***Could*** we leave Tesla no pizza, and achieve L4 once for all using e2e?

A: **Definitely not.**
Academia/industry should collaborate in whatsoever close manners.

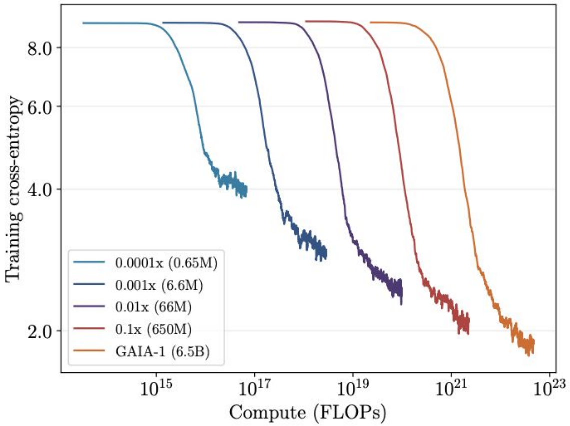
End-to-end Autonomy | True Incentive

+ **Global optimization:** when perception fails/inferior, planning still could work.



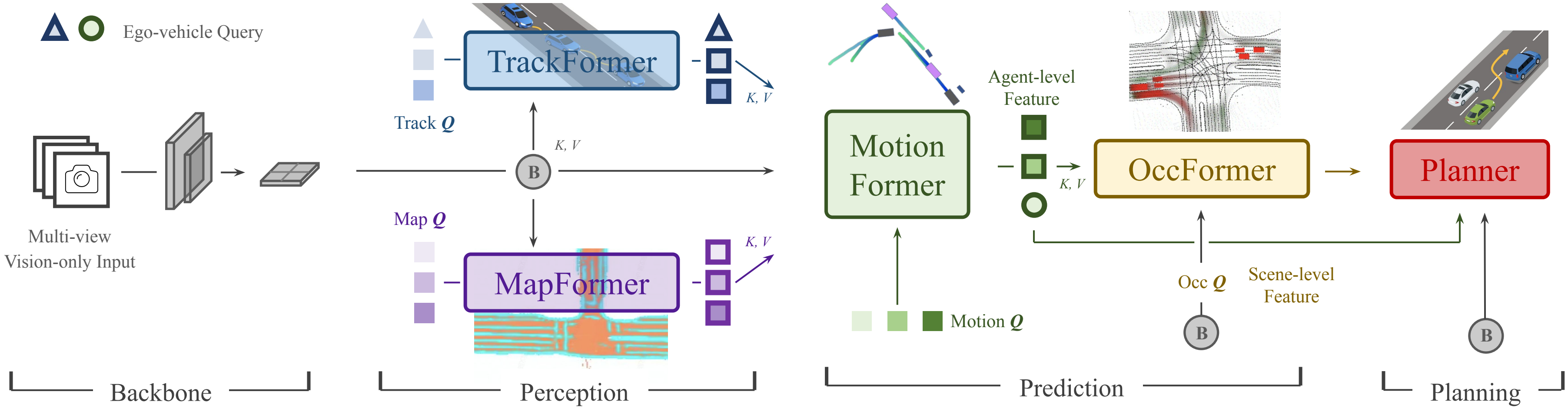
Chen et al., End-to-end Autonomous Driving:
Challenges and Frontiers, https://arxiv.org/abs/2306.16927

+ **Scaling law:** massive amount of data + infra/compute —> **strong generalization**



Hu et al. GAIA-1: A Generative World Model for Autonomous Driving https://arxiv.org/abs/2309.17080

Classic algorithm: UniAD



- Entire pipeline connected by queries
- Tasks coordinated with queries

Unified Query

- Interactions modeled by attention

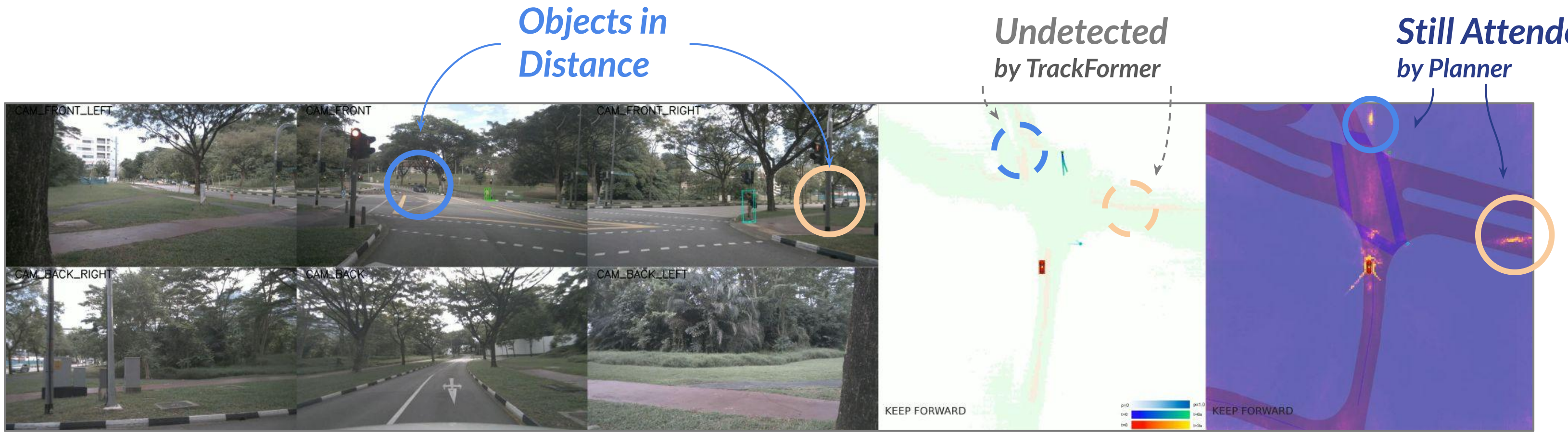
Transformer-based

First time to unify full-stack AD tasks!

Hu et al. Planning-oriented Autonomous Driving. CVPR 2023.

UniAD - Recover from Upstream Errors

Planner could still attend to ‘undetected’ regions/objects



Hu et al. Planning-oriented Autonomous Driving. CVPR 2023.

Core in UniAD: Planning-oriented, not a MTL framework.

Tasks benefit🚀each other and contribute to safe planning

ID	Modules					Tracking			Mapping		Motion Forecasting			Occupancy Prediction				Planning	
	Track	Map	Motion	Occ.	Plan	AMOTA↑	AMOTP↓	IDS↓	IoU-lane↑	IoU-road↑	minADE↓	minFDE↓	MR↓	IoU-n.↑	IoU-f.↑	VPQ-n.↑	VPQ-f.↑	avg.L2↓	avg.Col.↓
0*	✓	✓	✓	✓	✓	0.356	1.328	893	0.302	0.675	0.858	1.270	0.186	55.9	34.6	47.8	26.4	1.154	0.941
1	✓					0.348	1.333	791	-	-	-	-	-	-	-	-	-	-	-
2		✓				-	-	-	**0.305**	0.674	-	-	-	-	-	-	-	-	-
3	✓	✓				0.355	1.336	785	0.301	0.671	-	-	-	-	-	-	-	-	-
4			✓			-	-	-	-	-	0.815	1.224	0.182	-	-	-	-	-	-
5	✓		✓			0.360	1.350	919	-	-	0.751	1.109	0.162	-	-	-	-	-	-
6	✓	✓	✓			0.354	1.339	820	0.303	0.672	0.736(-9.7%)	1.066(-12.9%)	0.158	-	-	-	-	-	-
7				✓		-	-	-	-	-	-	-	-	60.5	37.0	52.4	29.8	-	-
8	✓			✓		0.360	**1.322**	809	-	-	-	-	-	62.1	38.4	52.2	32.1	-	-
9	✓	✓	✓	✓		0.359	1.359	1057	0.304	**0.675**	**0.710**(-3.5%)	**1.005**(-5.8%)	**0.146**	**62.3**	39.4	**53.1**	32.2	-	-
10					✓	-	-	-	-	-	-	-	-	-	-	-	-	1.131	0.773
11	✓	✓	✓		✓	**0.366**	1.337	889	0.303	0.672	0.741	1.077	0.157	-	-	-	-	1.014	0.717
12	✓	✓	✓	✓	✓	0.358	1.334	**641**	0.302	0.672	0.728	1.054	0.154	**62.3**	**39.5**	52.8	**32.3**	**1.004**	**0.430**

Task Synergy Effect:

- **ID. 4-6:** Track & Map → Motion🚀
- **ID. 7-9:** Motion🚀 ↔ Occupancy🚀
- **ID. 10-12:** Motion & Occupancy → Planning🚀

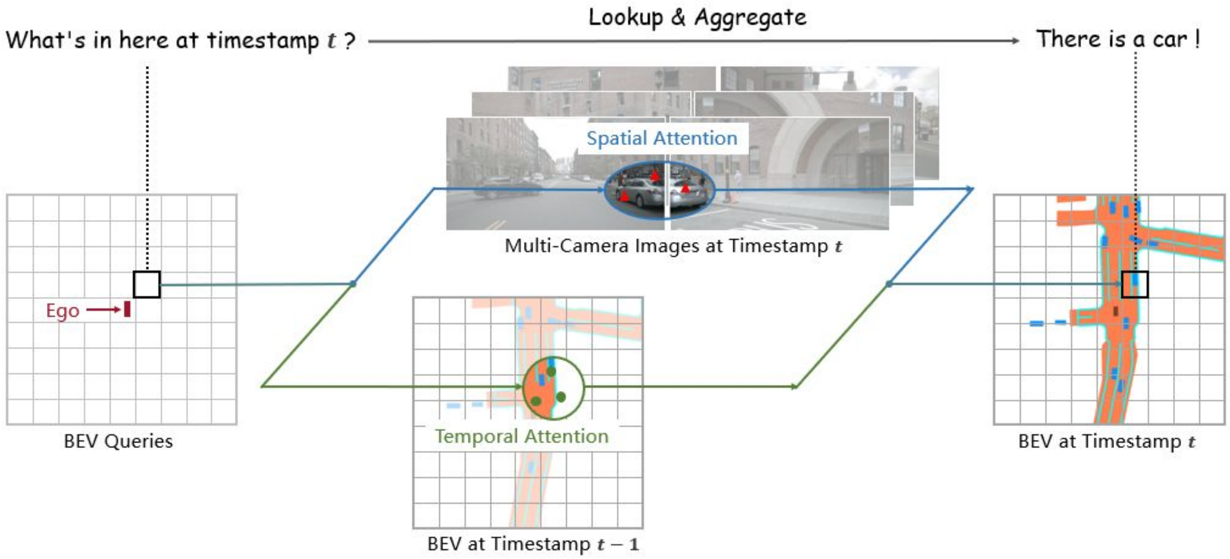
Hu et al. Planning-oriented Autonomous Driving. CVPR 2023.

BEVFormer: Motivation

Li et al. BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. ECCV 2022

In a nutshell…

A unified End-to-End framework which fuses multi-camera and temporal feature based on Deformable Attention and is suitable for various kinds of perception tasks in AD



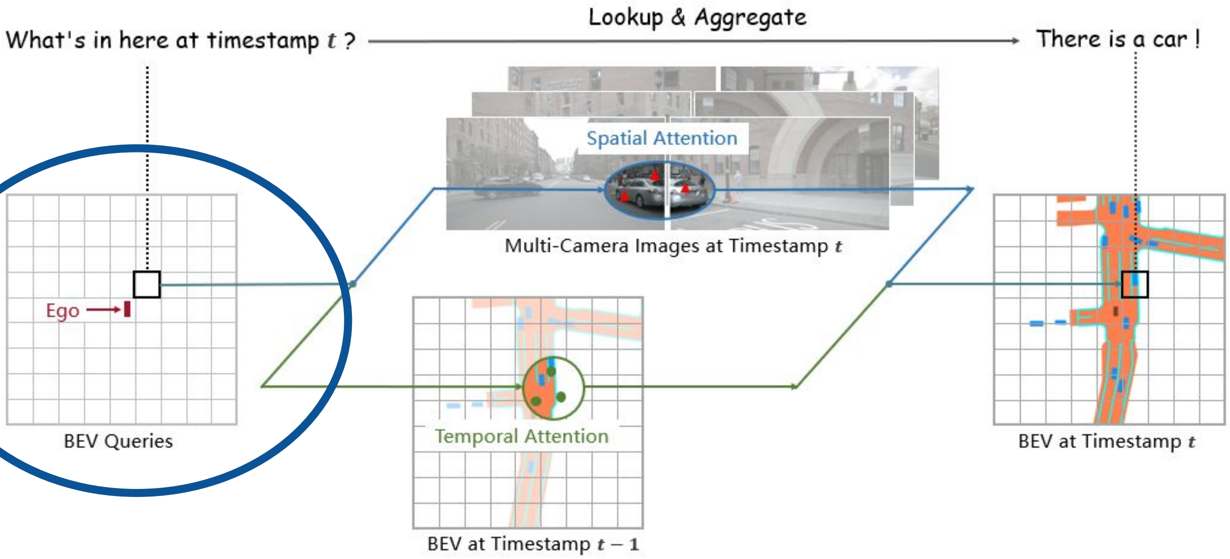
BEVFormer: Motivation

In a nutshell...

A unified End-to-End framework which fuses multi-camera and temporal feature based on Deformable Attention and is suitable for various kinds of perception tasks in AD

Key Recipe...

- **BEV Queries Q**: lookup to obtain BEV feature map



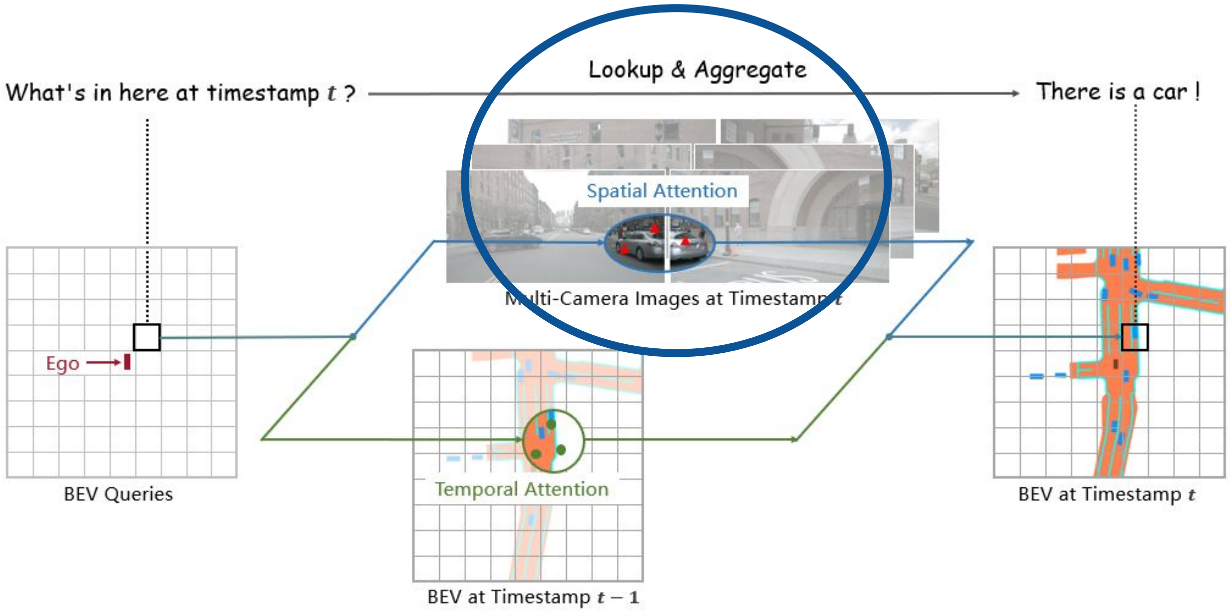
BEVFormer: Motivation

In a nutshell…

A unified End-to-End framework which fuses multi-camera and temporal feature based on Deformable Attention and is suitable for various kinds of perception tasks in AD

Key Recipe…

- **BEV Queries Q** : lookup to obtain BEV feature map
- **Spatial Cross-Attention** : fuse multi-camera feature



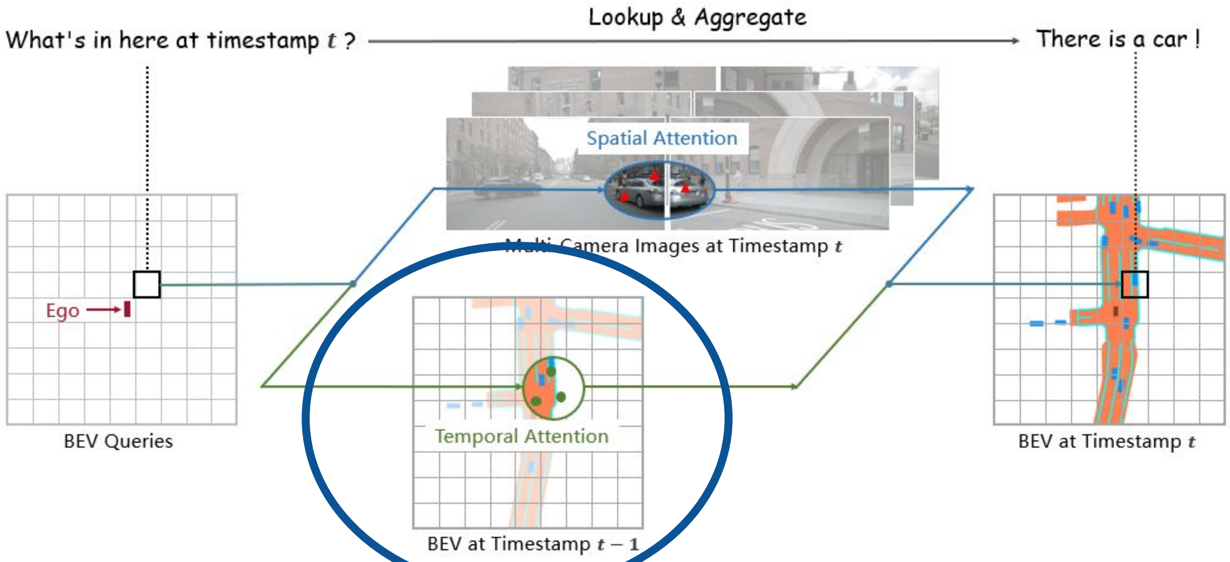
BEVFormer: Motivation

In a nutshell…

A unified End-to-End framework which fuses multi-camera and temporal feature based on Deformable Attention and is suitable for various kinds of perception tasks in AD

Key Recipe…

- **BEV Queries Q** : lookup to obtain BEV feature map
- **Spatial Cross-Attention** : fuse multi-camera feature
- **Temporal Self-Attention** : **aggregate temporal BEV feature**



BEVFormer: Motivation

In a nutshell...

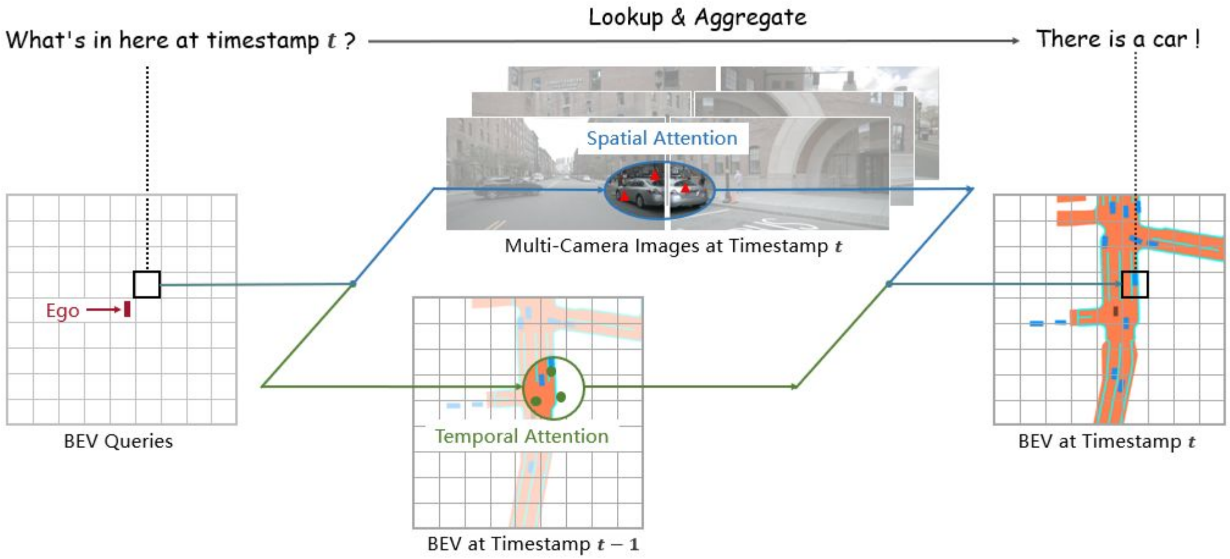
A unified End-to-End framework which fuses multi-camera and temporal feature based on Deformable Attention and is suitable for various kinds of perception tasks in AD

Key Recipe...

- **BEV Queries Q** : lookup to obtain BEV feature map
- **Spatial Cross-Attention** : fuse multi-camera feature
- **Temporal Self-Attention** : **aggregate temporal BEV feature**

Comment

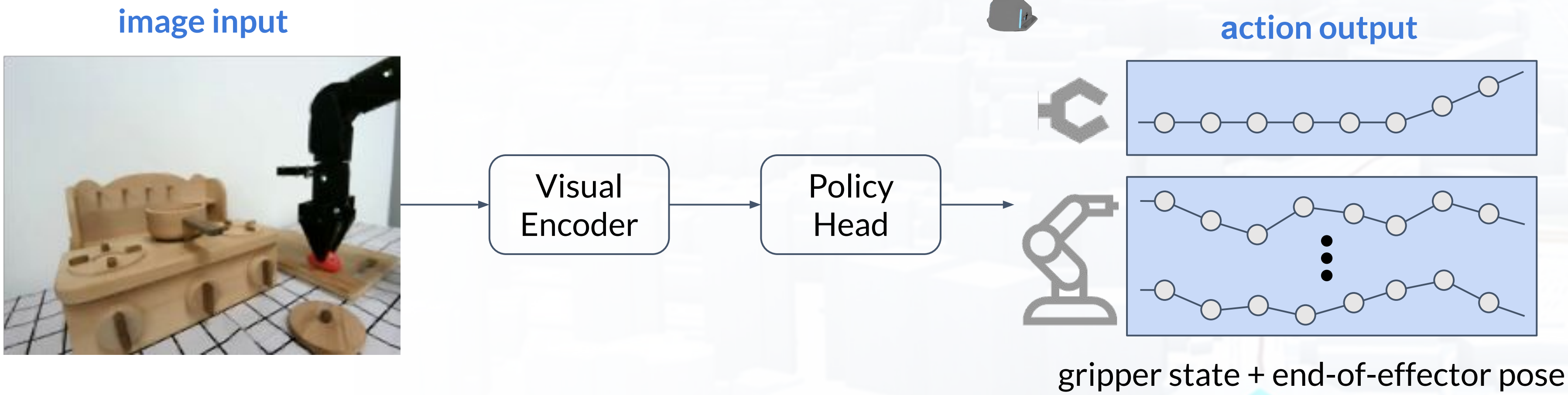
- Using **learnable** queries to represent real world from BEV view
- Look up spatial features in images and temporal features in previous BEV map, aka **Spatial-temporal**



BEVFormer: Demo

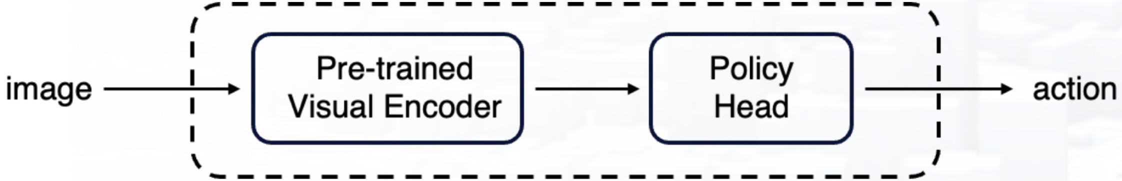
Introduction to Robotic Manipulation

Introduction | Visuomotor control



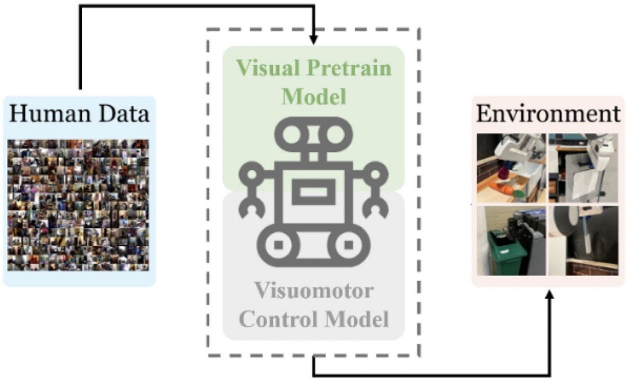
Introduction | Representation Learning for Robotic Manipulation

Pre-trained representations (e.g. CLIP feature) ***enable*** efficient robot learning.



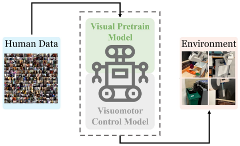
Rutav Shah, Vikash Kumar. RRL: Resnet as representation for Reinforcement Learning. In *ICML*, 2021.
Apoorv Khandelwal, et al. Simple but Effective: CLIP Embeddings for Embodied AI. In *CVPR*, 2022.

However, in-domain robot data is ***scarce***.



To address this limitation:
leverage large-scale human video datasets (e.g. Ego4D)
to extract generalizable features.

Research Roadmap



CoRL 2022
R3M[2]

RSS 2023 Best Paper Finalist
Voltron[4]

RSS 2024
MPI (Ours)

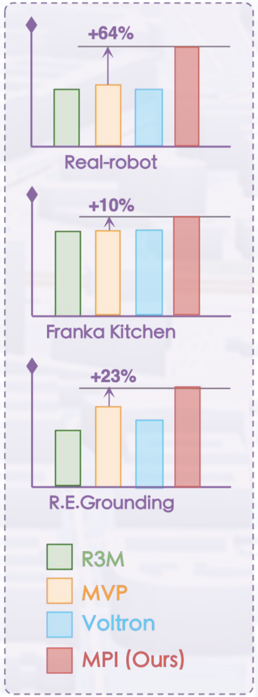
2022/03 — CoRL — 2022/09 — 2023/02 ROBOTICS SCIENCE AND SYSTEMS — 2023/12 — 2024/02 ROBOTICS SCIENCE AND SYSTEMS — Time

2022/03 CoRL — ICLR — ICLR

MVP [1]
CoRL 2022 Oral

VIP[3]
ICLR 2023 Spotlight

GR-1[5]
ICLR 2024



MPI exhibits stronger generalization capability c.f. R3M, MVP and Voltron.

OpenDriveLab

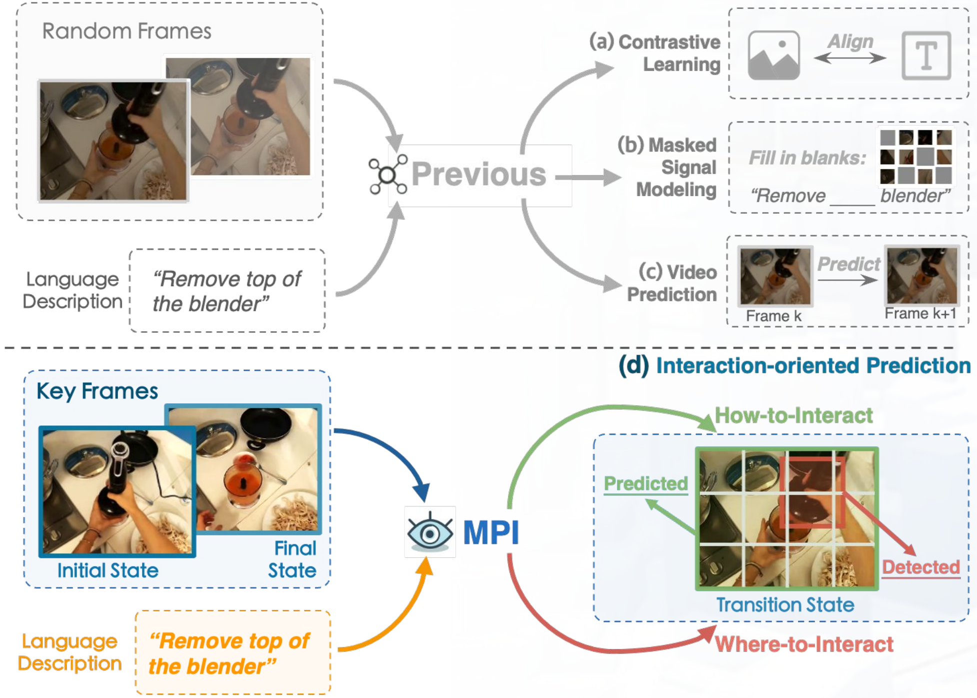
上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

星合联盟
STAR LEAGUE

How to learn a good visual representation in pre-training?

MPI: Manipulation by Prediction Interaction

Motivation



Lack of Explicit Interaction Modeling

Prior Work

- (a) R3M: utilize contrastive learning, focus on high-level semantics.
- (b) MVP: apply MAE, focus on low-level pixel-wise information.
- (c) GR-1: sequential video prediction, but introduce noise or redundant information
  ❌ *effectively capture the dynamic interactions*

Ours

- Reflect upon the pre-training objectives
- Instill interactive dynamics by proposing an interaction-oriented prediction paradigm

Interactive dynamics: *the patterns of behavior and physical interactions that occur between a robot and the environment*

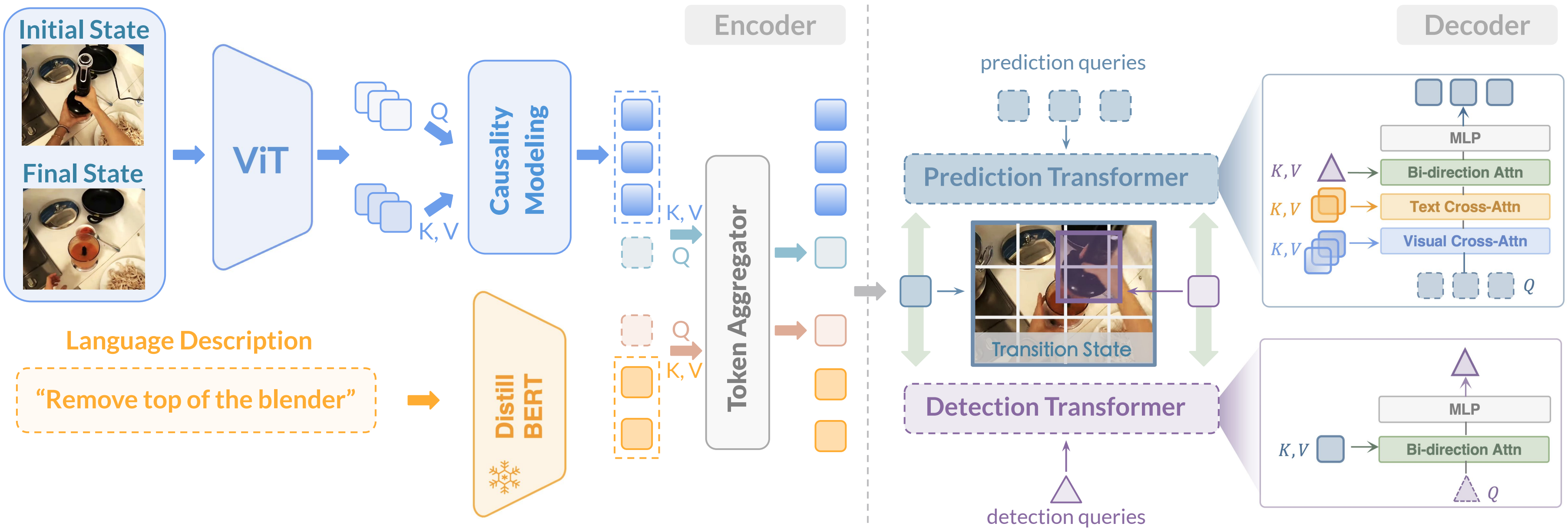
MPI | Pipeline and Framework

Pre-training dataset

Ego4d hand-and-object subset | Keyframes

Two Training Objectives

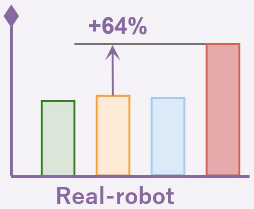
“how to interact” | “where to interact”



Zeng et al, Learning Manipulation by Predicting Interaction, RSS 2024

MPI | Experiments

Performance Comparison



R3M — Stanford University, Meta — CoRL 2022

MVP — Berkeley University of California — CoRL 2022 oral

Voltron — Stanford University, Toyota — RSS 2023 best paper finalist

MPI (Ours)

Generalization Validation

Robustness to Visual Distractions

(a) Original Setting

(b) BG. Distraction

(c) Obj. Variation

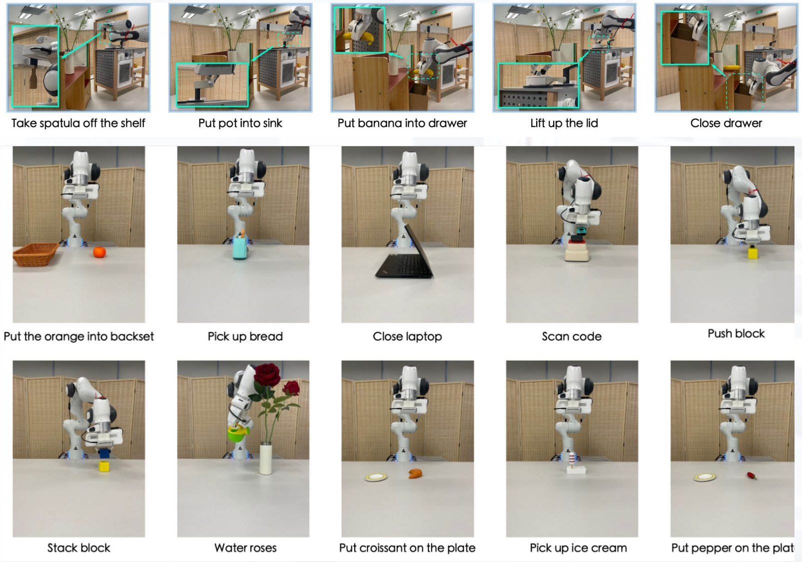
Evaluation Suite

- Evaluations on visuomotor control: both in real world and in simulation(Franka Kitchen, MetaWorld) 🦾
- Referring Expression Grounding 👀👉

Real-robot Experiment Setting

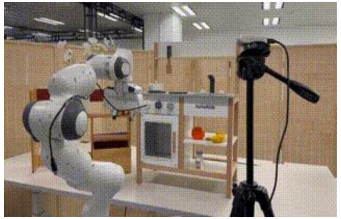
5 complex kitchen environment

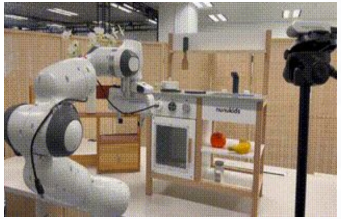
10 clean background

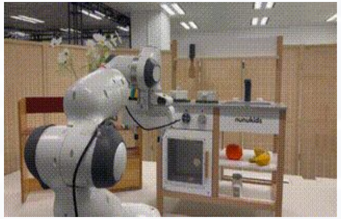


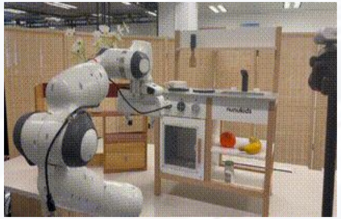
MPI - Testament on Real Robots

Demo in kitchen environment





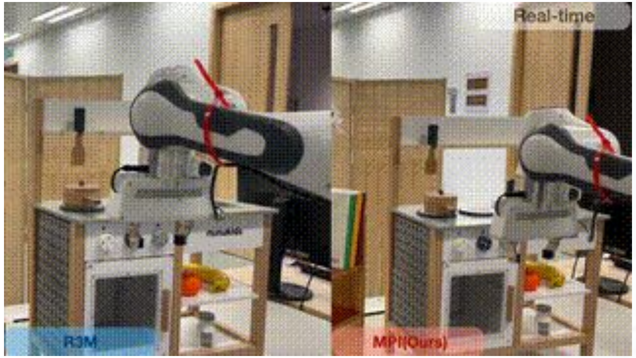




Validation on generalization

Object Variation

White plastic pot → Wooden pot

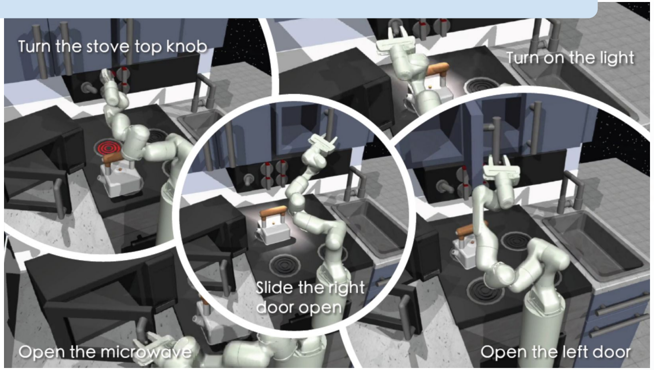


Background Distraction

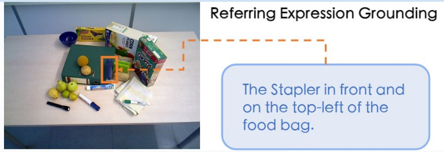
Daisies → Roses

MPI - Experiments

Visuomotor Control in Simulation



Referring Expression Grounding



Method	Embedding	Average Precision (AP) @0.25	@0.5	@0.75
R3M [36]	$\mathbb{R}^{2048}$	85.27	71.79	42.66
MVP [40]	$\mathbb{R}^{384}$	93.07	85.32	60.37
Voltron [24]	$\mathbb{R}^{196\times384}$	92.93	84.70	57.61
MPI (Ours)*	$\mathbb{R}^{384}$	**96.29**	**92.10**	71.87
MPI (Ours)	$\mathbb{R}^{196\times384}$	96.04	92.05	**74.40**

The experimental results reveal that MPI yields **state-of-the-art** performance on a broad spectrum of downstream tasks.

TABLE II: **Results of single-task visuomotor control on Franka Kitchen.** We report the success rate (%) over 50 randomly sampled trajectories. We **bold** the best result for models with similar parameters and underline the second. "INSUP." represents classification-based supervised learning on ImageNet. MPI consistently exhibits superior performance across multiple tasks.

Method	Backbone	Param.	Turn knob	Open door	Flip switch	Open microwave	Slide door	Average
INSUP. [21]	ResNet50	25.6M	28.0	18.0	50.0	26.7	75.7	39.7
CLIP [39]	ResNet50	25.6M	26.3	13.0	41.7	24.7	86.3	38.4
R3M [36]	ResNet50	25.6M	53.3	**50.7**	86.3	59.3	97.7	69.5
Voltron [24]	ViT-Small	22M	71.7	45.3	**95.3**	40.3	99.7	70.5
MPI (Ours)	ViT-Small	22M	**83.3**	50.3	89.0	**59.7**	**100.0**	**76.5**
MVP [40]	ViT-Base	86M	79.0	48.0	90.7	41.0	**100.0**	71.7
Voltron [24]	ViT-Base	86M	76.0	45.3	91.0	41.0	99.3	70.5
MPI (Ours)	ViT-Base	86M	**89.0**	**57.7**	**93.7**	**54.0**	**100.0**	**78.9**

TABLE III: **Results of single-task visuomotor control on Meta-World simulation environment.** We report the success rate (%) over 50 randomly sampled trajectories. The best results are **bolded** and the second highest are underlined. MPI showcases exemplary performance across three tasks, exhibiting a superior average success rate in comparison to prior methods.

Method	Backbone	Param.	Assemble	Pick & Place	Press Button	Open Drawer	Hammer	Average
R3M [36]	ResNet50	25.6M	**94.0**	60.3	66.3	100	93.7	82.9
MVP [40]	ViT-Base	86M	82.7	**82.0**	62.7	100	95.7	84.6
Voltron [24]	ViT-Small	22M	72.3	57.3	30.7	100	83.0	68.7
MPI (Ours)	ViT-Small	22M	69.0	64.0	**98.7**	**100**	**96.0**	**85.7**

MPI | Takeaways

- By instructing the model towards ***predicting*** transition frames and ***detecting*** manipulated objects, the model can foster better comprehension of "how-to-interact" and "where-to-interact".
- Interaction-level feature yields enhanced generalizability.
- The tasks of predicting transition frame and detecting manipulated objects can promote each other.

Project page

Code on Github

OpenDriveLab

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

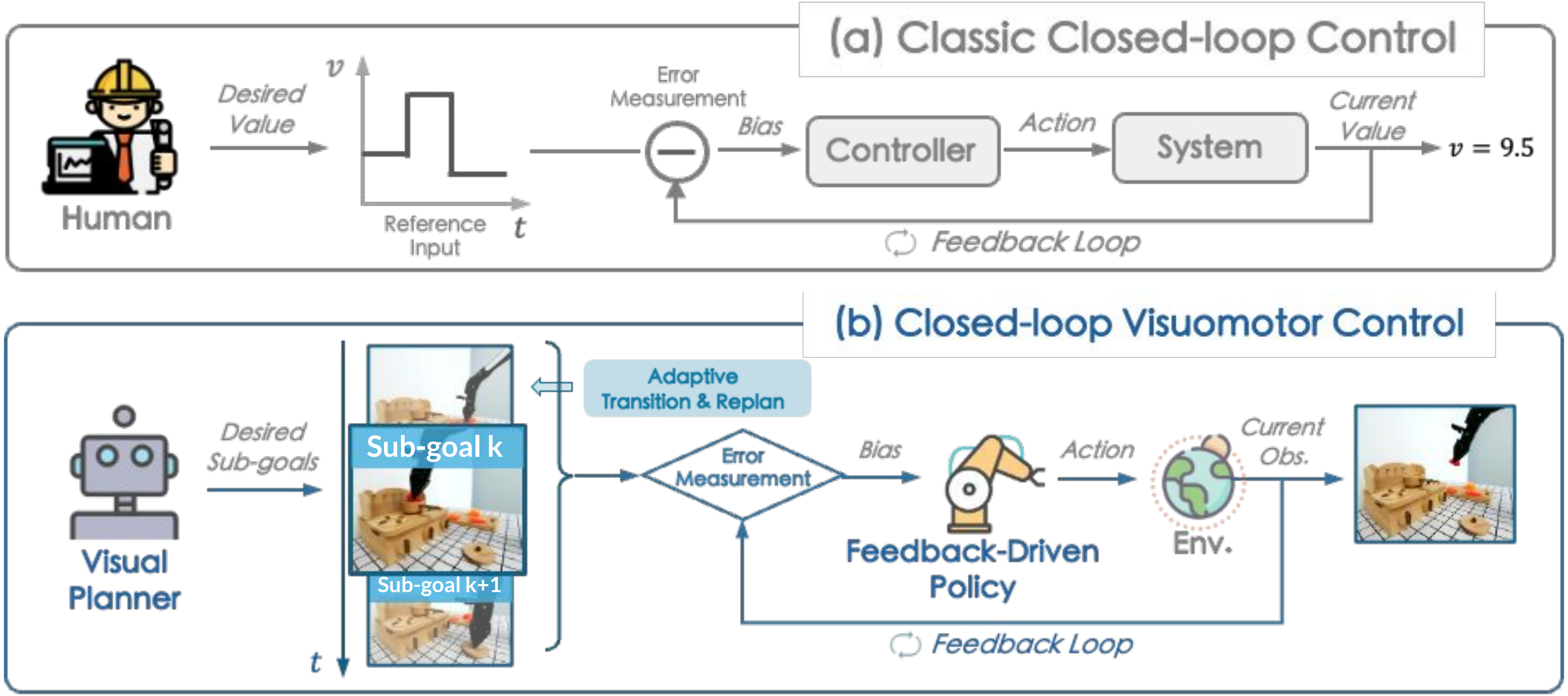
星合联盟
STAR LEAGUE

How to "calibrate" intermediate state towards better policy actions?

CLOVER: Closed-Loop Visuomotor Control with Generative Expectation for Robotic Manipulation

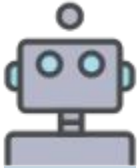
Motivation

From classic control to visuomotor policy



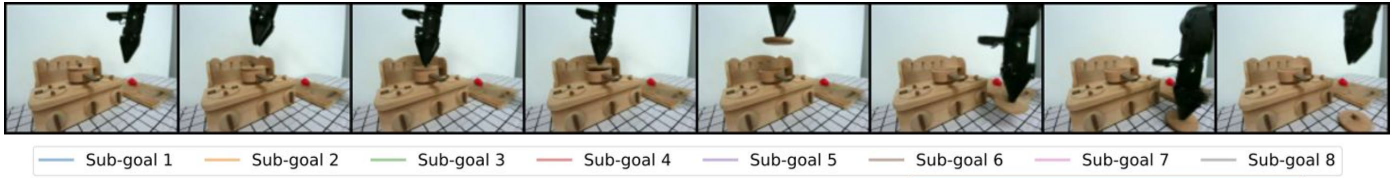
CLOVER: Adaptive Subgoal Transition + Re-planning + Completion Assessment

Key observation



Diffusion Model

Generated Sub-goals

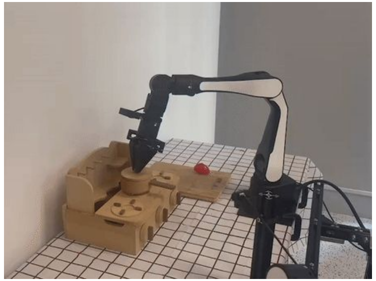


Guide

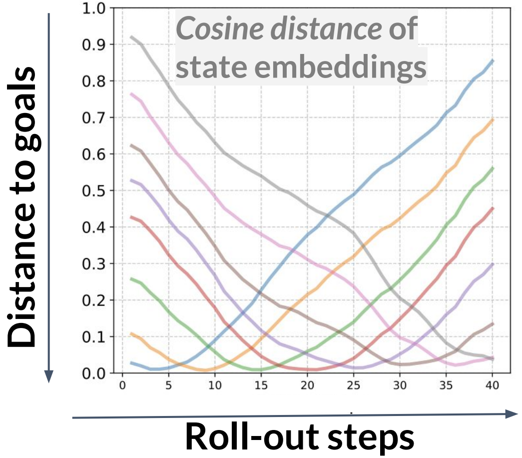
Align

Actual Roll-out

Policy



State Embeddings:
Effectively measures current-to-goal errors.

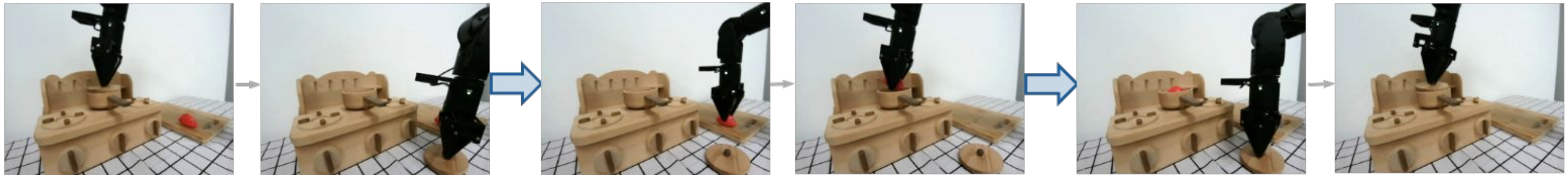


Bu, et al, Closed-Loop Visuomotor Control with Generative Expectation for Robotic Manipulation, in submission

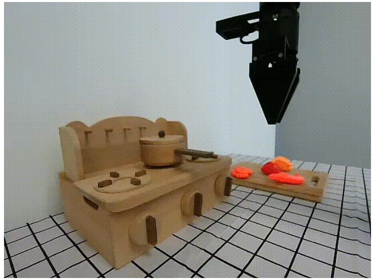
Experiments

Real-world Robots: Consecutive Tasks

Sub-task 1: lift up the pot lid　　Sub-task 2: put the fish into the pot　　Sub-task 3: put the lid on the pot



➔ Substantial improvement over language-conditioned behaviour cloning baselines



+30%

Method	Task completed in a row (%) ↑ 1	2	3	Avg. Len. ↑
ACT [53]	46.7	13.3	0.0	0.6
R3M [54]	53.3	20.0	0.0	0.7
RT-1 [49]	66.7	40.0	0.0	1.1
CLOVER (Ours)	**93.3**	**86.7**	**26.7**	**2.1**

Experiments

Simulation: CALVIN Benchmark

- **+8%** v.s. 3D Diffuser Actor (previous SOTA) with more inputs
- **+30%** v.s. Previous "Planner + Executor" Method

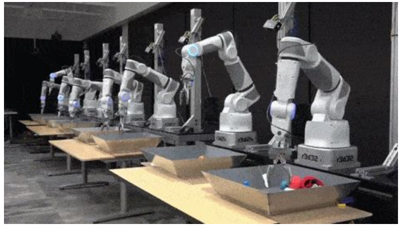
Method	Type	Train episodes	Task completed in a row (%) ↑ 1	2	3	4	5	Avg. Len. ↑
MCIL [47]	Language-conditioned Behaviour Cloning	All	30.4	1.3	0.2	0.0	0.0	0.31
HULC [48]		All	41.8	16.5	5.7	1.9	1.1	0.67
RT-1 [49]		Lang	53.3	22.2	9.4	3.8	1.3	0.90
RoboFlamingo [50]		Lang	82.4	61.9	46.6	33.1	23.5	2.48
GR-1 [51]		Lang	85.4	71.2	59.6	49.7	40.1	3.06
3D Diffuser Actor [52]	Diffusion Policy	Lang	92.2	78.7	63.9	51.2	41.2	3.27
UniPi* [14]	Planner + Executor	All	56.0	16.0	8.0	8.0	4.0	0.92
SuSIE [15]		All	87.0	69.0	49.0	38.0	26.0	2.69
CLOVER (Ours)		Lang	**96.0**	**83.5**	**70.8**	**57.5**	**45.4**	**3.53**

Concluding Remarks

What's Next for Robotic Manipulation

Data Collection

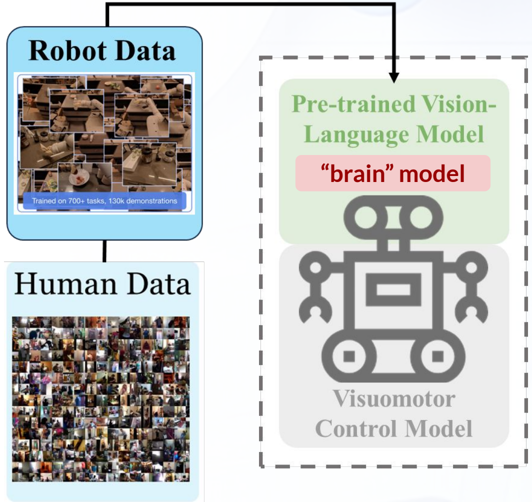
Programming



Tele-operation

Engine / Pre-training

Pre-train a "brain" model for robotic "upper body" tasks.



generali
-zation

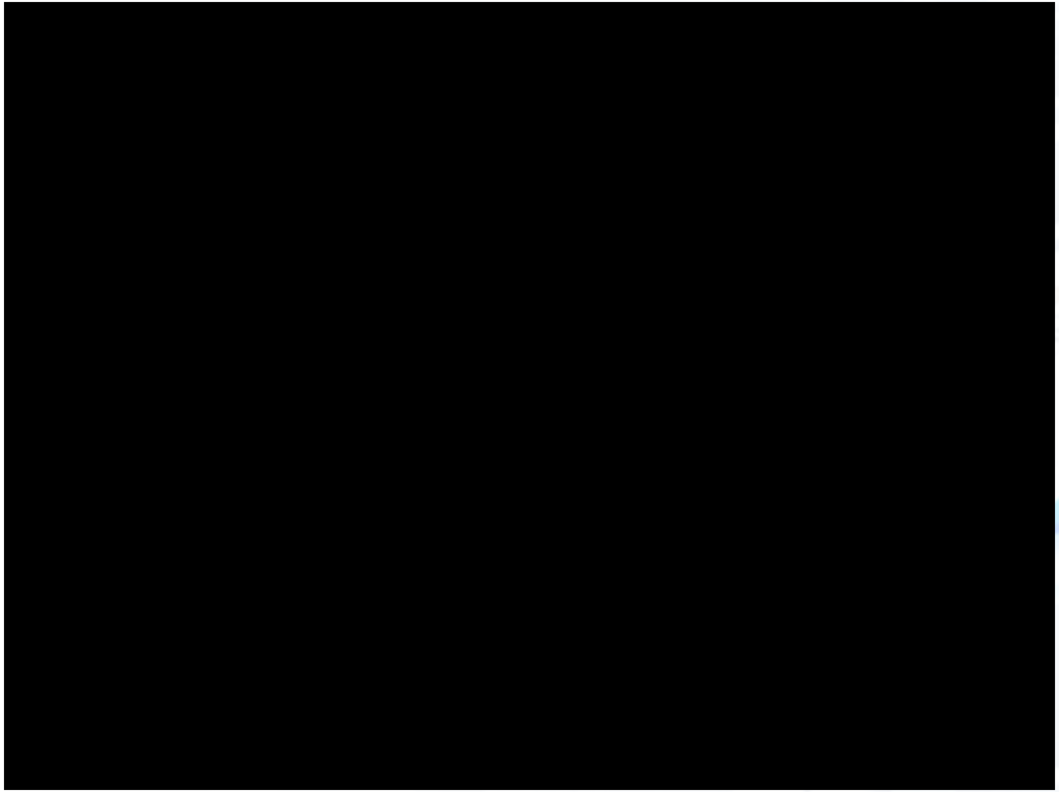
General-purpose, interpretable embodied foundation model with causal reasoning capabilities

Application / Task

Fixed Manipulation

Mobile Manipulation

Humanoid Robots for Manufacturing



Kudos to Our Fantastic Members / Collaborators

Also the slide credit

Meet our team in
Delft @RSS 2024!!!

Jia Zeng

MPI

Qingwen Bu

CLOVER

Li Chen

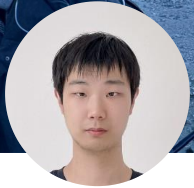
UniAD

Chonghao Sima

DriveLM

Huijie Wang

OpenLane



Zetong Yang

ViDAR

Yunsong Zhou

ELM

And many others …

Yihang Qiu

Tianyu Li

Shenyuan Gao

Jiazhi Yang