Karlsruhe Institute of Technology

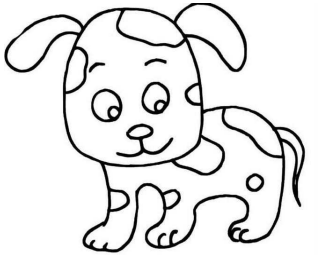
High Performance Humanoid Technologies

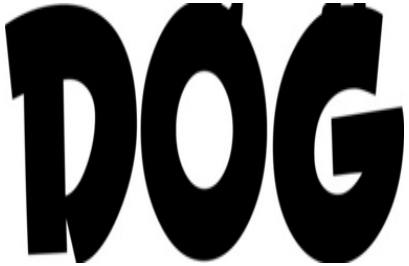
# Alignment of Modalities in Foundation Models

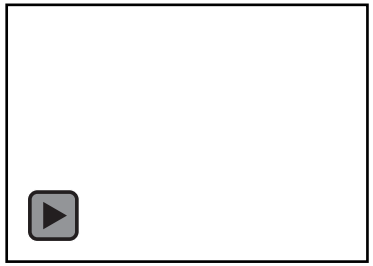
Weiheng Wang, Timo Birr and Tamim Asfour

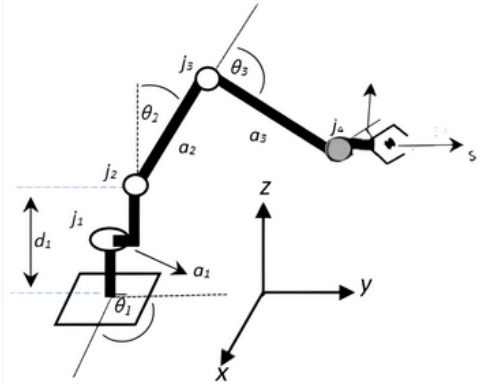
Seminar Humanoid Robots WS 2024/2025

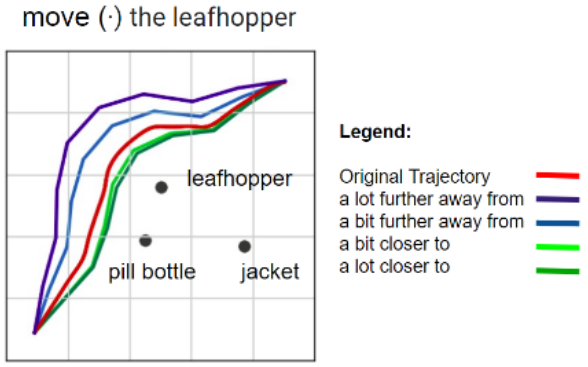
# Why Multi-modal





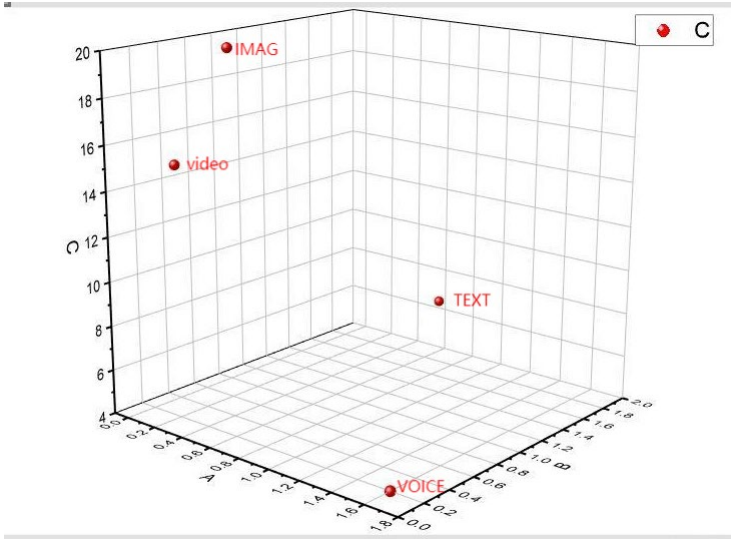



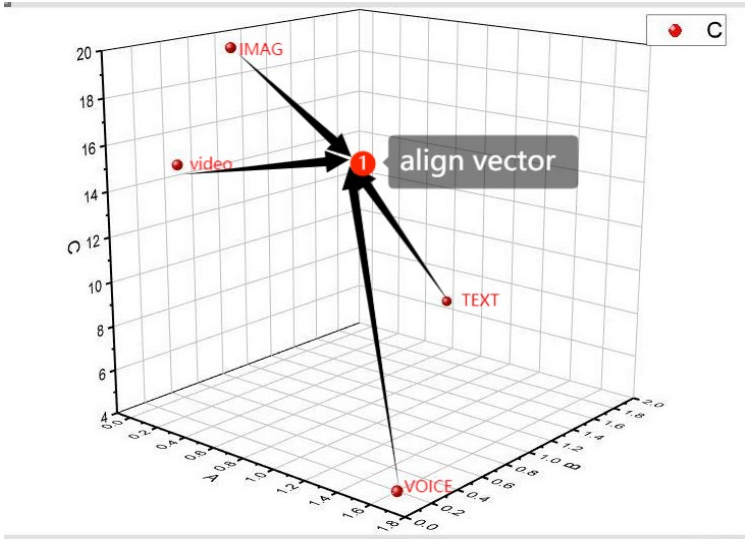


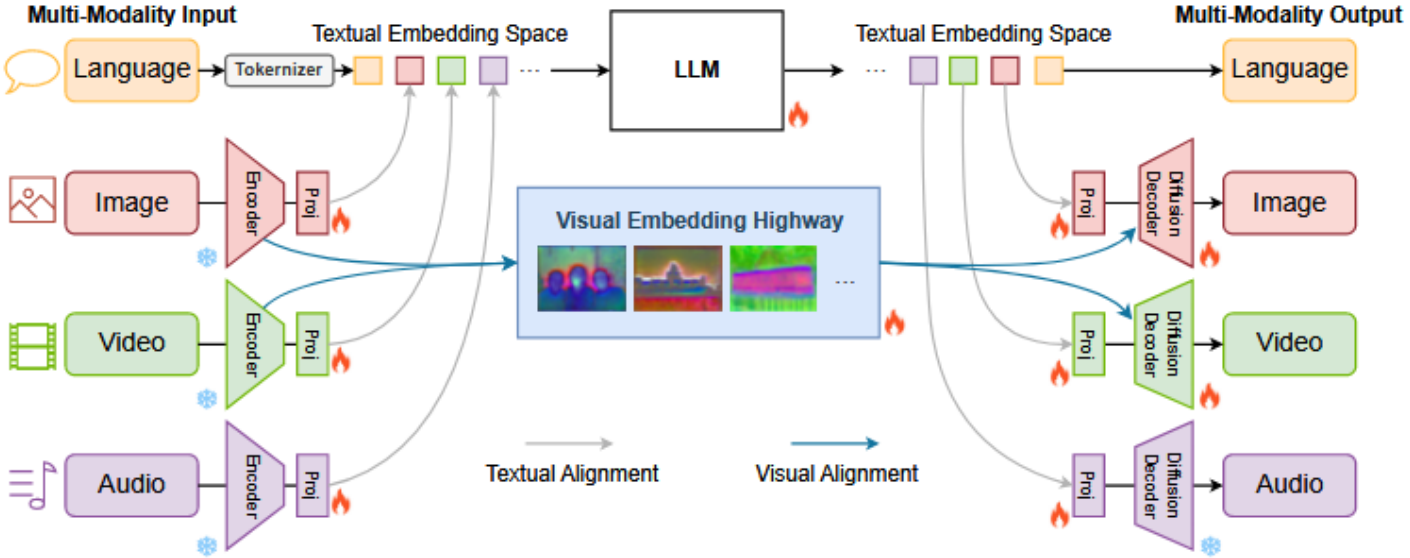

Reshaping Robot Trajectories Using Natural Language Commands: A Study of Multi-Modal Data Alignment Using Transformers

# Why Alignment of multi-modalities



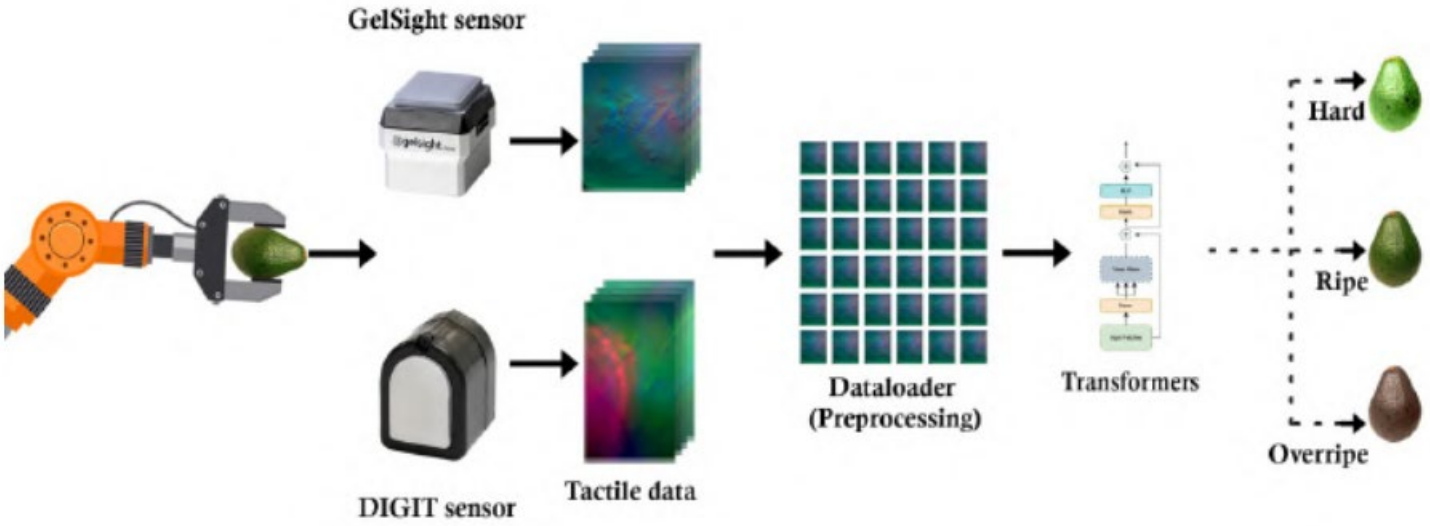
# The foundation structure of MM-LLMs




X-VILA: Cross-Modality Alignment for Large Language Model

- Encoder-Decoder Structure makes the alignment easier and promising[1]

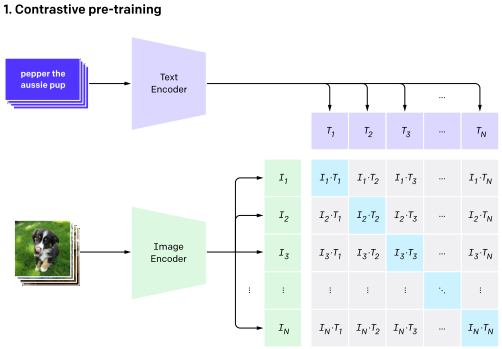
# The common sense reasoning ability of LLM



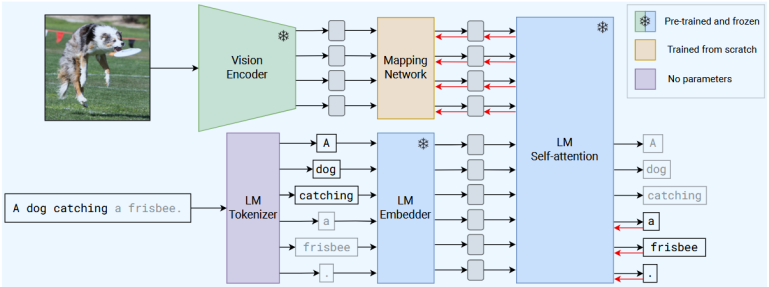

HapticFormers: Utilizing Transformers for Avocado Maturity Grading through Vision-based Tactile Assessment

- Using common sense to decide if the avocado is mature[2]

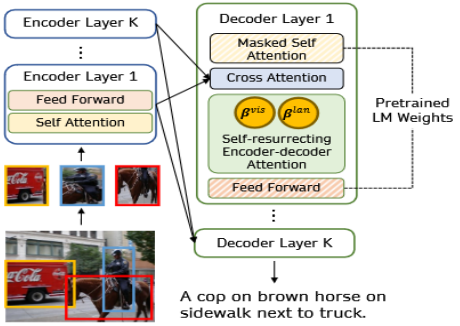
# Three different ways for alignment




Learning Transferable Visual Models From Natural Language Supervision



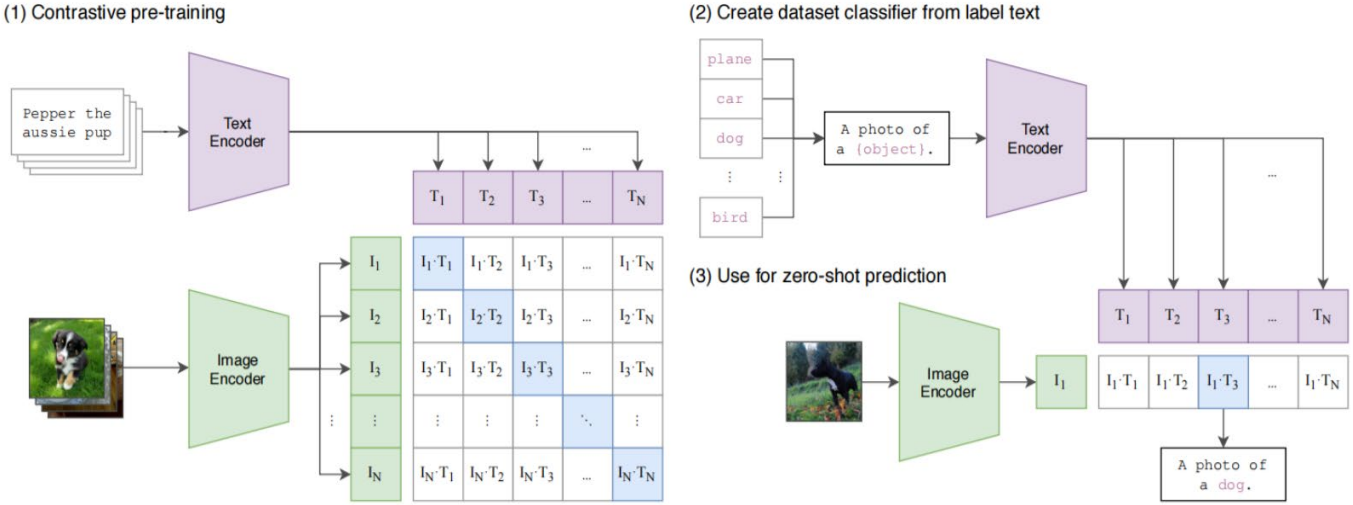

MAPL : Parameter-Efficient Adaptation of Unimodal Pre-Trained Models for Vision-Language Few-Shot Prompting




VisualGPT: Data-efficient Adaptation of Pretrained Language Models for Image Captioning

- Contrastive learning
  - CLIP[3]
  - CLOOB[4]
  - ALIGN[5]
  - DeCLIP[6]
- PrefixLM
  - MAPL[7]
  - Flamingo[8]
  - ClipCap[9]
- Cross attention
  - VisualGPT[10]
  - Flamingo[8]
  - VC-GPT[11]

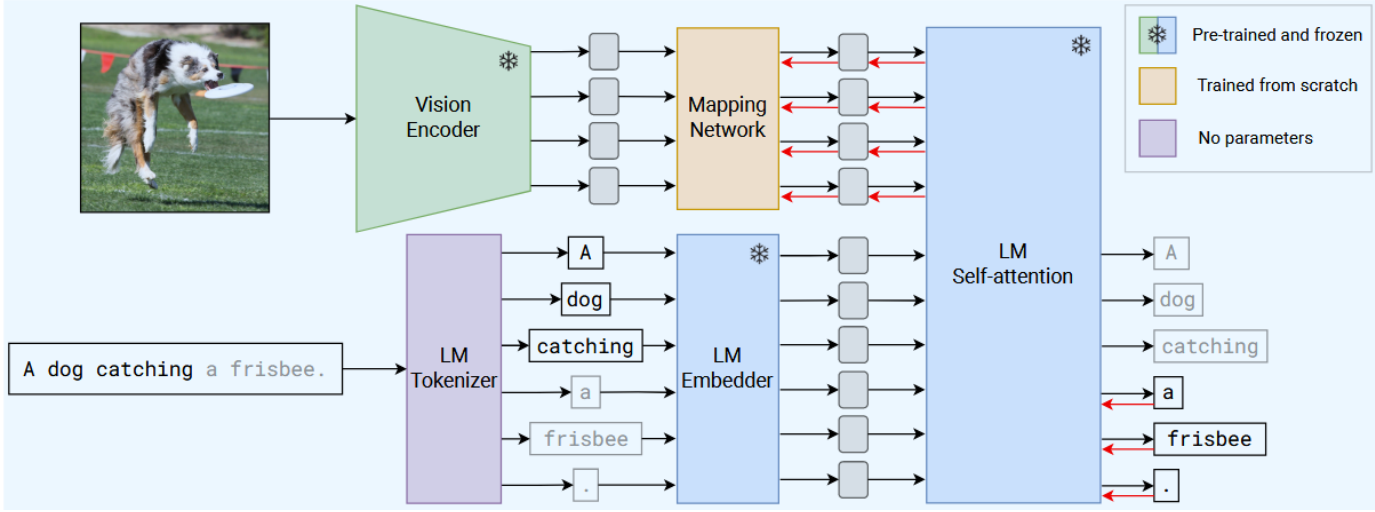
# Classification based on structure of alignment--Contrastive learning




Learning Transferable Visual Models From Natural Language Supervision

- CLIP[3]
- CLOOB[4]
- ALIGN[5]
- DeCLIP[6]
- Disadvantages: Training a new model from scratch with great amount of data

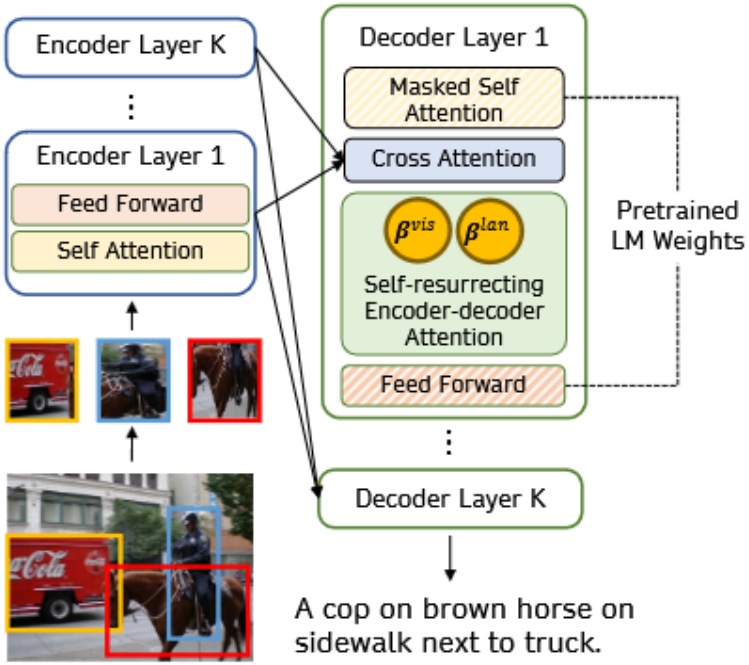
# Classification based on structure of alignment--PrefixLM




MAPL : Parameter-Efficient Adaptation of Unimodal Pre-Trained Models for Vision-Language Few-Shot Prompting

- MAPL[7]
- Flamingo[8]
- ClipCap[9]
- Advantage: Leveraging existed LLM
- Advatange: Adaptance for different modalities

# Classification based on structure of alignment—Cross Attention




VisualGPT: Data-efficient Adaptation of Pretrained Language Models for Image Captioning

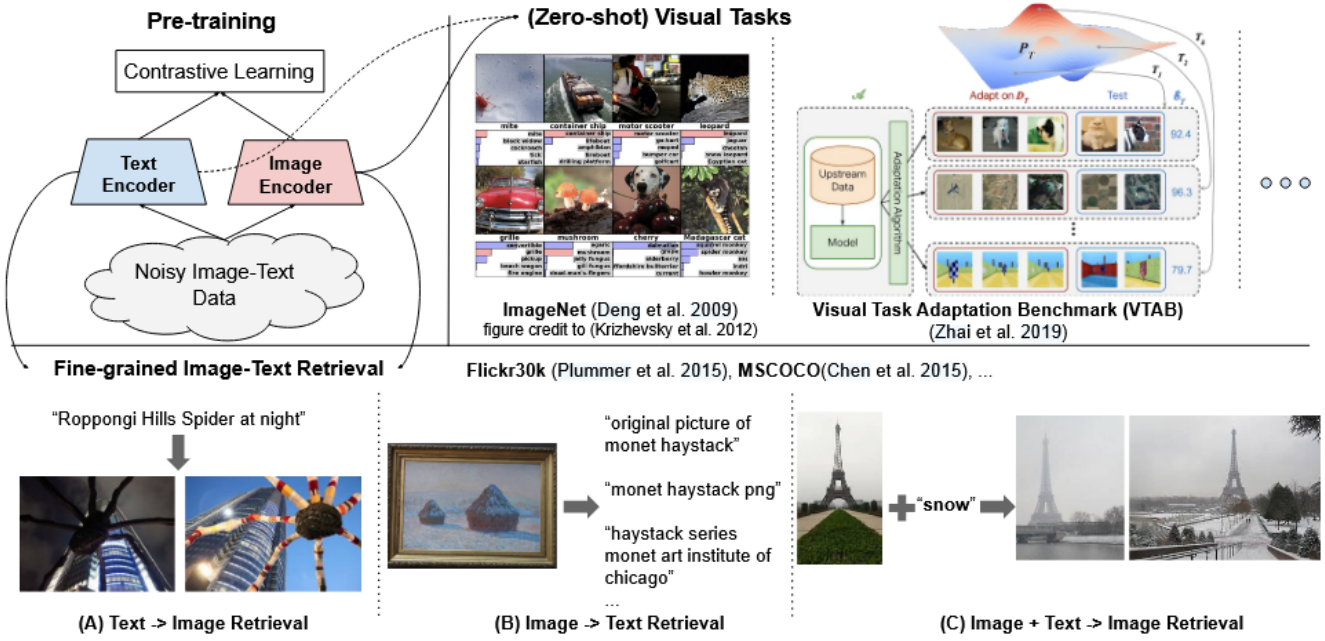
- VisualGPT[10]
- Flamingo[8]
- VC-GPT[11]

Balance the mixture of text generation capacity and visual information efficiently

# Classification based on Modality

- Non-robotic modalities
  - Image: CLIP[3], Flamingo[8], InternVL[12], PaLM-E[13]
  - Audio: AudioGPT[14], Qwen-Audio[15]
  - Any modalites: AnyGPT[16]
- Robot specific modalities (pose, joint angle, trajectory, states)
  - Point cloud: PointLLM[17], 3D-LLM[18]
  - Tactile: T3[19]
  - Image&Text&States: PaLM-E[13]
  - Trajectory:NL-trajectory-reshaper[20]
  - End Effector Pose: OpenVLA[21](MM-decoder)

# Classification based on Modality--Image

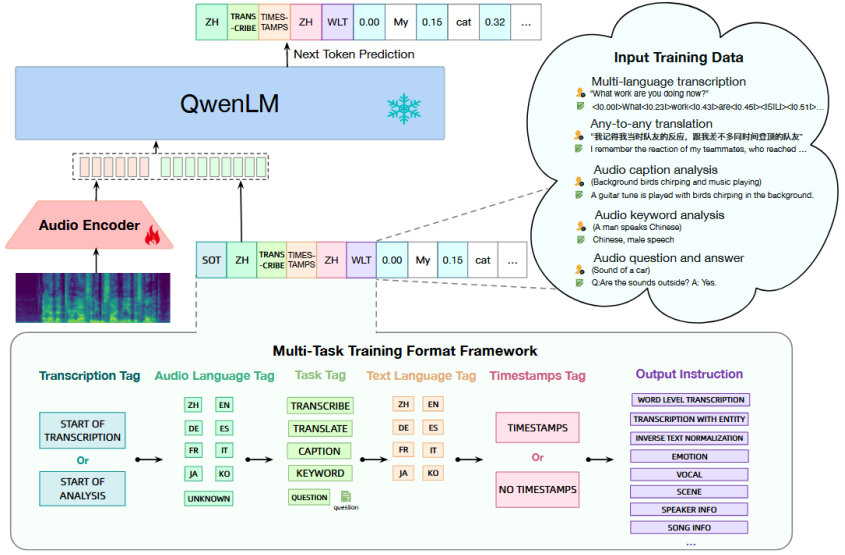



Learning Transferable Visual Models From Natural Language Supervision

CLIP, Flamingo, InternVL[12], PaLM-E[13]

CLIP: using Constrastive Learning

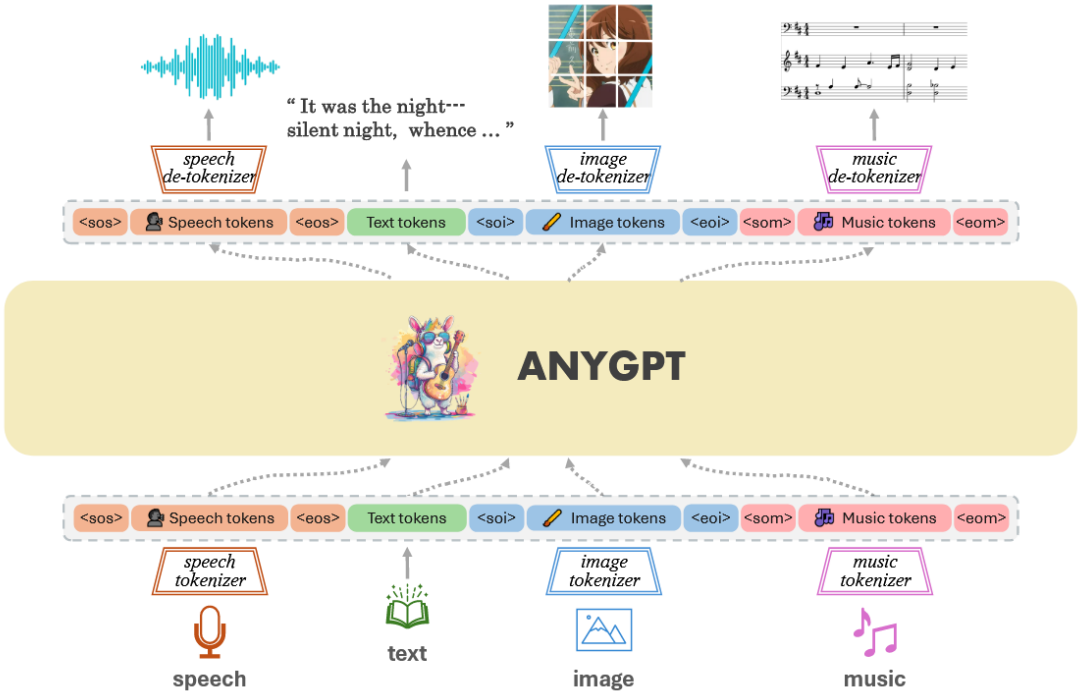
# Classification based on Modality--Audio




AudioGPT: Understanding and Generating Speech, Music, Sound, and Talking Head

AudioGPT[14], Qwen-Audio[15] PrefixLM Structure

# Classification based on Modality--Intergrated Modality

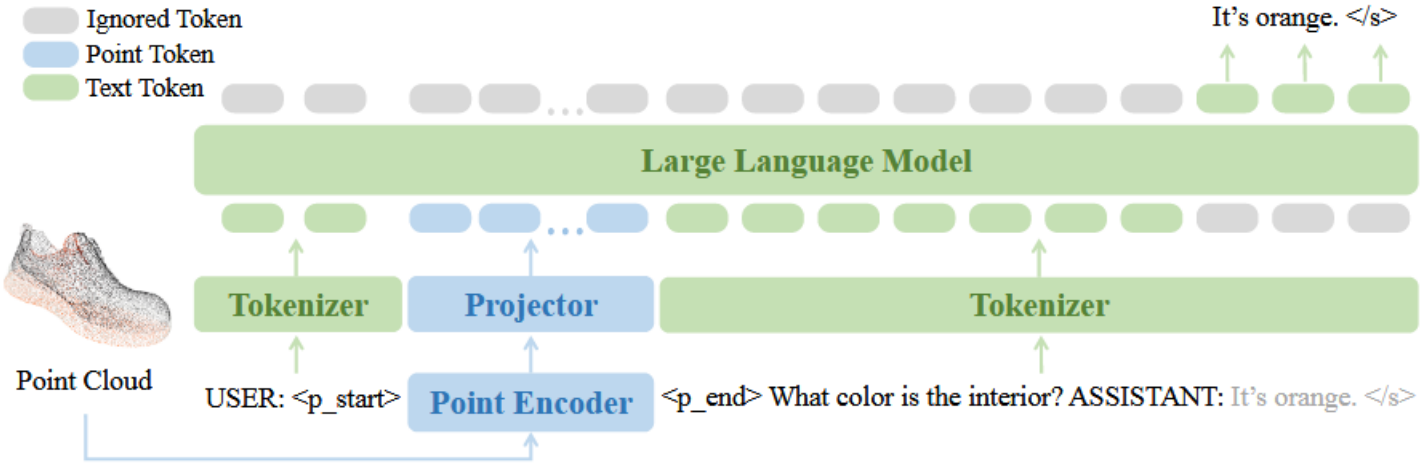



AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling

PrefixLM Structure

AnyGPT[16]

# Classification based on Modality--Point cloud

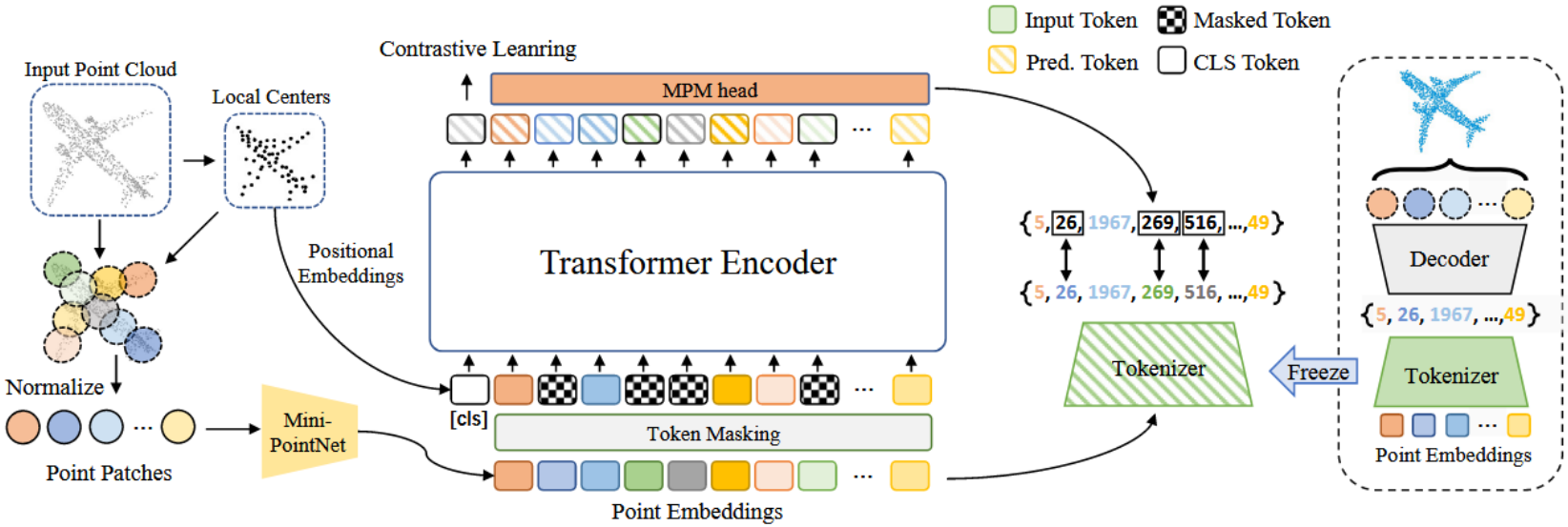



PointLLM: Empowering Large Language Models to Understand Point Clouds

PointLLM[17], 3D-LLM[18]

PrefixLM Structure
For 3D object captioning

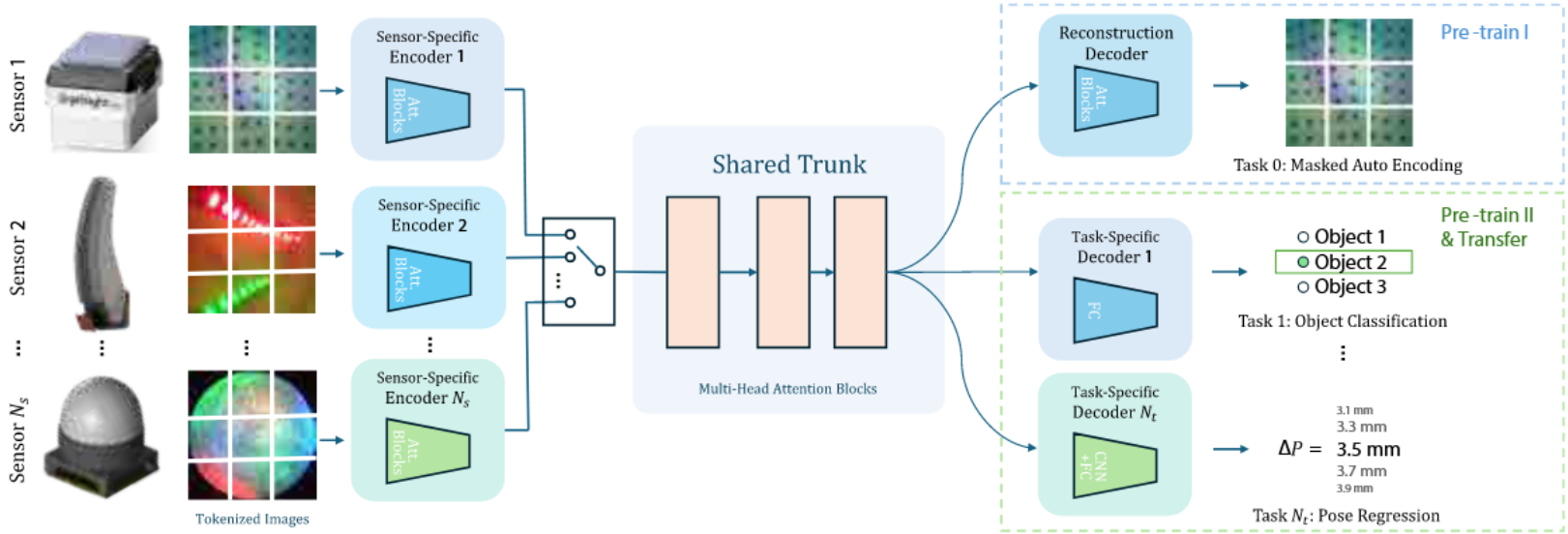
# Classification based on Modality--Point cloud




Point-BERT: Pre-training 3D Point Cloud Transformers with Masked Point Modeling

Point-BERT[22]

# Classification based on Modality--Tactile

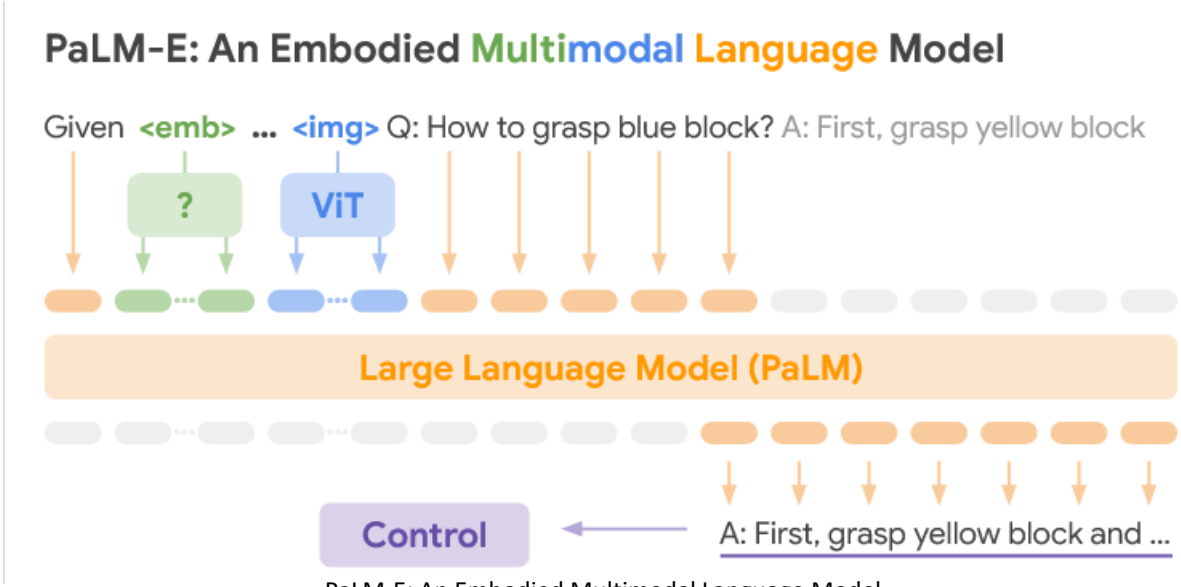



Transferable Tactile Transformers for Representation Learning Across Diverse Sensors and Tasks

T3[19]

Encoder-Decoder without LLM

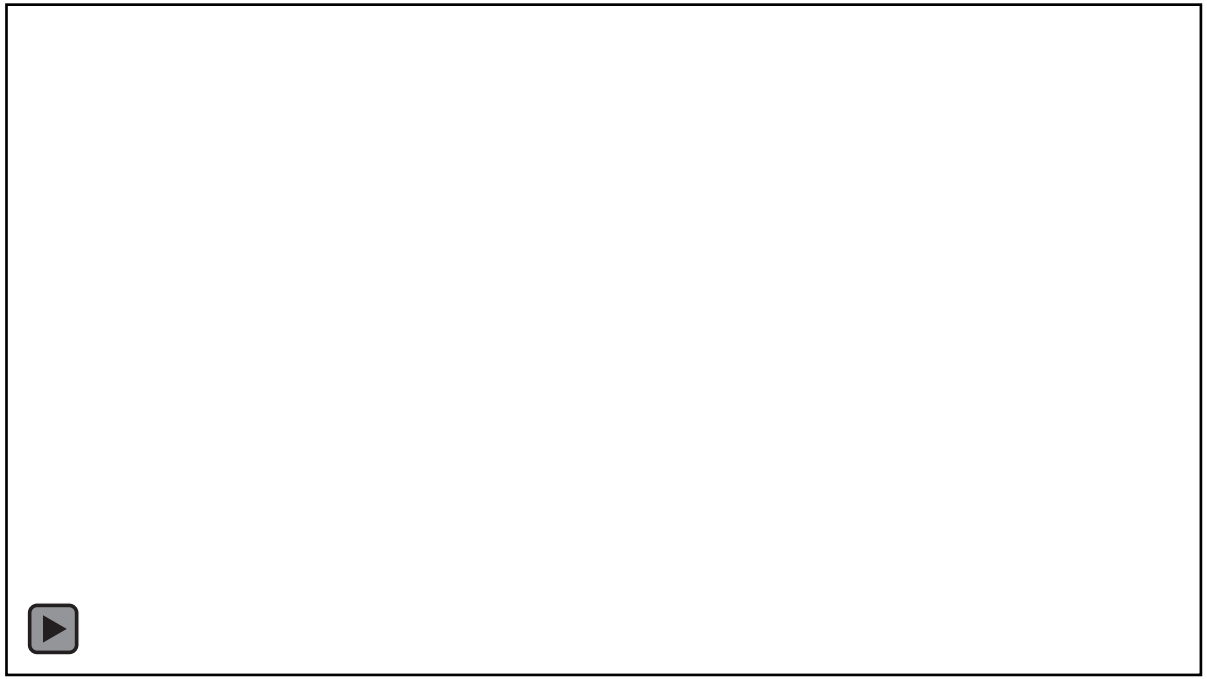
# Classification based on Modality--Image&Text&States




PaLM-E: An Embodied Multimodal Language Model

PaLM-E PrefixLM Structure

# Classification based on Modality--Image&Text&States

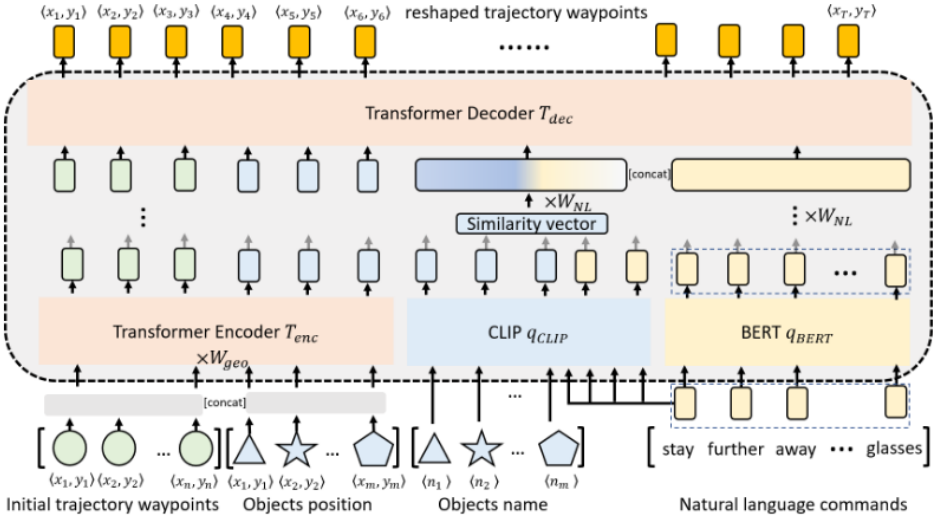


PaLM-E

PrefixLM  Structure

Shows Robustness
against disturbance

# Classification based on Modality--Trajectory

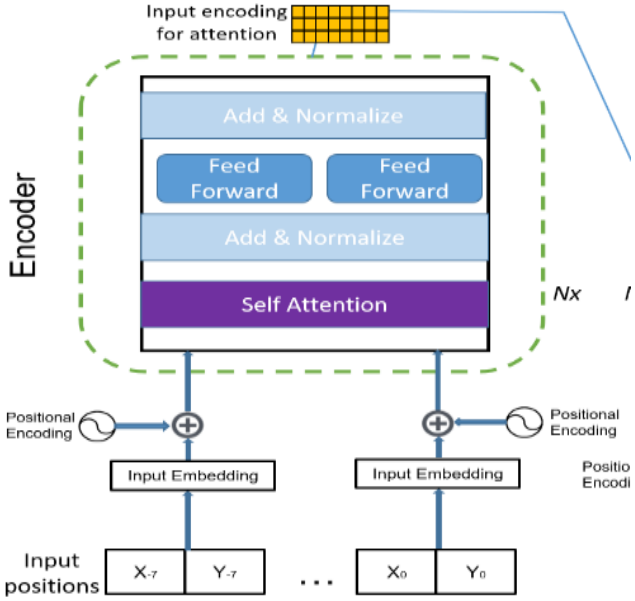



Reshaping Robot Trajectories Using Natural Language Commands: A Study of Multi-Modal Data Alignment Using Transformers

NL-trajectory-reshaper[20]

PrefixLM Structure

# Classification based on Modality--Trajectory

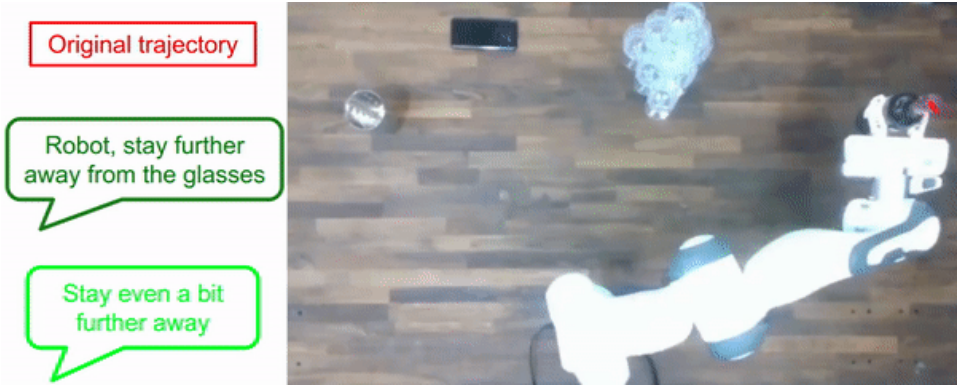


Transformer Networks for Trajectory Forecasting

Trajectory Encoder[23]

Introducing the Positional Encoding to add the time stamp for trajectory sequence

# Classification based on Modality--Trajectory

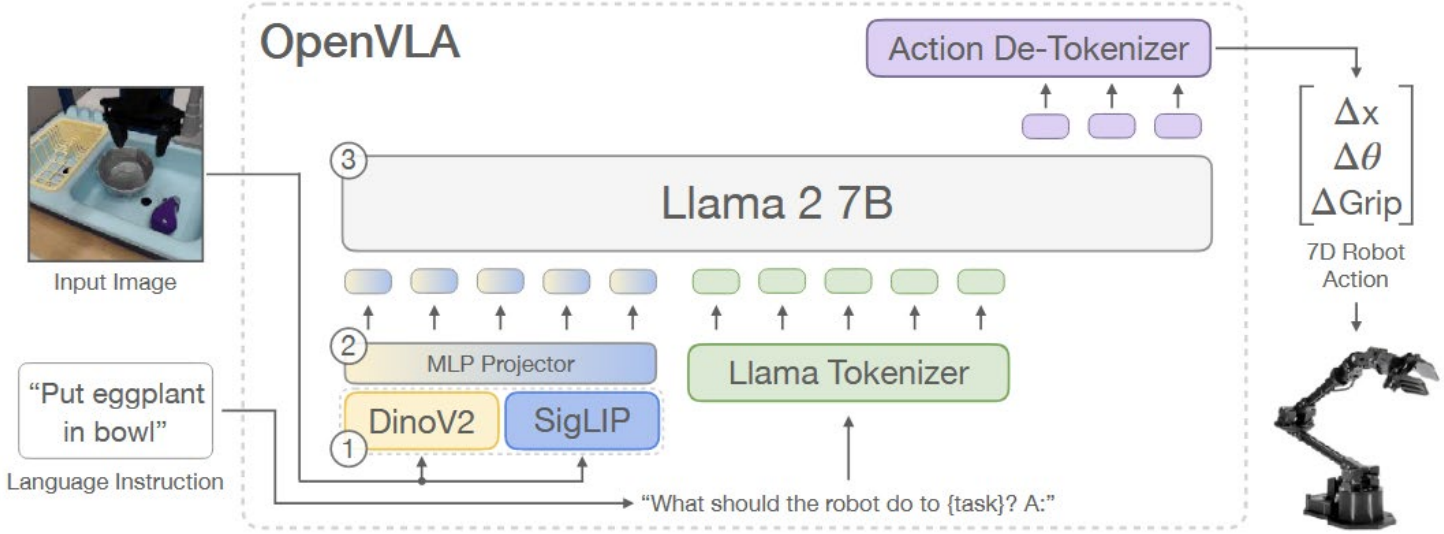



Reshaping Robot Trajectories Using Natural Language Commands: A Study of Multi-Modal Data Alignment Using Transformers

NL-trajectory-reshaper[19]

Function: Using natural language to reshape the trajectory, from collision

# Classification based on Modality—End Effector Pose

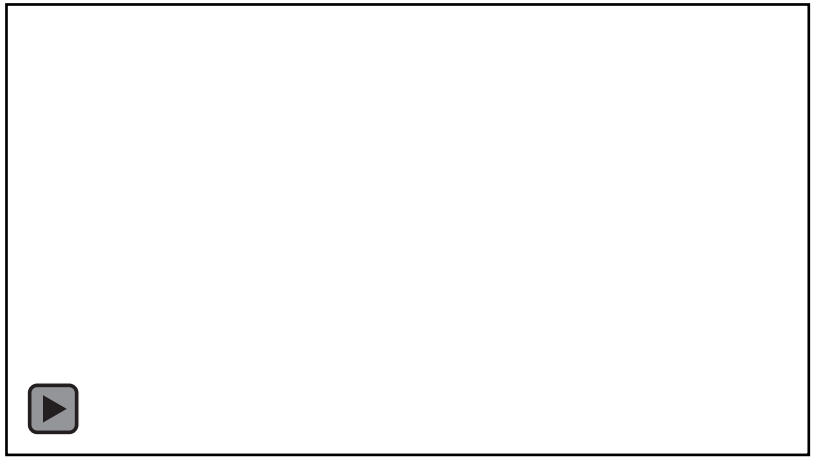


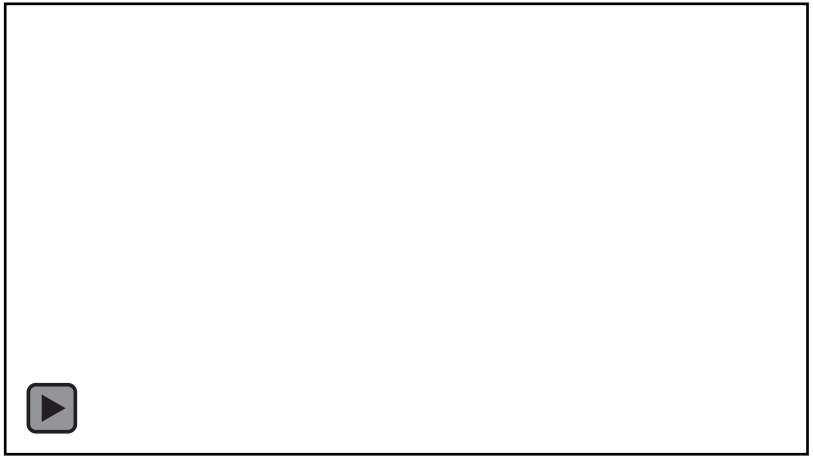

OpenVLA: An Open-Source Vision-Language-Action Model

OpenVLA[21]

PrefixLM Structure

# Classification based on Modality—End Effector Pose



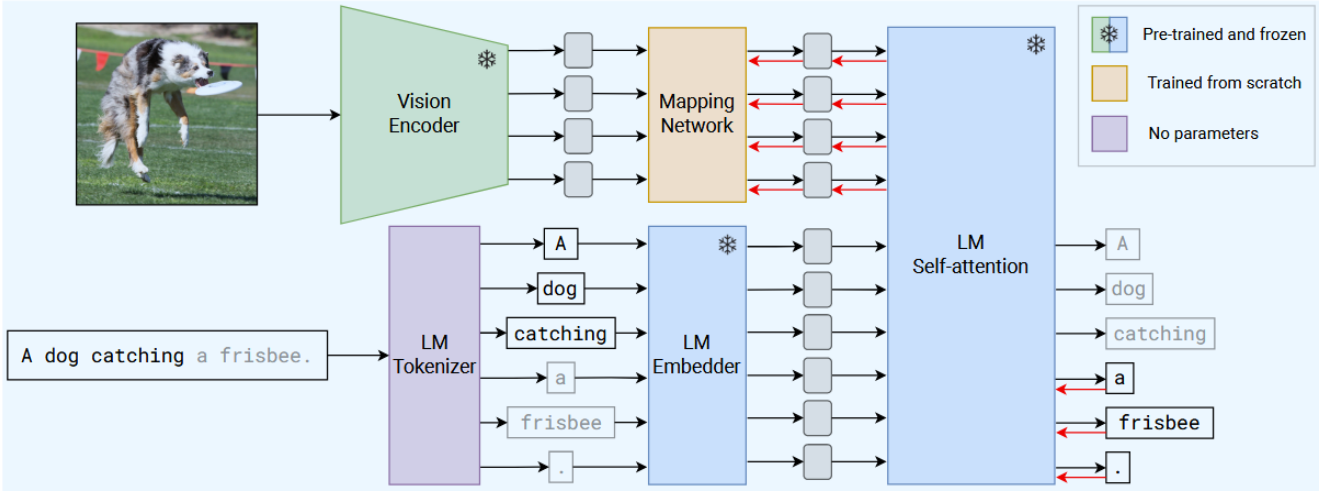


OpenVLA: An Open-Source Vision-Language-Action Model

It shows the ability for end-to-end Action generate

# Conclusion

- PrefixLM seems like the most promising method for robotic multimodal learning, leveraging the pre-trained LLM.
- More modalities can be aligned to improve the performance of robots.




MAPL : Parameter-Efficient Adaptation of Unimodal Pre-Trained Models for Vision-Language Few-Shot Prompting

# References


[1] Ye, Hanrong, et al. "X-VILA: Cross-Modality Alignment for Large Language Model." arXiv preprint arXiv:2405.19335 (2024).

[2] I. Fahmy, T. Hassan, I. Hussain, N. Werghi and L. Seneviratne, "HapticFormers: Utilizing Transformers for Avocado Maturity Grading through Vision-based Tactile Assessment," *2024 IEEE Haptics Symposium (HAPTICS)*, Long Beach, CA, USA, 2024, pp. 347-352, doi: 10.1109/HAPTICS59260.2024.10520836.

[3]A. Radford *et al.*, “Learning Transferable Visual Models From Natural Language Supervision” Feb. 26, 2021, *arXiv*: arXiv:2103.00020. doi: 10.48550/arXiv.2103.00020.

[4] A. Fürst *et al.*, “CLOOB: Modern Hopfield Networks with InfoLOOB Outperform CLIP,” Nov. 07, 2022, *arXiv*: arXiv:2110.11316. doi: 10.48550/arXiv.2110.11316.

[5]C. Jia et al., “Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision,” Jun. 11, 2021, arXiv: arXiv:2102.05918. doi: 10.48550/arXiv.2102.05918.

[6] Li Y, Liang F, Zhao L, et al. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm[J]. arXiv preprint arXiv:2110.05208, 2021.

[7] O. Mañas, P. Rodriguez, S. Ahmadi, A. Nematzadeh, Y. Goyal, and A. Agrawal, “MAPL: Parameter-Efficient Adaptation of Unimodal Pre-Trained Models for Vision-Language Few-Shot Prompting,” Mar. 15, 2023, arXiv: arXiv:2210.07179. doi: 10.48550/arXiv.2210.07179.

[8] J.-B. Alayrac et al., “Flamingo: a Visual Language Model for Few-Shot Learning,” Nov. 15, 2022, arXiv: arXiv:2204.14198. doi: 10.48550/arXiv.2204.14198.

[9] Mokady, Ron, Amir Hertz, and Amit H. Bermano. "Clipcap: Clip prefix for image captioning." arXiv preprint arXiv:2111.09734 (2021).

# References


[10] Chen, Jun, et al. "Visualgpt: Data-efficient adaptation of pretrained language models for image captioning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[11] Luo, Ziyang, et al. "A frustratingly simple approach for end-to-end image captioning." arXiv preprint arXiv:2201.12723 (2022).

[12] Z. Chen et al., “InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks,” Jan. 15, 2024, arXiv: arXiv:2312.14238. doi: 10.48550/arXiv.2312.14238.

[13] D. Driess et al., “PaLM-E: An Embodied Multimodal Language Model,” Mar. 06, 2023, arXiv: arXiv:2303.03378. doi: 10.48550/arXiv.2303.03378.

[14] R. Huang et al., “AudioGPT: Understanding and Generating Speech, Music, Sound, and Talking Head,” Apr. 25, 2023, arXiv: arXiv:2304.12995. doi: 10.48550/arXiv.2304.12995. arXiv:2308.16911. doi: 10.48550/arXiv.2308.16911.

[15] Y. Chu et al., “Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models,” Dec. 21, 2023, arXiv: arXiv:2311.07919. doi: 10.48550/arXiv.2311.07919.

[16] J. Zhan et al., “AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling,” Mar. 07, 2024, arXiv: arXiv:2402.12226. doi: 10.48550/arXiv.2402.12226.

[17] R. Xu, X. Wang, T. Wang, Y. Chen, J. Pang, and D. Lin, “PointLLM: Empowering Large Language Models to Understand Point Clouds,” Sep. 06, 2024, arXiv:

# References

[18] Hong, Yining, et al. "3d-llm: Injecting the 3d world into large language models." Advances in Neural Information Processing Systems 36 (2023): 20482-20494.

[19] J. Zhao, Y. Ma, L. Wang, and E. H. Adelson, “Transferable Tactile Transformers for Representation Learning Across Diverse Sensors and Tasks,” Oct. 06, 2024, arXiv: arXiv:2406.13640. doi: 10.48550/arXiv.2406.13640.A. [20] Bucker, L. Figueredo, S. Haddadin, A. Kapoor, S. Ma, and R. Bonatti, “Reshaping Robot Trajectories Using Natural Language Commands: A Study of Multi-Modal Data Alignment Using Transformers,” Mar. 25, 2022, arXiv: arXiv:2203.13411. doi: 10.48550/arXiv.2203.13411.

[21] M. J. Kim et al., “OpenVLA: An Open-Source Vision-Language-Action Model,” Sep. 05, 2024, arXiv: arXiv:2406.09246. doi: 10.48550/arXiv.2406.09246.

[22] Yu X, Tang L, Rao Y, et al. Point-bert: Pre-training 3d point cloud transformers with masked point modeling[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 19313-19322.

[23] Giuliari F, Hasan I, Cristani M, et al. Transformer networks for trajectory forecasting[C]//2020 25th international conference on pattern recognition (ICPR). IEEE, 2021: 10335-10342.