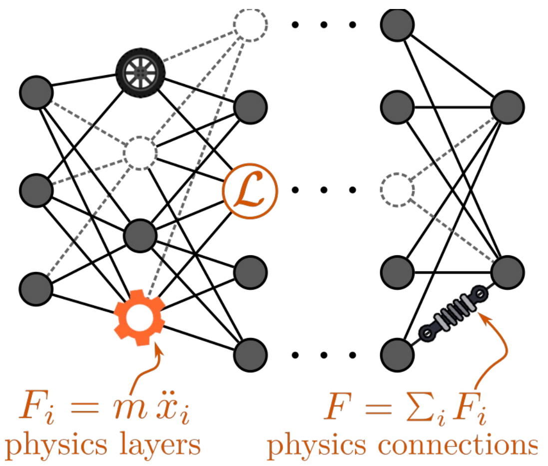
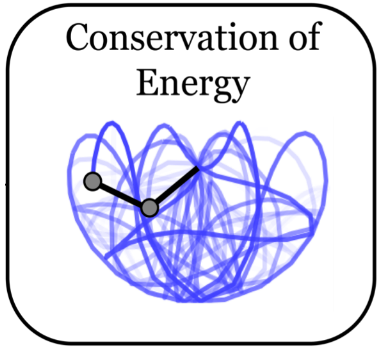
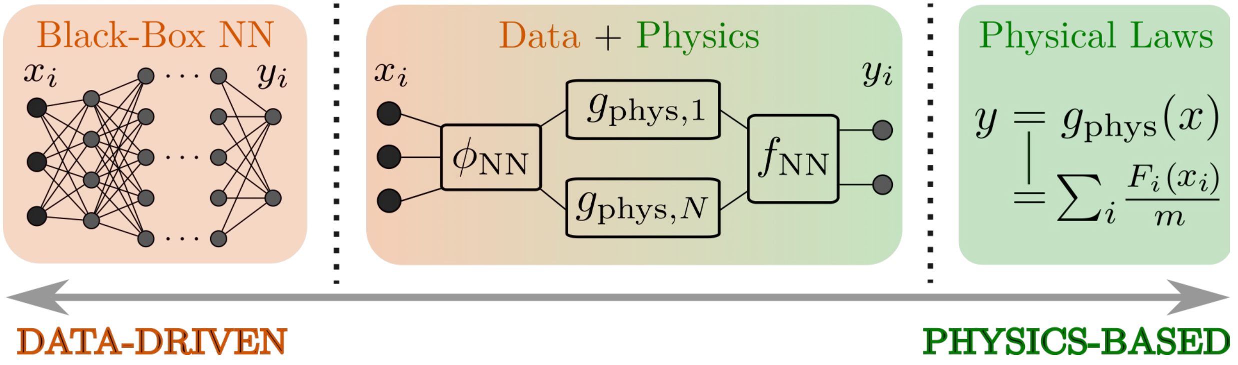
Outlook: Challenges & Future Directions

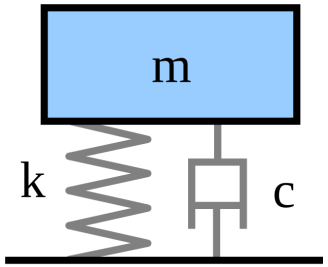


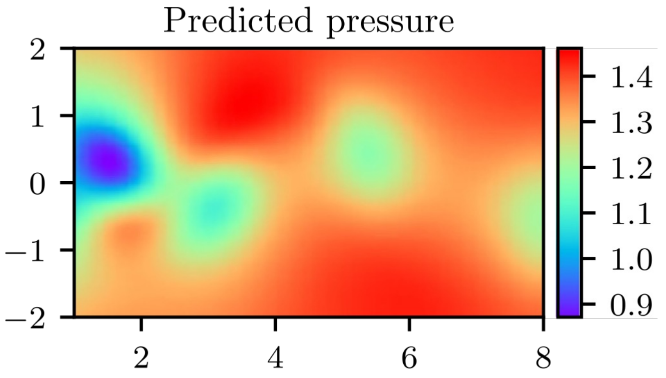




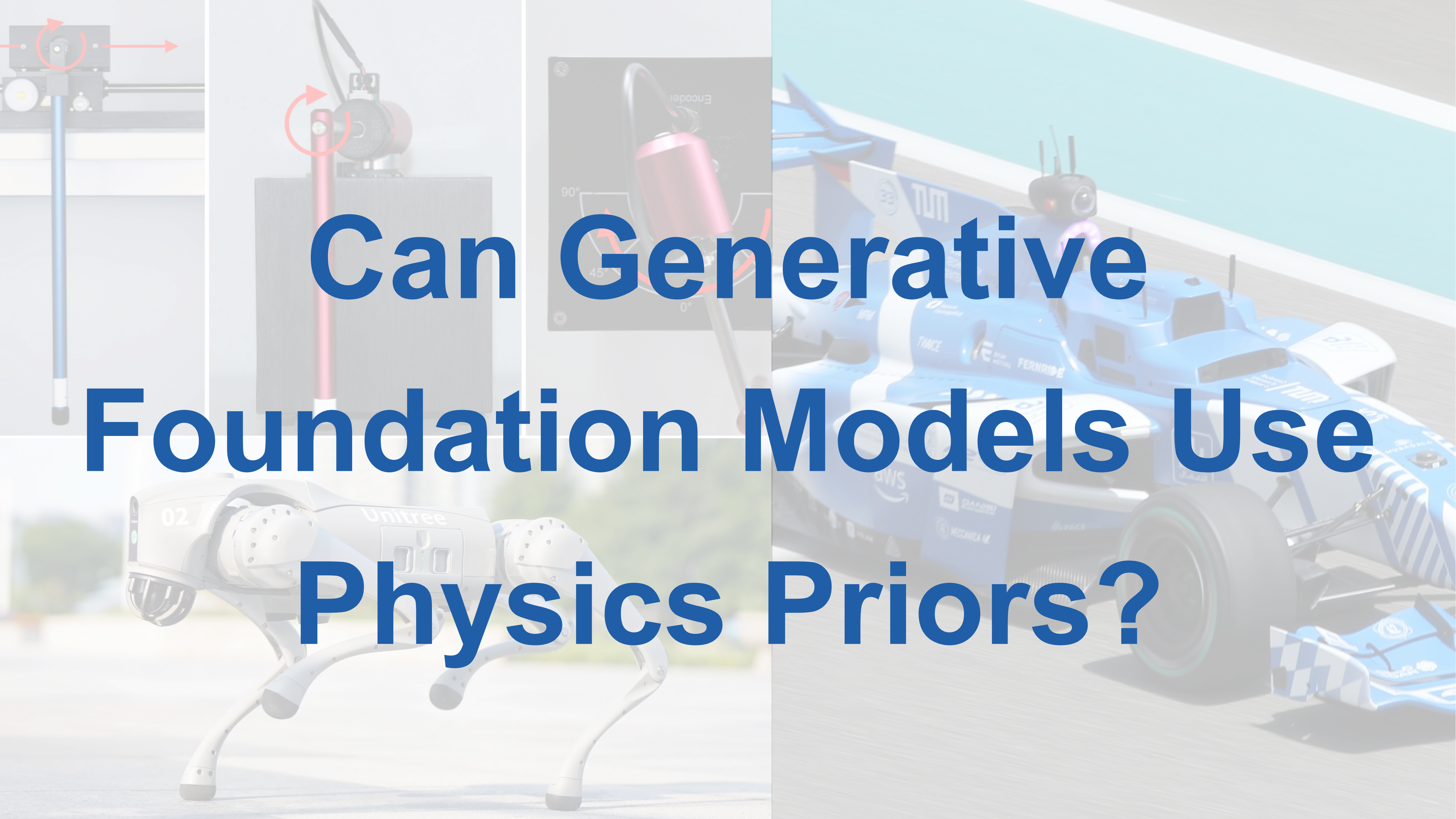
Dr. Mattia Piccinini &
Prof. Gastone Pietro Rosati Papini

University of Trento &
Technical University of Munich,
Autonomous Vehicle Systems Lab



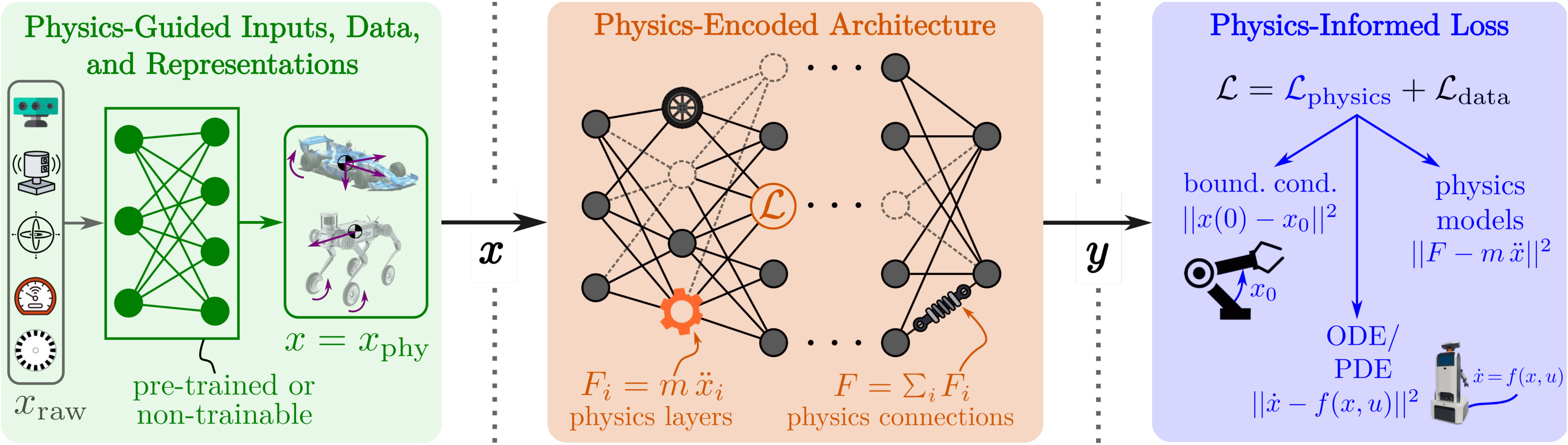


Can Generative Foundation Models Use Physics Priors?



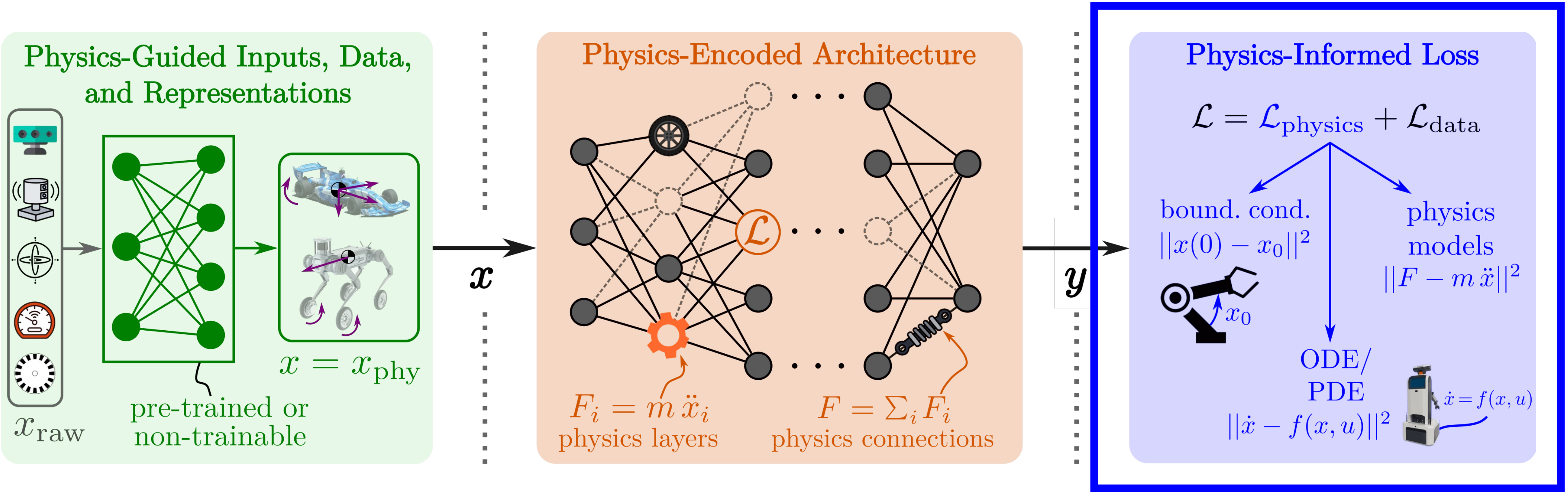
What About Generative Foundation Models with Physics Priors?

- Very few examples, but the topic is becoming very **hyped**!
- For embodied AI with **physical systems**, the data we have is often not enough
- Physics priors can „**fill the data gap**“ & provide better **generalization**



Physics-Informed Video-Language-Action Model (VLA)

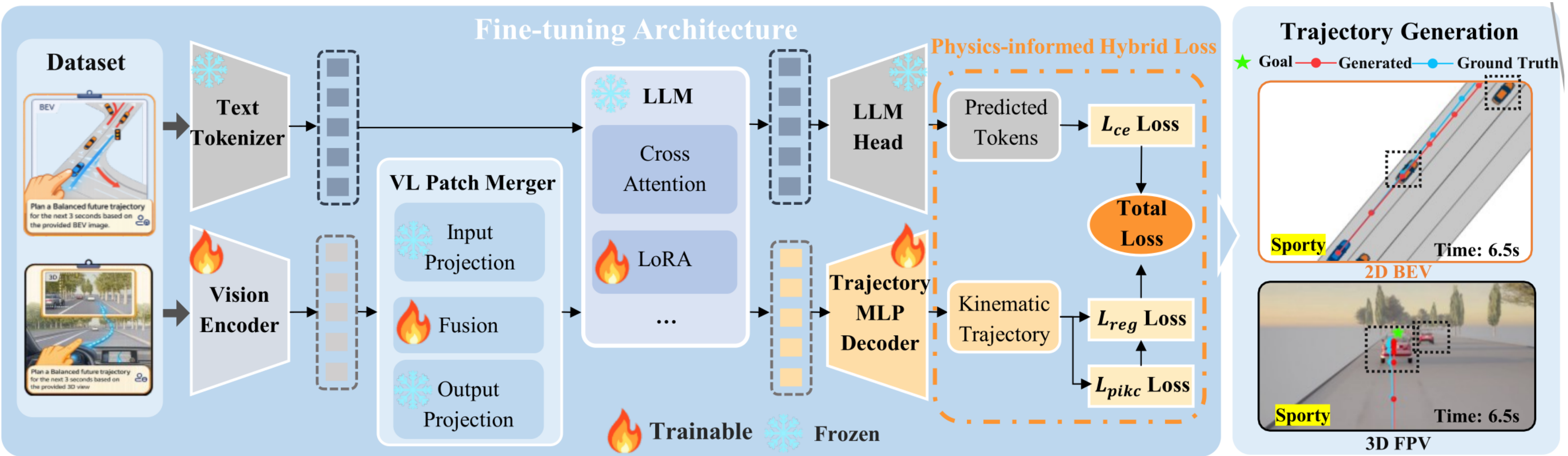
An example from our **TUM-AVS lab** in the context of **physics-informed learning** with vision-based semantic reasoning for **trajectory planning**



Y. Gao, D. Hua, **M. Piccinini**, F. R. Schäfer, K. Moller, L. Li, J. Betz, “StyleVLA: Driving Style-Aware Vision Language Action Model for Autonomous Driving”, *available on arXiv,* 2026

Physics-Informed Video-Language-Action Model (VLA)

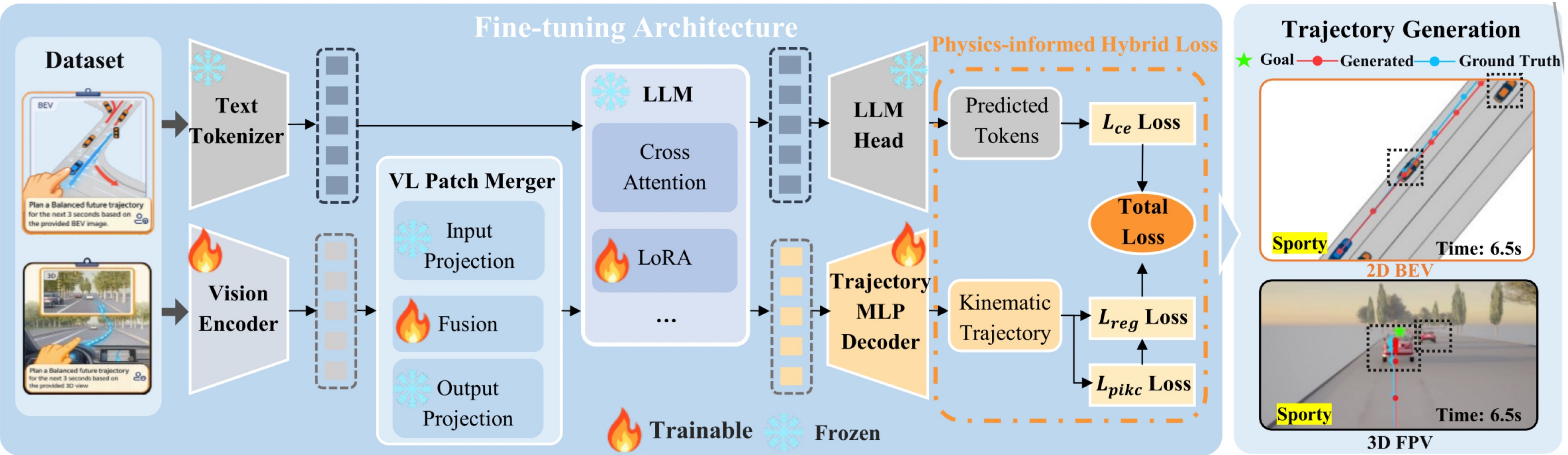
- Ask a **VLA** to drive a car with a **desired style**
- **VLA inputs**: ego states (0.5 sec history), goal state (x,y coords), image of the current time step: 2D bird-eye-view (BEV) or 3D first-person-view (FPV)
- **VLA outputs**: $[x, y, \theta, v, a]$ for the next 3 sec



Y. Gao, D. Hua, **M. Piccinini**, F. R. Schäfer, K. Moller, L. Li, J. Betz, "StyleVLA: Driving Style-Aware Vision Language Action Model for Autonomous Driving", *available on arXiv*, 2026

Physics-Informed Video-Language-Action Model (VLA)

- Ask a **VLA** to drive a car with a **desired style**
- Using the Qwen3-VL-4B VLA (4 billion parameters)
- VLA **fine-tuning** on a custom dataset
- Fine-tuning only the vision encoder, lightweight adaptor matrices (QLoRA), MLP decoder



Y. Gao, D. Hua, **M. Piccinini**, F. R. Schäfer, K. Moller, L. Li, J. Betz, "StyleVLA: Driving Style-Aware Vision Language Action Model for Autonomous Driving", *available on arXiv,* 2026

Physics-Informed Video-Language-Action Model (VLA)

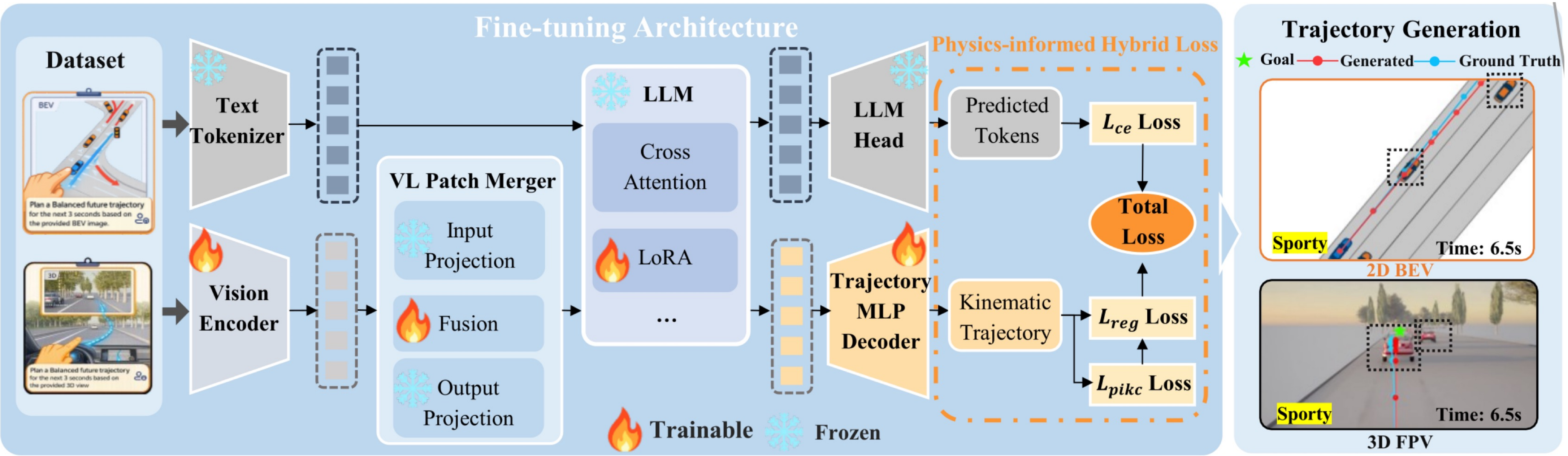
Loss function design:

- Standard cross-entropy loss
- **Physics-informed** kinematic (**PIKC**)
  - $\hat{x}, \hat{y}$ = predictions from prev. time step

    ➔ unsupervised physics loss

$$\hat{x}_{t+1} = x_t + v_t \cos\theta_t \Delta t + 0.5\, a_t \cos\theta_t (\Delta t)^2,$$
$$\hat{y}_{t+1} = y_t + v_t \sin\theta_t \Delta t + 0.5\, a_t \sin\theta_t (\Delta t)^2.$$

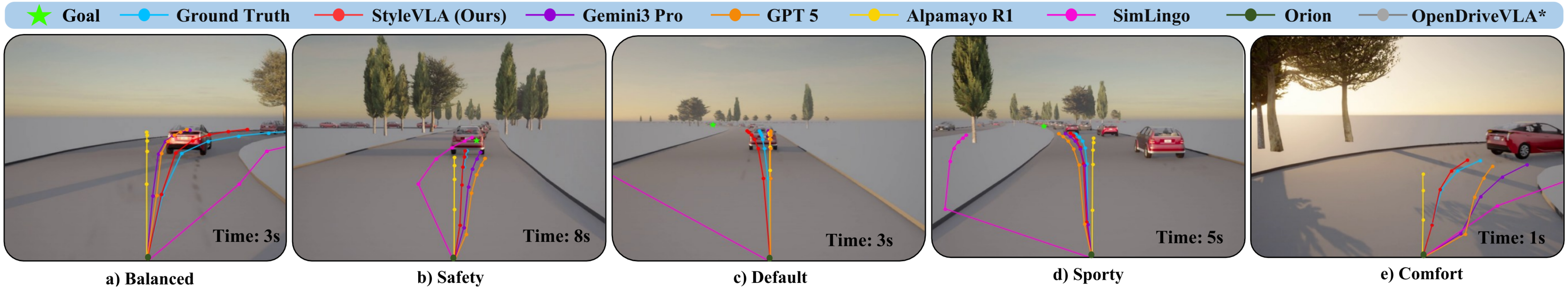
$$\boxed{\mathcal{L}_{\text{pikc}} = \frac{1}{T-1} \sum_{t=0}^{T-1} \left(||x_{t+1} - \hat{x}_{t+1}||^2 + ||y_{t+1} - \hat{y}_{t+1}||^2 \right)}$$



Y. Gao, D. Hua, **M. Piccinini**, F. R. Schäfer, K. Moller, L. Li, J. Betz, "StyleVLA: Driving Style-Aware Vision Language Action Model for Autonomous Driving", *available on arXiv*, 2026

Physics-Informed Video-Language-Action Model (VLA)

Outperforming proprietary VLAs (including Alpamayo R1)



Model	Score ↑ (0-1)	PSR ↑ (ADE< 1m)	MR ↓ (FDE> 2m)	ADE ↓ (m)	FDE ↓ (m)	KCE ↓ (m)	Time ↓ (s)
Open Source Baselines							
Qwen3-VL-4B (Base)	0.00	-	-	-	-	-	4.97
Proprietary Models							
Gemini 2.5 Flash	0.12	9.04%	87.21%	9.39	18.21	3.42	1.80
Gemini 2.5 Pro	0.29	16.63%	70.74%	2.23	5.76	0.06	35.48
GPT-5 Nano	0.29	16.67%	69.70%	2.36	6.06	0.11	49.05
Gemini 3 Pro	0.35	17.65%	62.35%	1.54	3.94	**0.06**	91.39
SOTA Models							
SimLingo (1B)	0.00	0.30%	99.40%	8.01	17.58	/	0.55
Orion (7B)	0.05	2.10%	96.40%	11.13	21.35	/	**0.36**
OpenDriveVLA (0.5B)	0.13	7.38%	87.25%	5.83	9.82	/	0.51
Alpamayo-R1 (10B)	0.19	13.60%	73.10%	3.37	5.85	/	0.65
Qwen3-VL-4B (StyleVLA)	**0.51**	**38.60%**	**36.90%**	**1.17**	**3.13**	0.11	2.13

Y. Gao, D. Hua, **M. Piccinini**, F. R. Schäfer, K. Moller, L. Li, J. Betz, "StyleVLA: Driving Style-Aware Vision Language Action Model for Autonomous Driving", *available on arXiv,* 2026

Physics-Informed Video-Language-Action Model (VLA)

- Impact of **dataset** size: the more, the better
- Impact of **physics-informed loss** component (**PIKC**): better generalization

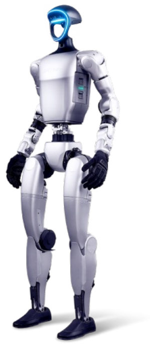
Configuration	**ADE** (m) ↓	**FDE** (m) ↓	**PSR** ↑	**Heading MAE** (rad) ↓
Impact of Training Data Scaling (with Physics-Informed Hybrid Loss)				
Small (4.5k)	2.08	5.43	20.60%	0.073
Medium (20k)	1.51	3.92	27.14%	0.046
Large (40k)	1.47	3.81	29.37%	0.044
Standard (50k)	**1.17**	**3.06**	**33.19%**	**0.035**
Impact of Loss Components (50k Training dataset)				
CE	1.47	3.82	29.00%	0.043
CE + REG	1.21	3.17	32.08%	0.036
CE + REG + PIKC	**1.17**	**3.06**	**33.19%**	**0.035**

Y. Gao, D. Hua, **M. Piccinini**, F. R. Schäfer, K. Moller, L. Li, J. Betz, "StyleVLA: Driving Style-Aware Vision Language Action Model for Autonomous Driving", *available on arXiv,* 2026

What Are Promising Future Directions?

Future Directions to Address Open Challenges

- Combining **multiple physics priors routes**: physics-guided inputs and data, physics-encoded architectures, physics-informed loss functions
- **Foundation models** in Embodied AI: use **physics priors to fill the data gap**
- **Differentiable end-to-end training** to boost performance & rapid adaptation
- **Online lifelong** physics-aware adaptation
- Provably **safe** and verifiable system performance
- **Interpretability** and social acceptance

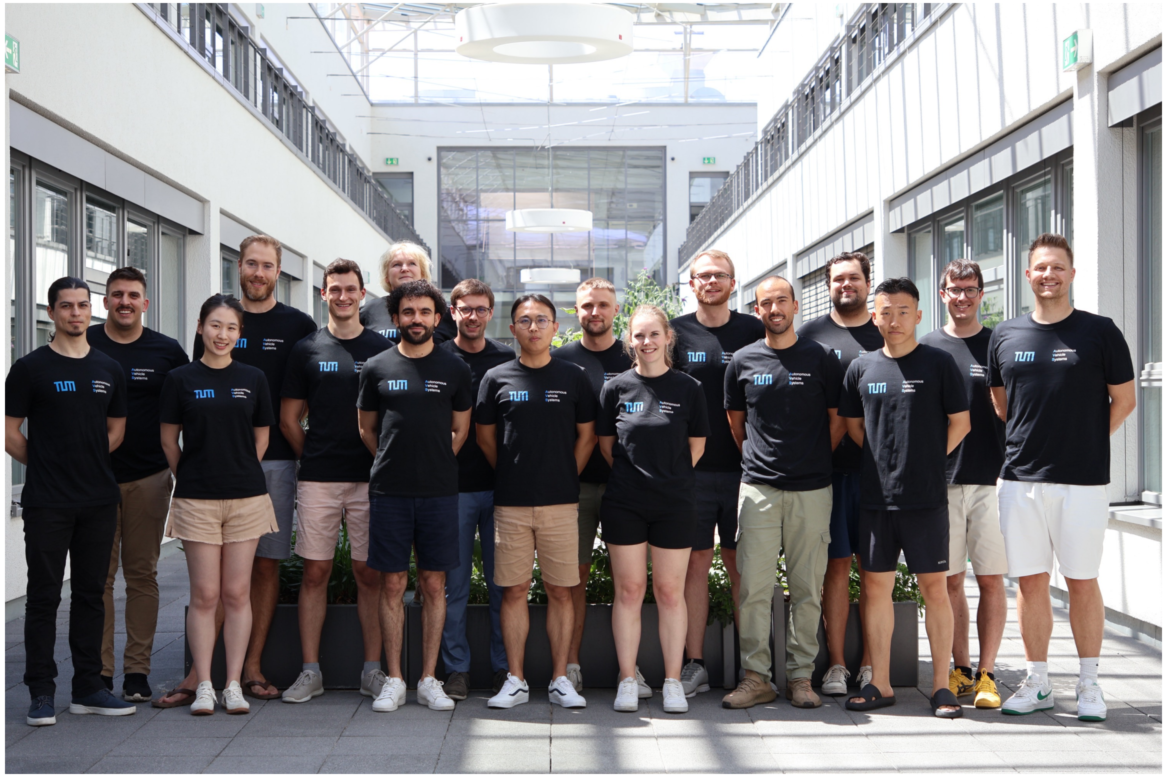


Thank You for Attending!

- Remember to send us your **report** to receive the credits by the **end of August**
- We are happy to **support** you in developing a good **research idea** (both for the report and the following implementation, if you like!)
- We will create an Excel file in the shared Google Drive folder, where you can write if you need to pass the final „exam“ (report), or if you just need an attendance certificate → please use the shared Excel for this

Feedback needed: We prepared a **short questionnaire** (5 min) to help us improve next year's edition. Thanks a lot!

Technical University of Munich



Let’s work together!