Artificial Intelligence

Computer Vision and Robotics

1141AI08
MBA, IM, NTPU (M5276) (Fall 2025)
Tue 2, 3, 4 (9:10-12:00) (B3F17)

https://meet.google.com/paj-zhhj-mya

NVIDIA University Ambassador
Certified Instructor
aws educate Cloud Ambassador 2020 Cohort
aws academy Accredited Educator
aws certified Cloud Practitioner
aws certified Solutions Architect Associate

Min-Yuh Day, Ph.D,
Professor and Director
Institute of Information Management, National Taipei University
https://web.ntpu.edu.tw/~myday

2025-11-25

Syllabus

Week Date Subject/Topics

1 2025/09/09 Introduction to Artificial Intelligence

2 2025/09/16 Artificial Intelligence and Intelligent Agents;
Problem Solving

3 2025/09/23 Knowledge, Reasoning and Knowledge Representation;
Uncertain Knowledge and Reasoning

4 2025/09/30 Case Study on Artificial Intelligence I

5 2025/10/07 Machine Learning: Supervised and Unsupervised Learning;
The Theory of Learning and Ensemble Learning

Syllabus

Week Date Subject/Topics

6 2025/10/14 NVIDIA Fundamentals of Deep Learning I:
Deep Learning; Neural Networks

7 2025/10/21 NVIDIA Fundamentals of Deep Learning II:
Convolutional Neural Networks;
Data Augmentation and Deployment

8 2025/10/28 Self-Learning

9 2025/11/04 Midterm Project Report

10 2025/11/11 NVIDIA Fundamentals of Deep Learning III:
Pre-trained Models; Natural Language Processing

Syllabus

Week Date Subject/Topics

11 2025/11/18 Case Study on Artificial Intelligence II

12 2025/11/25 Computer Vision and Robotics

13 2025/12/02 Generative AI, Agentic AI, and Physical AI

14 2025/12/09 Philosophy and Ethics of AI and the Future of AI

15 2025/12/16 Final Project Report I

16 2025/12/23 Final Project Report II

Computer Vision
and
Robotics

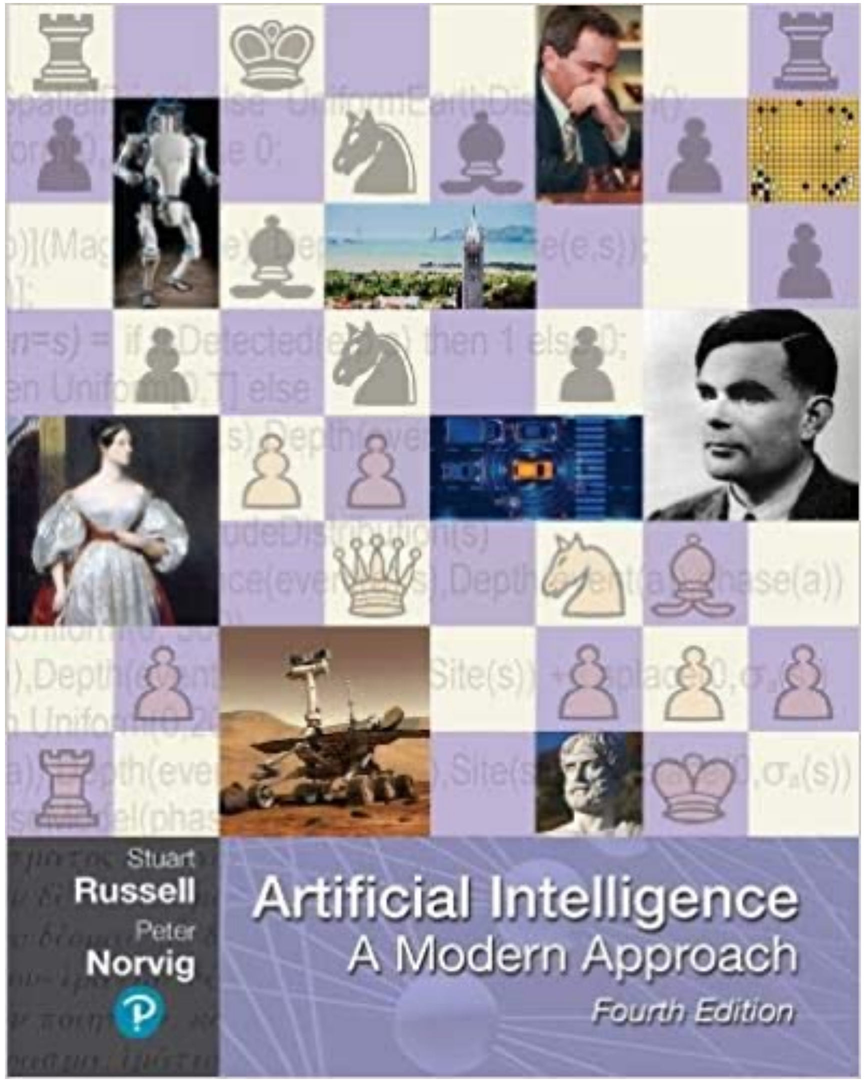
Outline

- Computer Vision
  - Classifying Images
  - Detecting Objects
  - The 3D World
- Robotics
  - Robotic Perception
  - Planning and Control
  - Planning Uncertain Movements
  - Reinforcement Learning in Robotics

Stuart Russell and Peter Norvig (2020),

Artificial Intelligence: A Modern Approach,

4th Edition, Pearson



Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

https://www.amazon.com/Artificial-Intelligence-A-Modern-Approach/dp/0134610997/

Artificial Intelligence: A Modern Approach

1. Artificial Intelligence
2. Problem Solving
3. Knowledge and Reasoning
4. Uncertain Knowledge and Reasoning
5. Machine Learning
6. Communicating, Perceiving, and Acting
7. Philosophy and Ethics of AI

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Artificial Intelligence: Communicating, perceiving, and acting

Artificial Intelligence:

6. Communicating, Perceiving, and Acting

- Natural Language Processing
- Deep Learning for Natural Language Processing
- Computer Vision
- Robotics

Artificial Intelligence:
Computer Vision

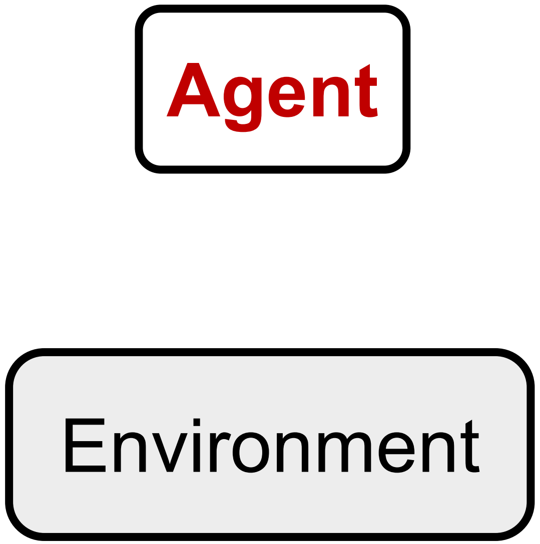
- Image Formation
- Simple Image Features
- Classifying Images
- Detecting Objects
- The 3D World
- Using Computer Vision

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Artificial Intelligence: Robotics

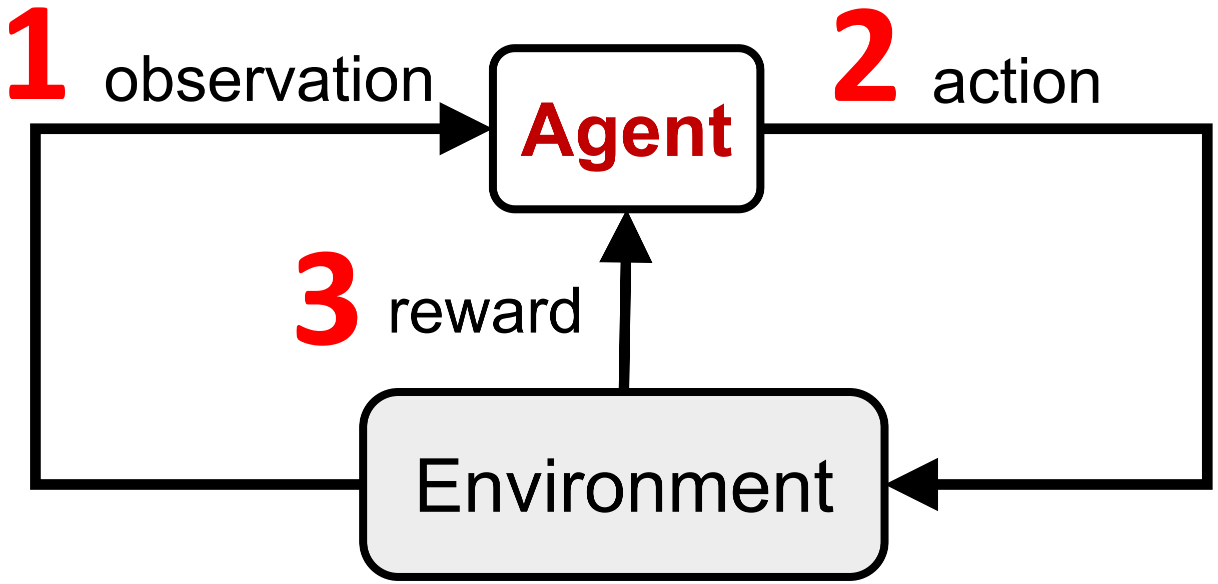
- Robots
- Robotic Perception
- Planning and Control
- Planning Uncertain Movements
- Reinforcement Learning in Robotics
- Humans and Robots

Reinforcement Learning (DL)



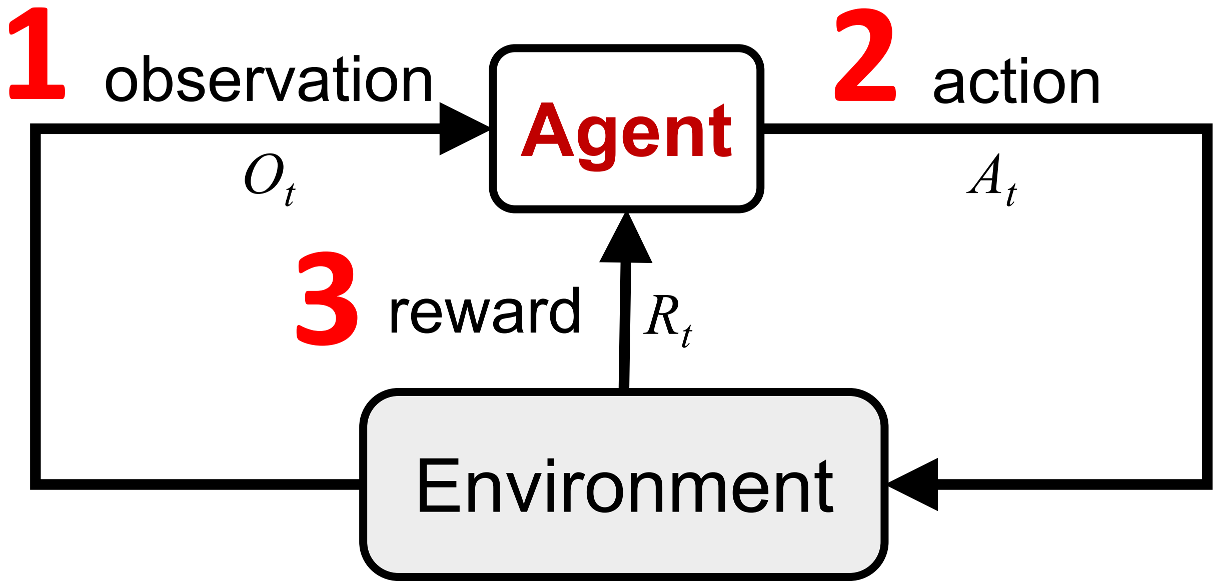
Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Reinforcement Learning (DL)



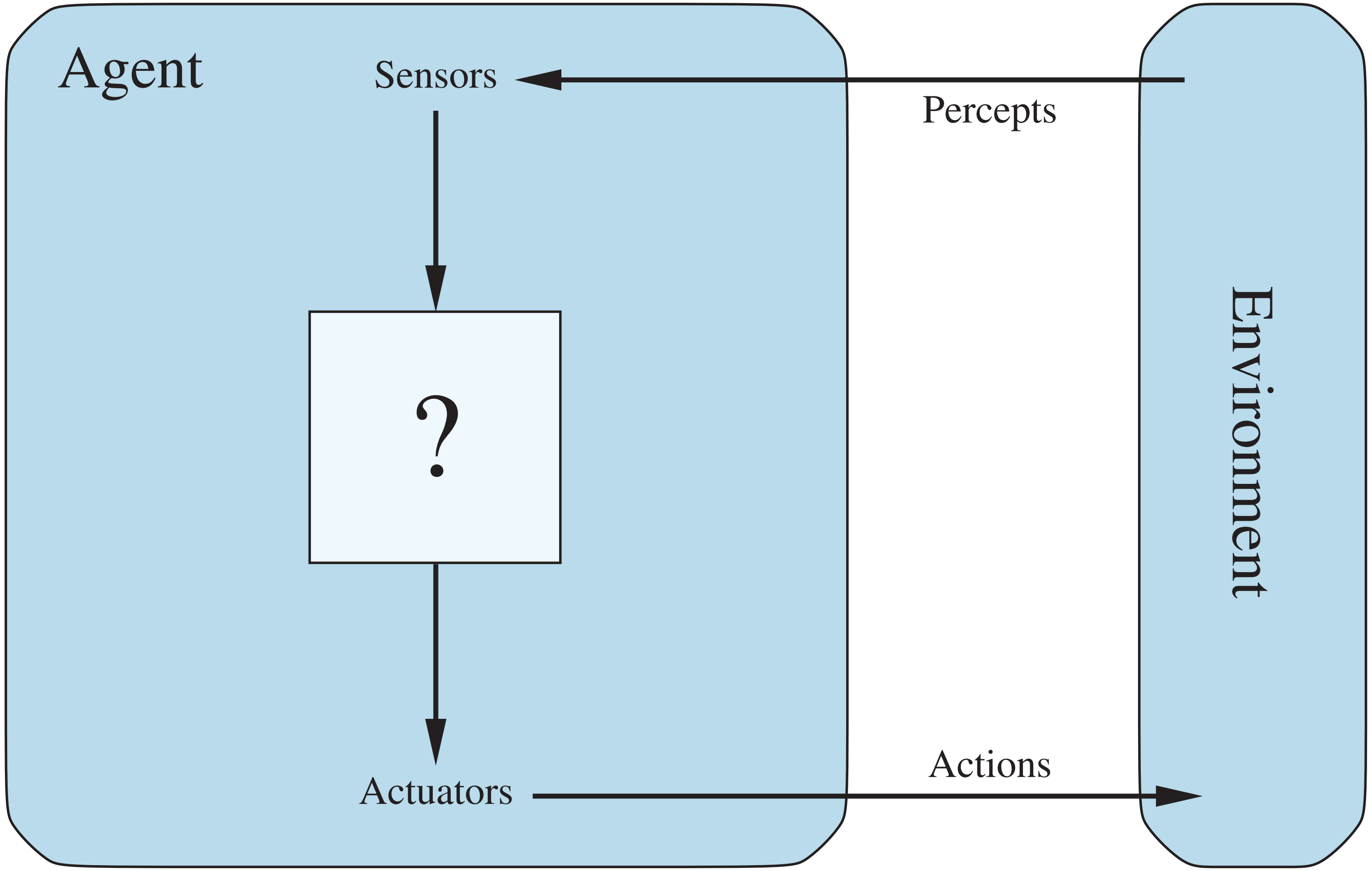
Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Reinforcement Learning (DL)



Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Agents interact with environments through sensors and actuators



Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

AI Acting Humanly: The Turing Test Approach

(Alan Turing, 1950)

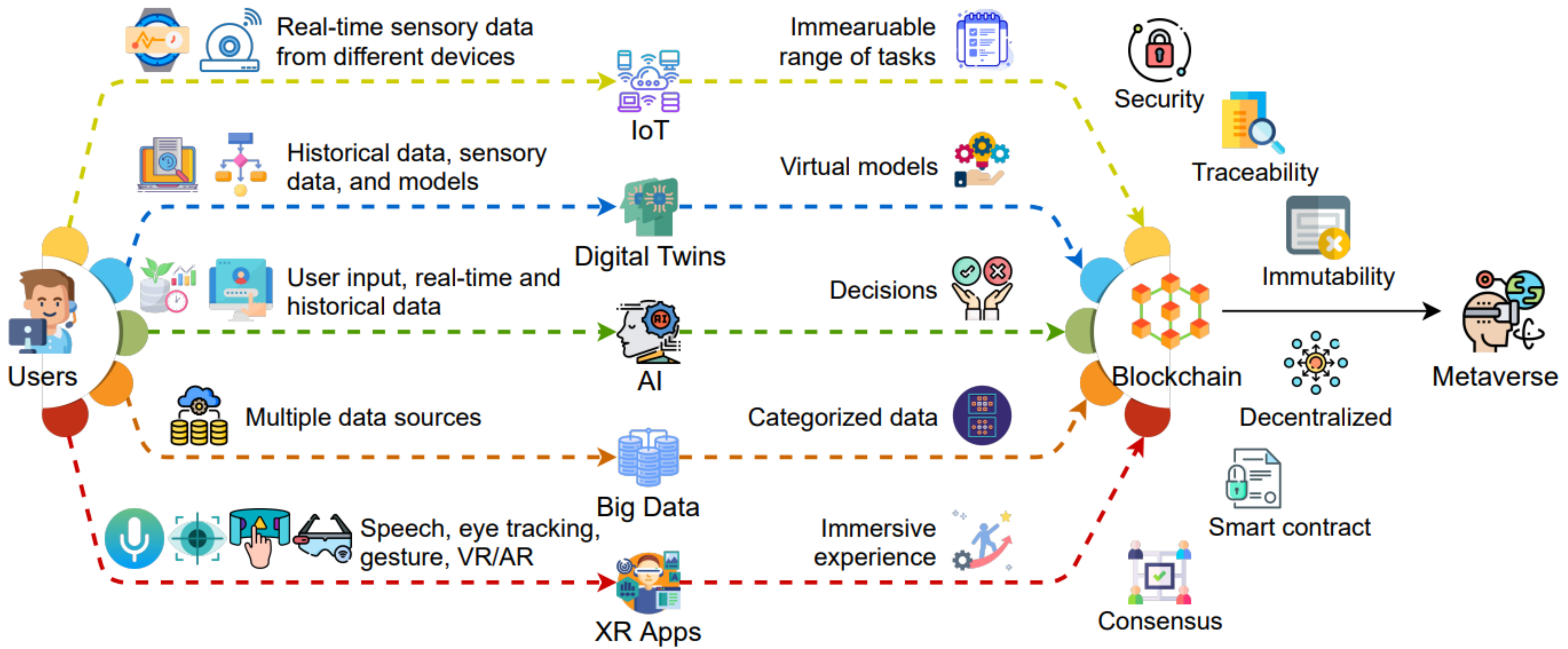
- Knowledge Representation
- Automated Reasoning
- Machine Learning (ML)
  - Deep Learning (DL)
- Computer Vision (Image, Video)
- Natural Language Processing (NLP)
- Robotics

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Artificial Intelligence: Communicating, Perceiving, and Acting

- **Computer vision and speech recognition**
  - **to perceive the world**
- **Robotics**
  - **to manipulate objects and move about**

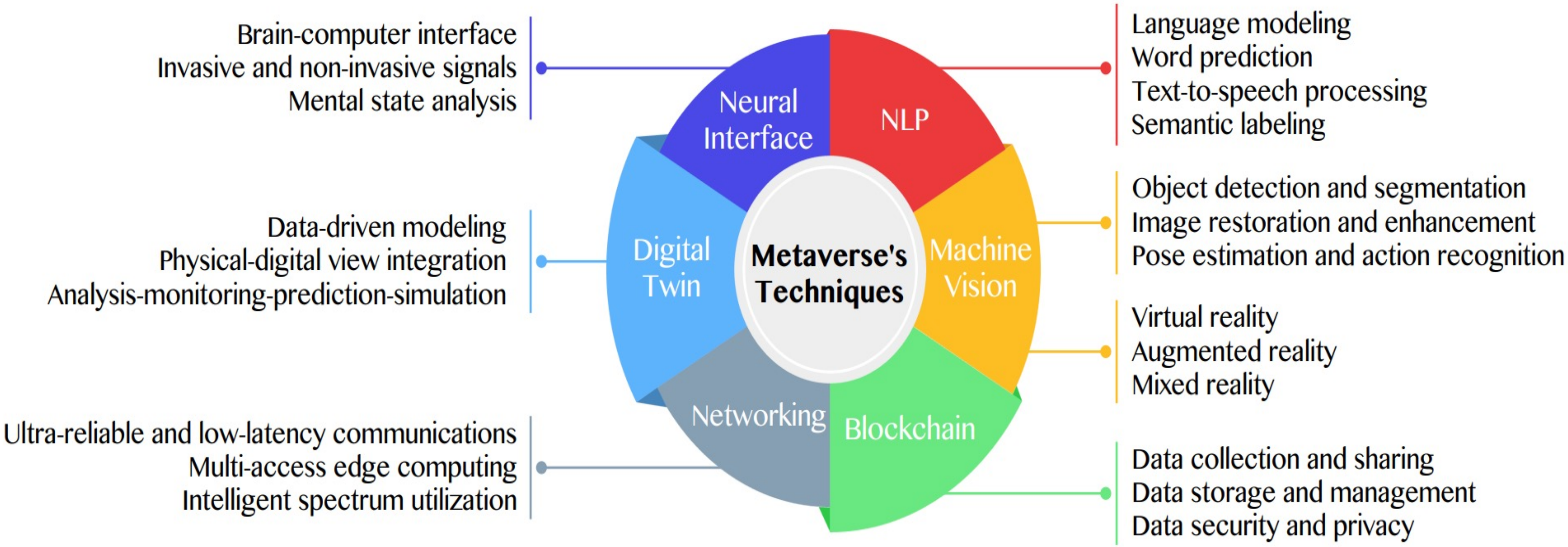
Key Enabling Technologies of the Metaverse



Source: Gadekallu, Thippa Reddy, Thien Huynh-The, Weizheng Wang, Gokul Yenduri, Pasika Ranaweera, Quoc-Viet Pham, Daniel Benevides da Costa, and Madhusanka Liyanage (2022). "Blockchain for the Metaverse: A Review." arXiv preprint arXiv:2203.09738..

Primary Technical Aspects in the Metaverse

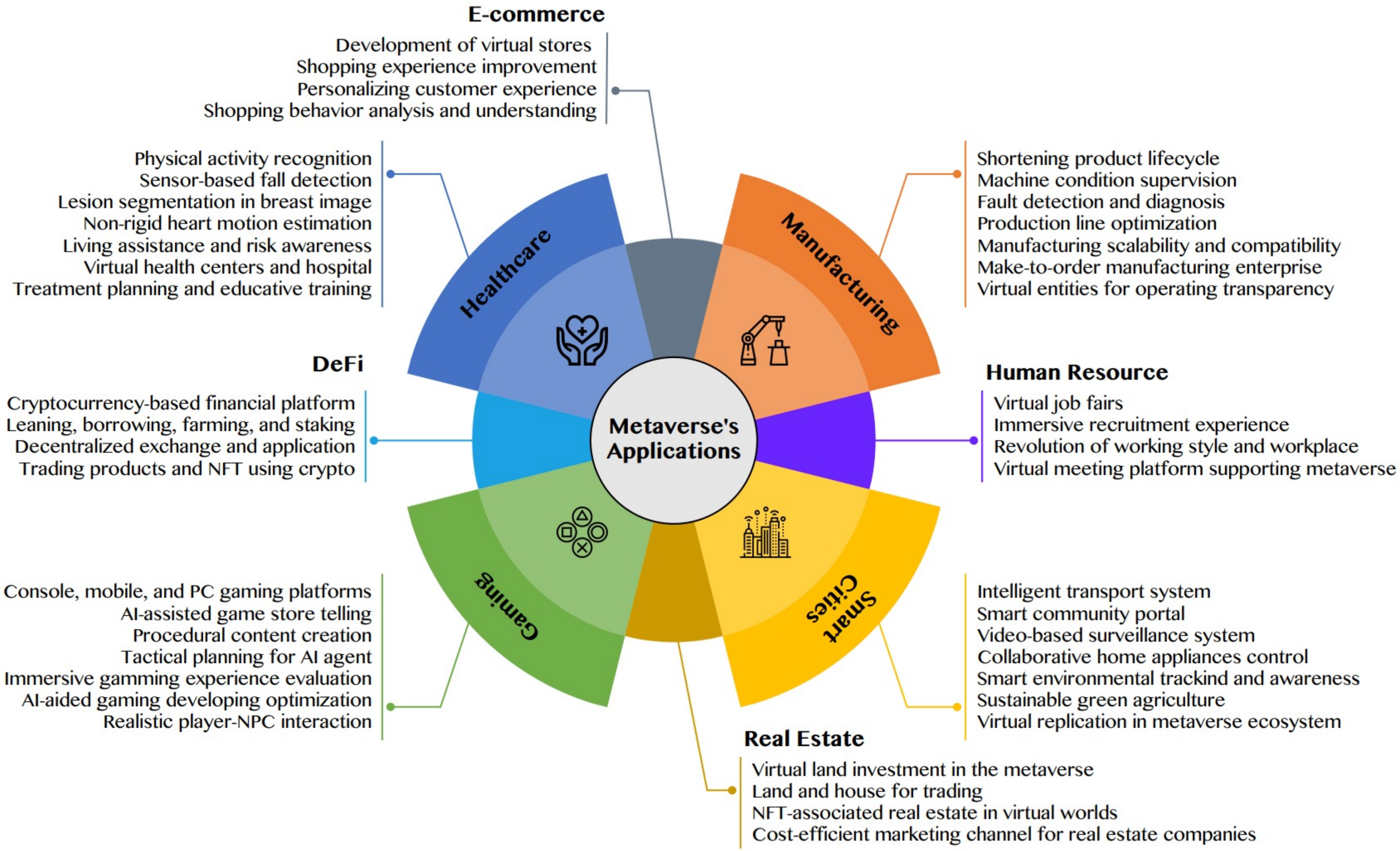
AI with ML algorithms and DL architectures is advancing the user experience in the virtual world



Source: Huynh-The, Thien, Quoc-Viet Pham, Xuan-Qui Pham, Thanh Thi Nguyen, Zhu Han, and Dong-Seong Kim (2022). "Artificial Intelligence for the Metaverse: A Survey." arXiv preprint arXiv:2202.10336.

AI for the Metaverse in the Application Aspects

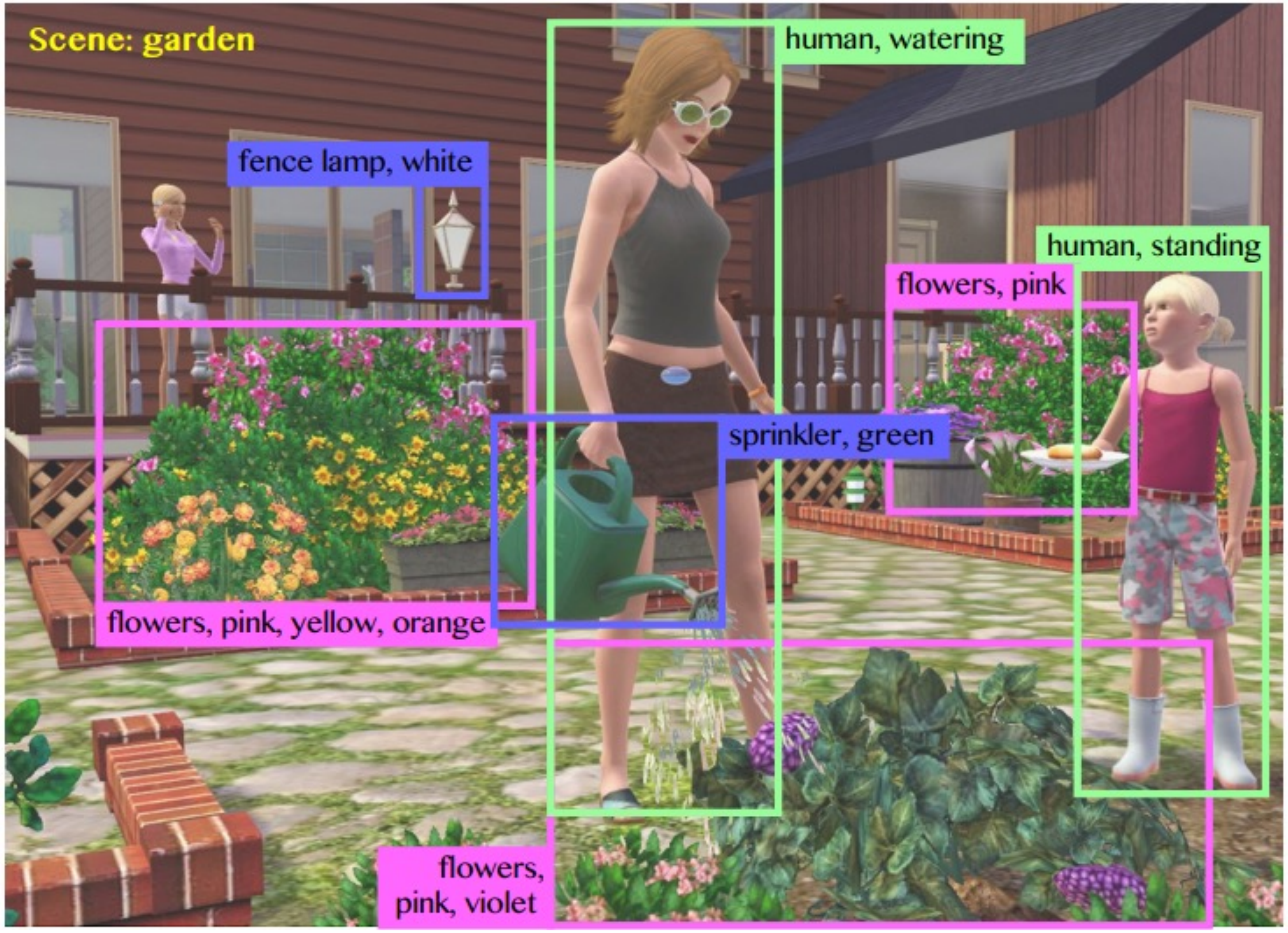
healthcare, manufacturing, smart cities, gaming
E-commerce, human resources, real estate, and DeFi



Source: Huynh-The, Thien, Quoc-Viet Pham, Xuan-Qui Pham, Thanh Thi Nguyen, Zhu Han, and Dong-Seong Kim (2022). "Artificial Intelligence for the Metaverse: A Survey." arXiv preprint arXiv:2202.10336.

Computer Vision in the Metaverse

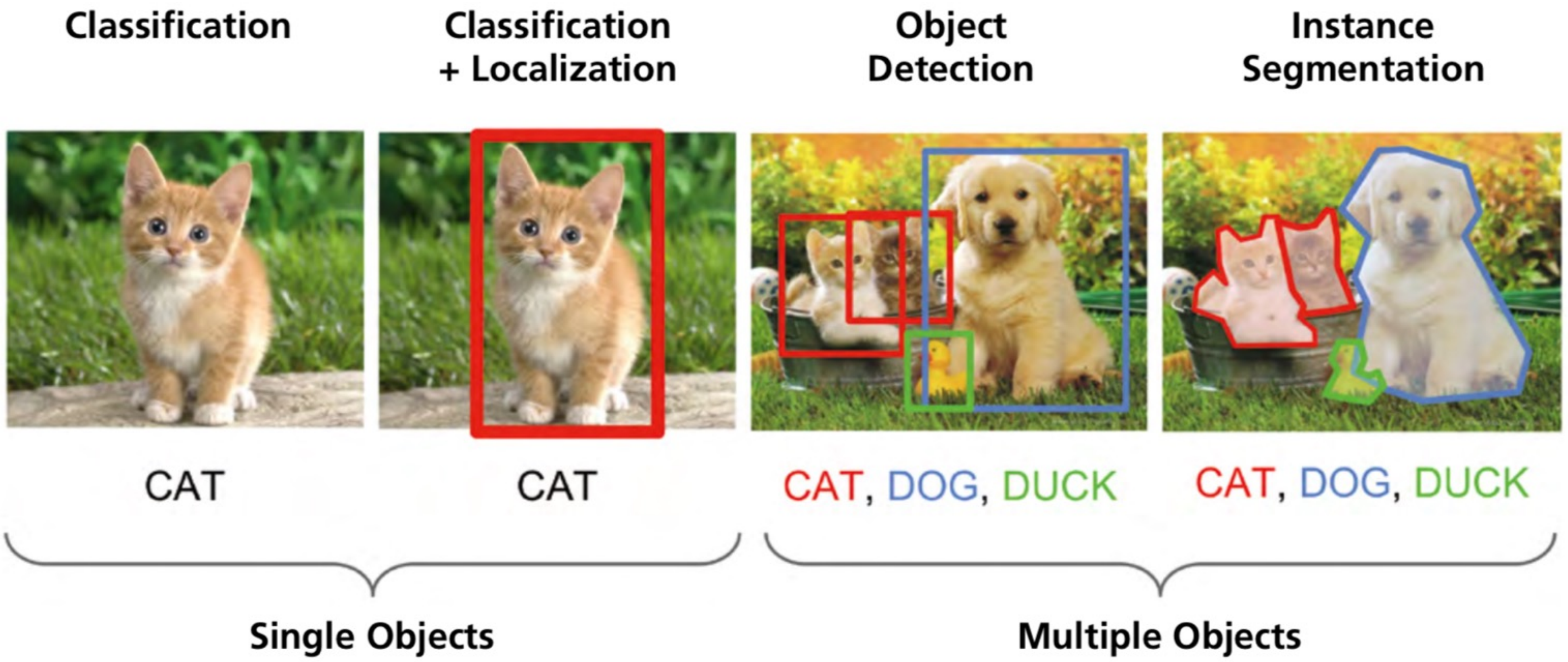
with scene understanding, object detection, and human action/activity recognition



Source: Huynh-The, Thien, Quoc-Viet Pham, Xuan-Qui Pham, Thanh Thi Nguyen, Zhu Han, and Dong-Seong Kim (2022). "Artificial Intelligence for the Metaverse: A Survey." arXiv preprint arXiv:2202.10336.

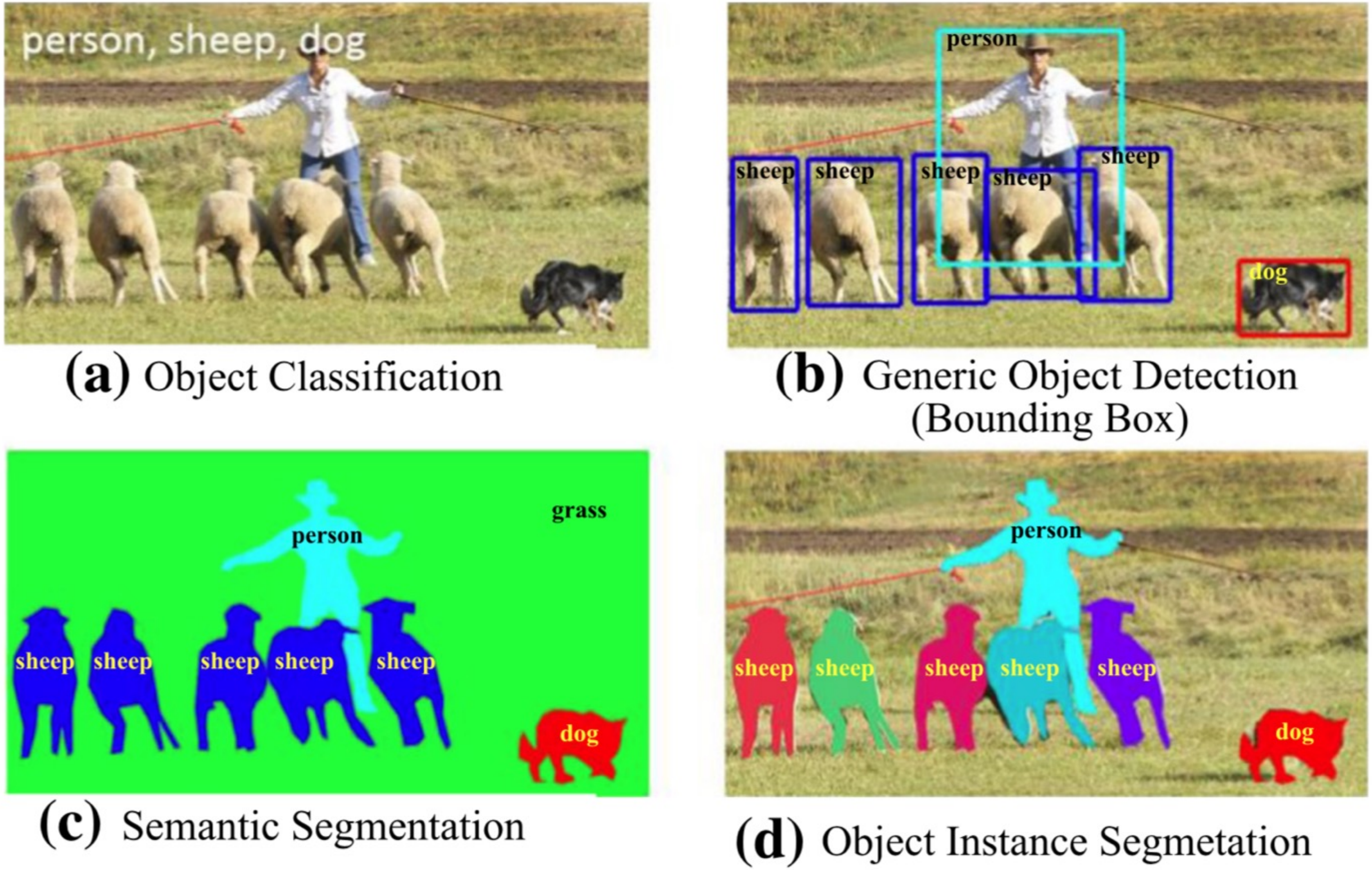
Computer Vision

Computer Vision:
Image Classification, Object Detection, Object Instance Segmentation



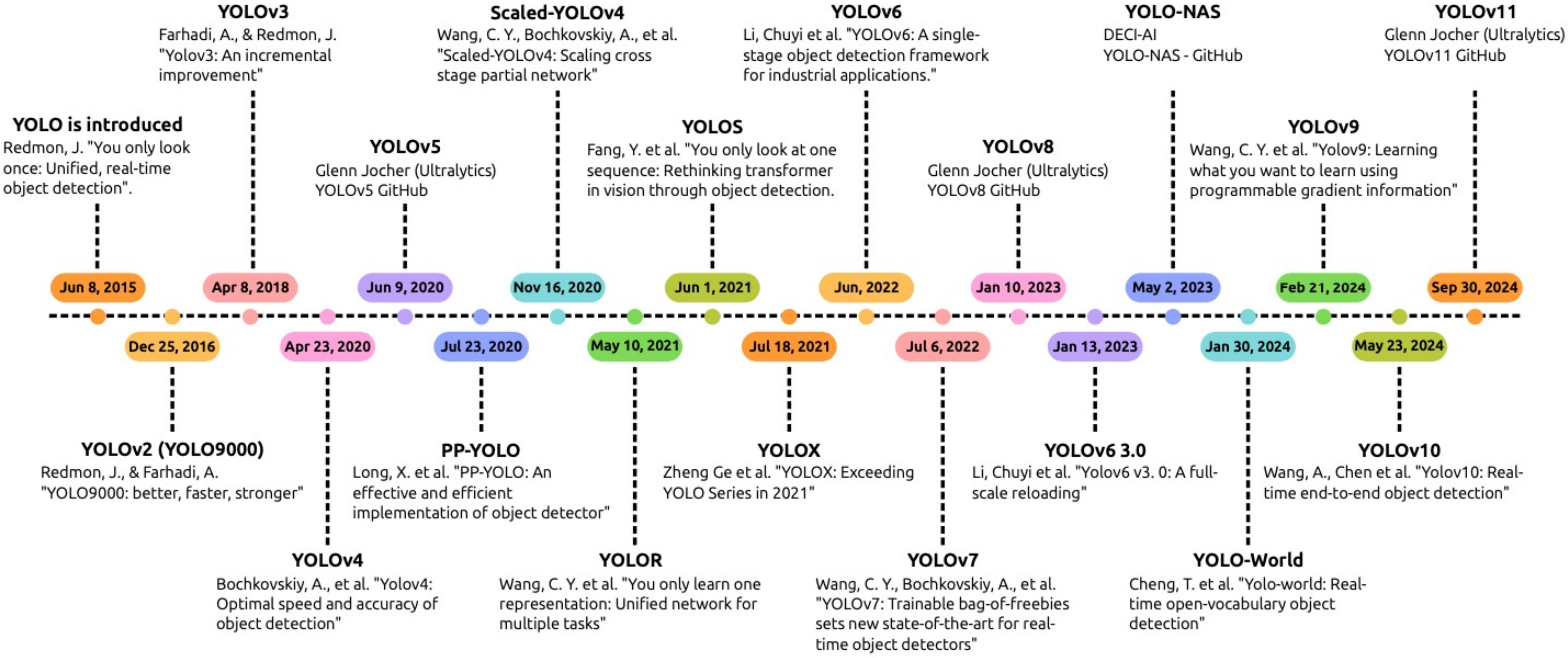
Source: DHL (2018), Artificial Intelligence in Logistics,
http://www.globalhha.com/doclib/data/upload/doc_con/5e50c53c5bf67.pdf/

Computer Vision: Object Detection



Source: Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. "Deep learning for generic object detection: A survey." International journal of computer vision 128, no. 2 (2020): 261-318.

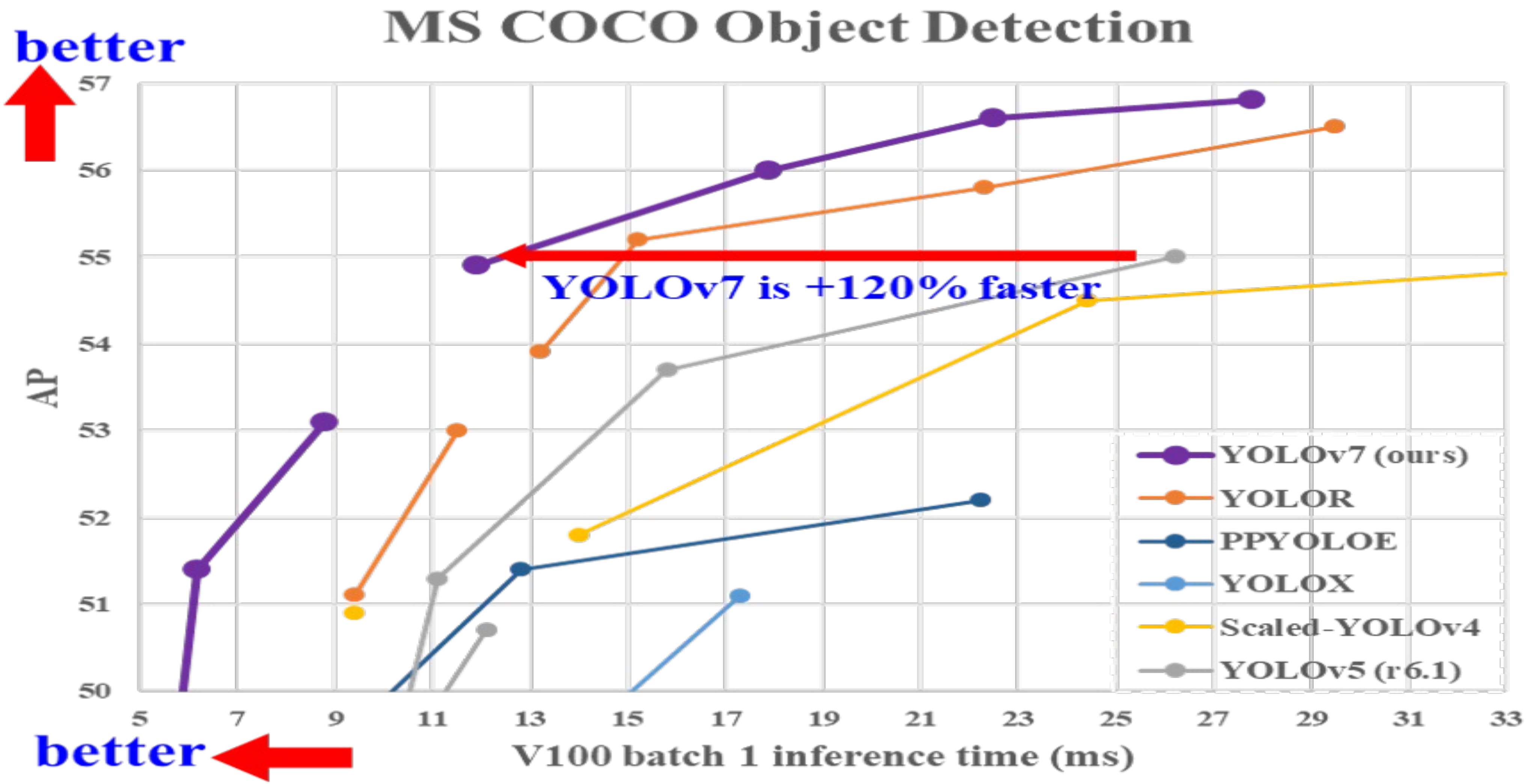
Evolution of YOLO Algorithms (YOLOv7, YOLOv8, YOLOv9, YOLOv10, YOLOv11)



Source: Nidhal Jegham, Chan Young Koh, Marwan Abdelatti, and Abdeltawab Hendawi. (2024) "Evaluating the Evolution of YOLO (You Only Look Once) Models: A Comprehensive Benchmark Study of YOLO11 and Its Predecessors." arXiv preprint arXiv:2411.00201.

YOLOv7:

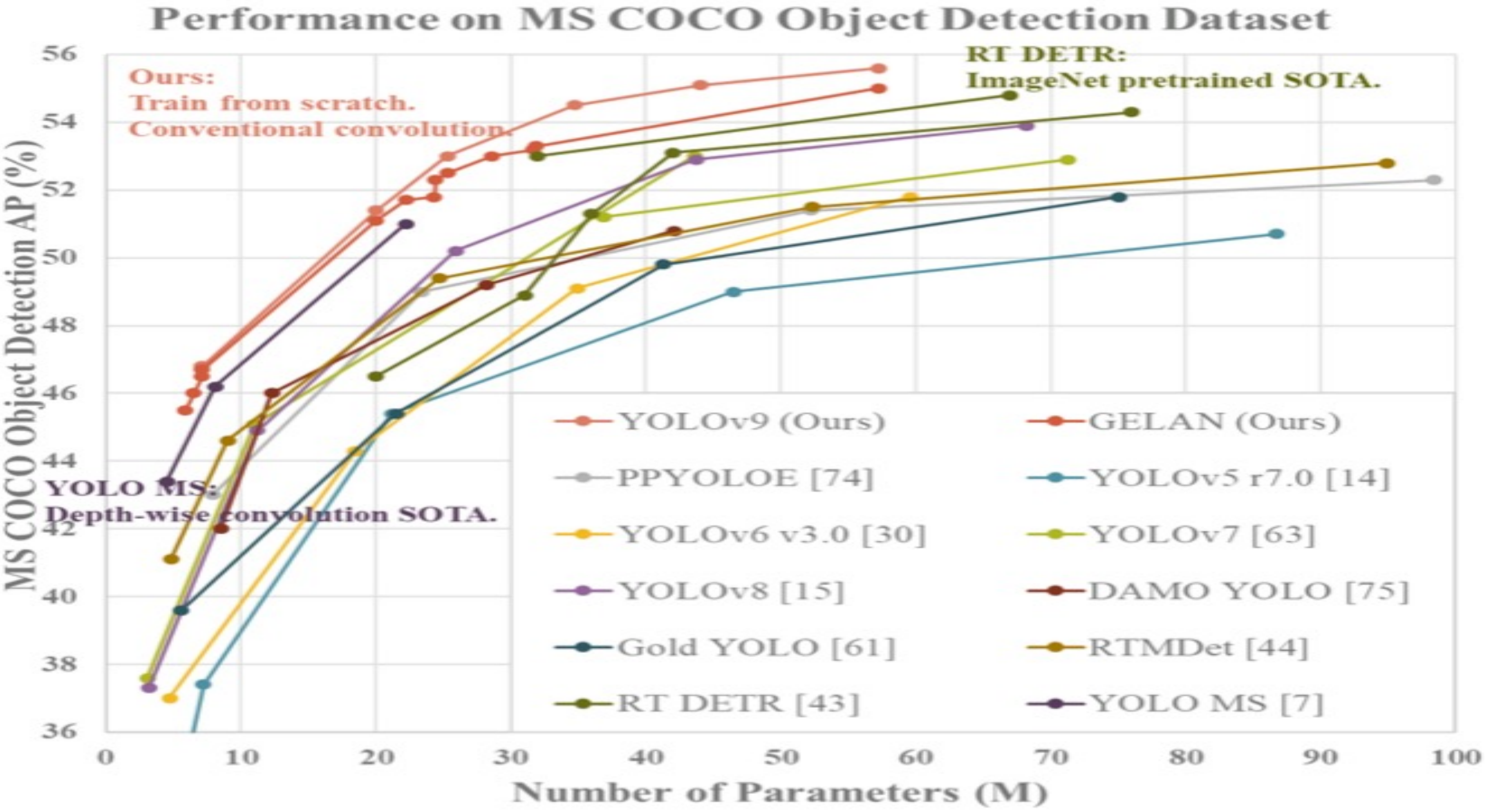
Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors



Source: Wang, Chien-Yao, Alexey Bochkovskiy, and Hong-Yuan Mark Liao.
"YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors." arXiv preprint arXiv:2207.02696 (2022).

YOLOv9:

Learning what you want to learn using programmable gradient information



Source: Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. (2025)
"Yolov9: Learning what you want to learn using programmable gradient information." In European Conference on Computer Vision, pp. 1-21. Springer, Cham.

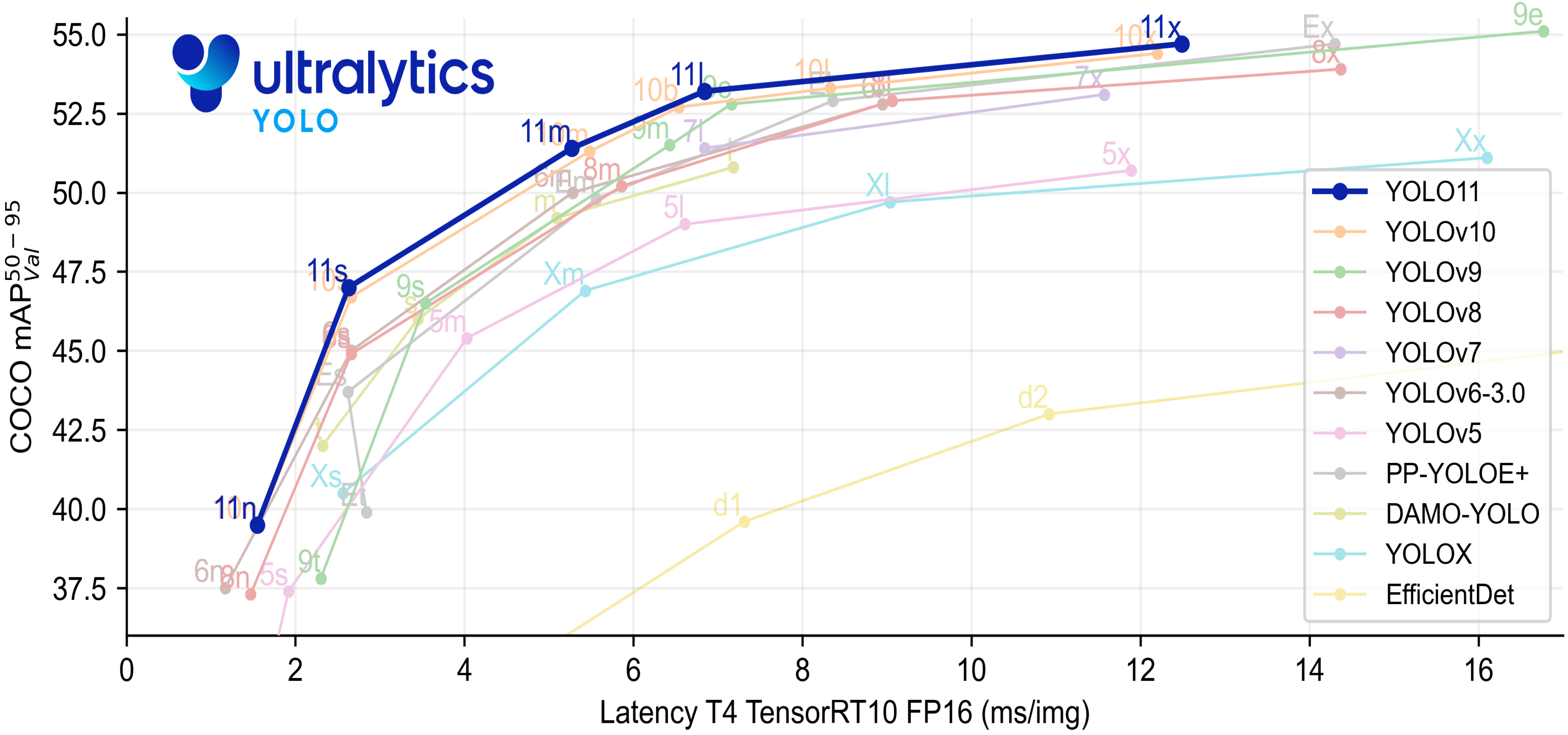
YOLOv9:

Comparison of state-of-the-art real-time object detectors

Model	#Param. (M)	FLOPs (G)	$AP^{val}_{50:95}$ (%)	AP^{val}_{50} (%)	AP^{val}_{75} (%)	AP^{val}_{S} (%)	AP^{val}_{M} (%)	AP^{val}_{L} (%)
YOLOv7 [63]	36.9	104.7	51.2	69.7	55.9	31.8	55.5	65.0
YOLOv7-X [63]	71.3	189.9	52.9	71.1	51.4	36.9	57.7	68.6
YOLOv7-N AF [63]	3.1	8.7	37.6	53.3	40.6	18.7	41.7	52.8
YOLOv7-S AF [63]	11.0	28.1	45.1	61.8	48.9	25.7	50.2	61.2
YOLOv7 AF [63]	43.6	130.5	53.0	70.2	57.5	35.8	58.7	68.9
YOLOv8-N [15]	3.2	8.7	37.3	52.6	–	–	–	–
YOLOv8-S [15]	11.2	28.6	44.9	61.8	–	–	–	–
YOLOv8-M [15]	25.9	78.9	50.2	67.2	–	–	–	–
YOLOv8-L [15]	43.7	165.2	52.9	69.8	57.5	35.3	58.3	69.8
YOLOv8-X [15]	68.2	257.8	53.9	71.0	58.7	35.7	59.3	70.7
YOLOv9-S (Ours)	7.1	26.4	46.8	63.4	50.7	26.6	56.0	64.5
YOLOv9-M (Ours)	20.0	76.3	51.4	68.1	56.1	33.6	57.0	68.0
YOLOv9-C (Ours)	25.3	102.1	53.0	70.2	57.8	36.2	58.5	69.3
YOLOv9-E (Ours)	57.3	189.0	55.6	72.8	60.6	40.2	61.0	71.4

Source: Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. (2025)
"Yolov9: Learning what you want to learn using programmable gradient information." In European Conference on Computer Vision, pp. 1-21. Springer, Cham.

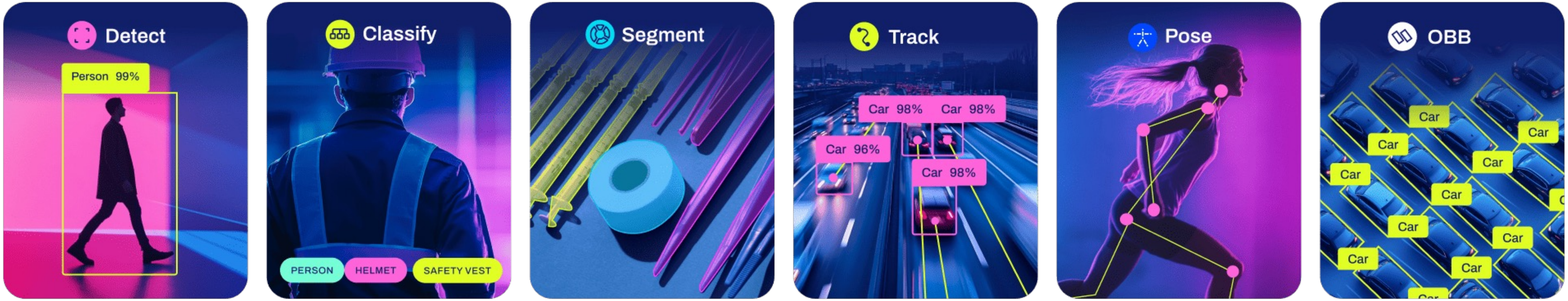
Ultralytics YOLO11



Source: https://docs.ultralytics.com/models/yolo11/

Ultralytics YOLO11 for Computer Vision

YOLO11 can train, val, predict and export models for the most common tasks in vision AI: Detect, Segment, Classify and Pose.

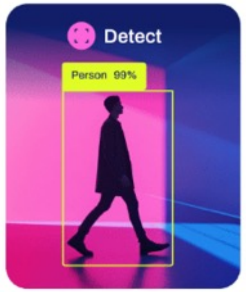


https://colab.research.google.com/github/ultralytics/ultralytics/blob/main/examples/tutorial.ipynb

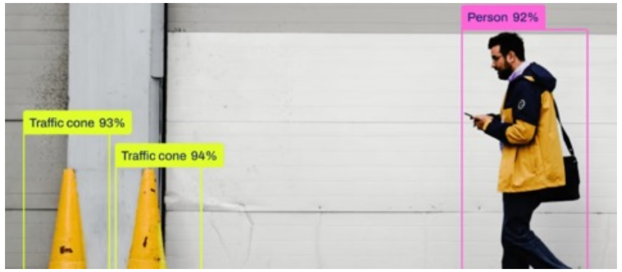
YOLO11 Supported Tasks and Modes

Model	Filenames	Task
YOLO11	yolo11n.pt yolo11s.pt yolo11m.pt yolo11l.pt yolo11x.pt	[Detection]()
YOLO11-seg	yolo11n-seg.pt yolo11s-seg.pt yolo11m-seg.pt yolo11l-seg.pt yolo11x-seg.pt	[Instance Segmentation]()
YOLO11-pose	yolo11n-pose.pt yolo11s-pose.pt yolo11m-pose.pt yolo11l-pose.pt yolo11x-pose.pt	[Pose/Keypoints]()
YOLO11-obb	yolo11n-obb.pt yolo11s-obb.pt yolo11m-obb.pt yolo11l-obb.pt yolo11x-obb.pt	[Oriented Detection]()
YOLO11-cls	yolo11n-cls.pt yolo11s-cls.pt yolo11m-cls.pt yolo11l-cls.pt yolo11x-cls.pt	[Classification]()

Source: https://docs.ultralytics.com/models/yolo11/



YOLO11 Performance Detection (COCO)



Object detection: identifying the location and class of objects in an image or video stream (80 pre-trained classes)

Model	size (pixels)	mAP^{val} 50-95	Speed CPU ONNX (ms)	Speed T4 TensorRT10 (ms)	params (M)	FLOPs (B)
YOLO11n	640	39.5	56.1 ± 0.8	1.5 ± 0.0	2.6	6.5
YOLO11s	640	47.0	90.0 ± 1.2	2.5 ± 0.0	9.4	21.5
YOLO11m	640	51.5	183.2 ± 2.0	4.7 ± 0.1	20.1	68.0
YOLO11l	640	53.4	238.6 ± 1.4	6.2 ± 0.1	25.3	86.9
YOLO11x	640	**54.7**	462.8 ± 6.7	11.3 ± 0.2	56.9	194.9

Source: https://docs.ultralytics.com/models/yolo11/

YOLO11 Performance Segmentation (COCO)

80 pre-trained classes

Model	size (pixels)	mAP^{box} 50-95	mAP^{mask} 50-95	Speed CPU ONNX (ms)	Speed T4 TensorRT10 (ms)	params (M)	FLOPs (B)
YOLO11n-seg	640	38.9	32.0	65.9 ± 1.1	1.8 ± 0.0	2.9	10.4
YOLO11s-seg	640	46.6	37.8	117.6 ± 4.9	2.9 ± 0.0	10.1	35.5
YOLO11m-seg	640	51.5	41.5	281.6 ± 1.2	6.3 ± 0.1	22.4	123.3
YOLO11l-seg	640	53.4	42.9	344.2 ± 3.2	7.8 ± 0.2	27.6	142.2
YOLO11x-seg	640	**54.7**	**43.8**	664.5 ± 3.2	15.8 ± 0.7	62.1	319.0

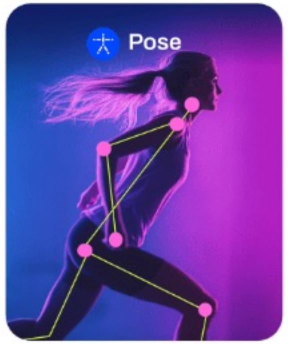
Source: https://docs.ultralytics.com/models/yolo11/

YOLO11 Performance Classification (ImageNet)

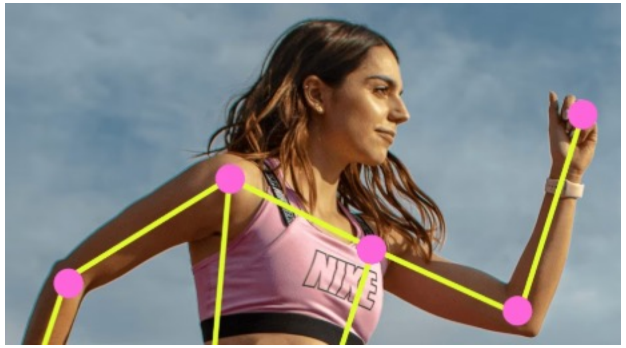
1000 pre-trained classes

Model	size (pixels)	acc top1	acc top5	Speed CPU ONNX (ms)	Speed T4 TensorRT10 (ms)	params (M)	FLOPs (B) at 640
YOLO11n-cls	224	70.0	89.4	5.0 ± 0.3	1.1 ± 0.0	1.6	3.3
YOLO11s-cls	224	75.4	92.7	7.9 ± 0.2	1.3 ± 0.0	5.5	12.1
YOLO11m-cls	224	77.3	93.9	17.2 ± 0.4	2.0 ± 0.0	10.4	39.3
YOLO11l-cls	224	78.3	94.3	23.2 ± 0.3	2.8 ± 0.0	12.9	49.4
YOLO11x-cls	224	79.5	94.9	41.4 ± 0.9	3.8 ± 0.0	28.4	110.4

Source: https://docs.ultralytics.com/models/yolo11/



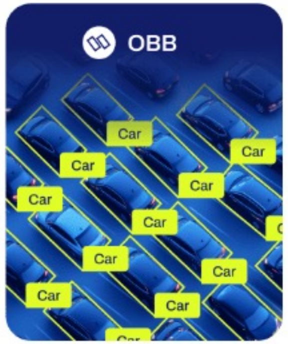
YOLO11 Performance Pose / Keypoint (COCO)



1 pre-trained class: 'person'

Model	size (pixels)	mAP^{pose} 50-95	mAP^{pose} 50	Speed CPU ONNX (ms)	Speed T4 TensorRT10 (ms)	params (M)	FLOPs (B)
YOLO11n-pose	640	50.0	81.0	52.4 ± 0.5	1.7 ± 0.0	2.9	7.6
YOLO11s-pose	640	58.9	86.3	90.5 ± 0.6	2.6 ± 0.0	9.9	23.2
YOLO11m-pose	640	64.9	89.4	187.3 ± 0.8	4.9 ± 0.1	20.9	71.7
YOLO11l-pose	640	66.1	89.9	247.7 ± 1.1	6.4 ± 0.1	26.2	90.7
YOLO11x-pose	640	69.5	91.1	488.0 ± 13.9	12.1 ± 0.2	58.8	203.3

Source: https://docs.ultralytics.com/models/yolo11/



YOLO11 Performance
OBB (DOTAv1)

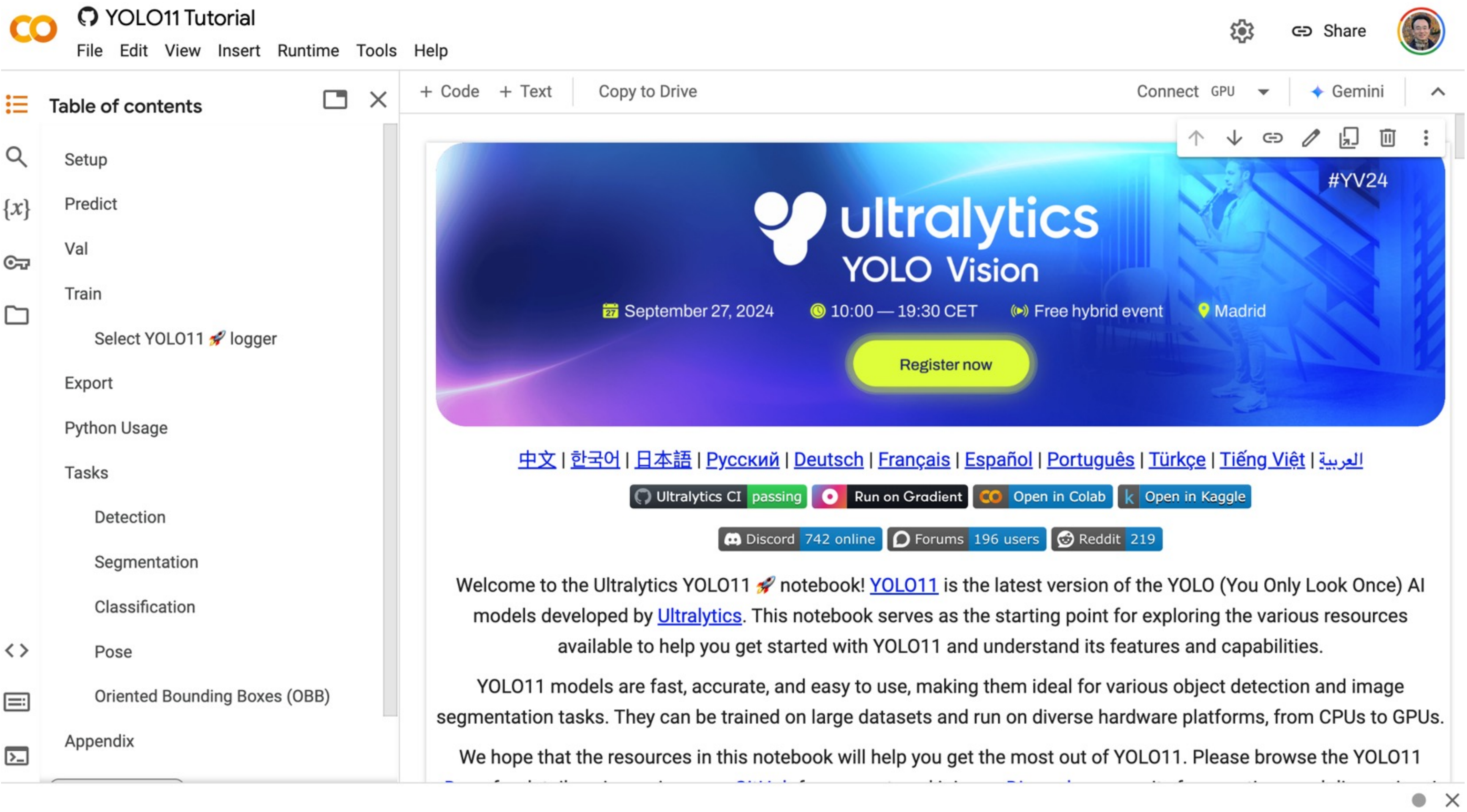
Oriented Bounding Boxes Object Detection

15 pre-trained classes

Model	size (pixels)	mAP^{test}_{50}	Speed CPU ONNX (ms)	Speed T4 TensorRT10 (ms)	params (M)	FLOPs (B)
YOLO11n-obb	1024	78.4	117.6 ± 0.8	4.4 ± 0.0	2.7	17.2
YOLO11s-obb	1024	79.5	219.4 ± 4.0	5.1 ± 0.0	9.7	57.5
YOLO11m-obb	1024	80.9	562.8 ± 2.9	10.1 ± 0.4	20.9	183.5
YOLO11l-obb	1024	81.0	712.5 ± 5.0	13.5 ± 0.6	26.2	232.0
YOLO11x-obb	1024	81.3	1408.6 ± 7.7	28.6 ± 1.0	58.8	520.2

Source: https://docs.ultralytics.com/models/yolo11/

Ultralytics YOLO11 Tutorial



https://colab.research.google.com/github/ultralytics/ultralytics/blob/main/examples/tutorial.ipynb

YOLO11 Tutorial (1. Predict)

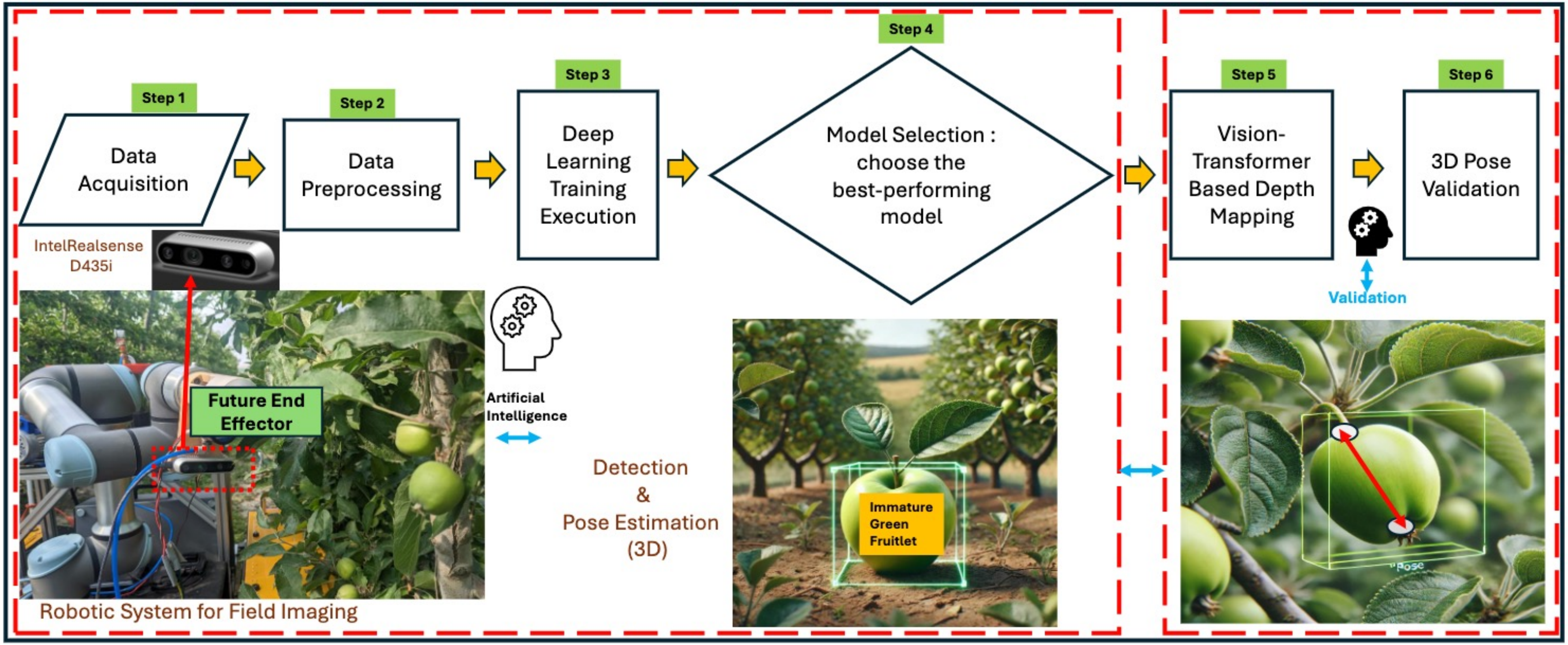
```
%pip install ultralytics
import ultralytics
ultralytics.checks()
```

```
# Run inference on an image with YOLO11n
!yolo predict model=yolo11n.pt source='https://ultralytics.com/images/zidane.jpg'
```

https://colab.research.google.com/github/ultralytics/ultralytics/blob/main/examples/tutorial.ipynb

Robotic System with Computer Vision

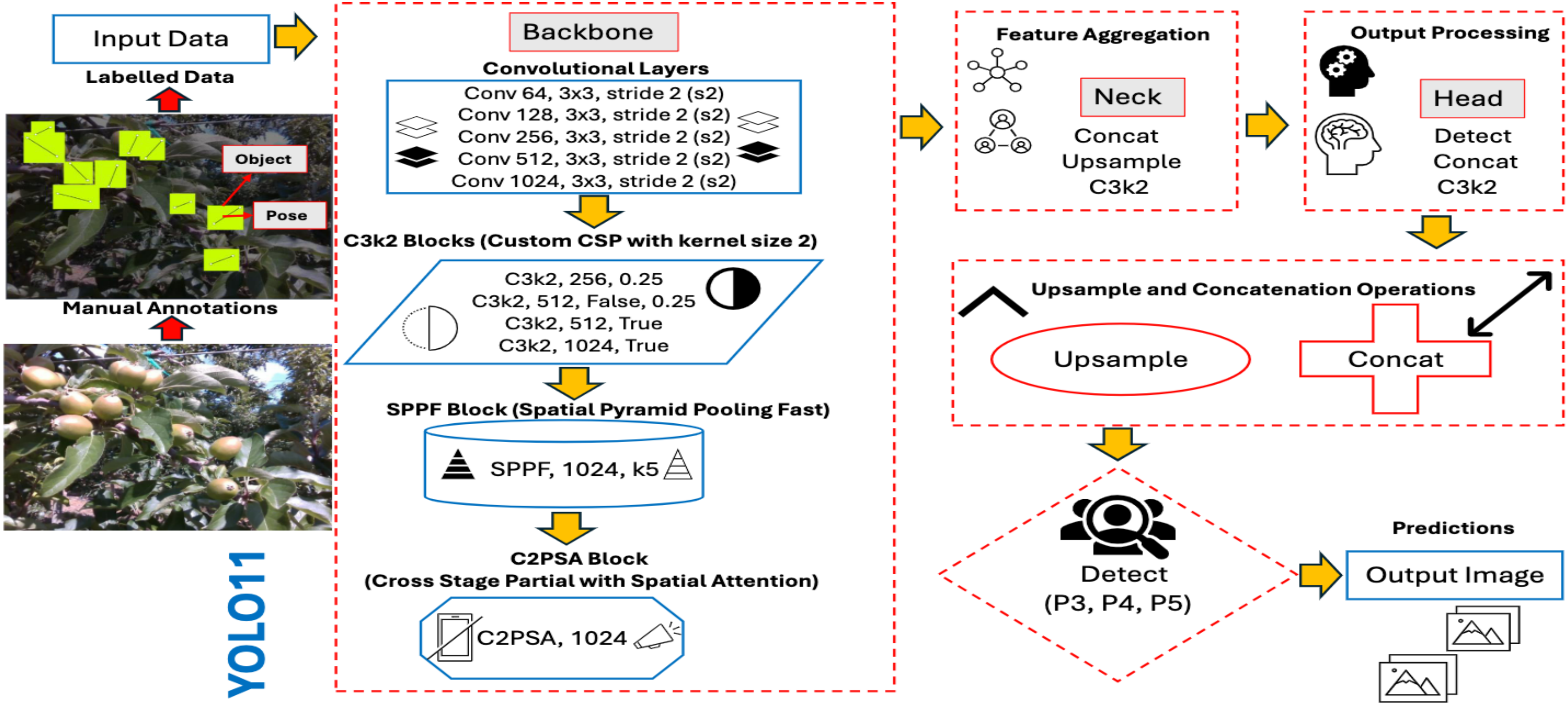
Precise pose estimation of immature green fruitlets



Source: Sapkota, Ranjan, and Manoj Karkee. "Yolo11 and vision transformers based 3d pose estimation of immature green fruits in commercial apple orchards for robotic thinning." arXiv preprint arXiv:2410.19846 (2024)..

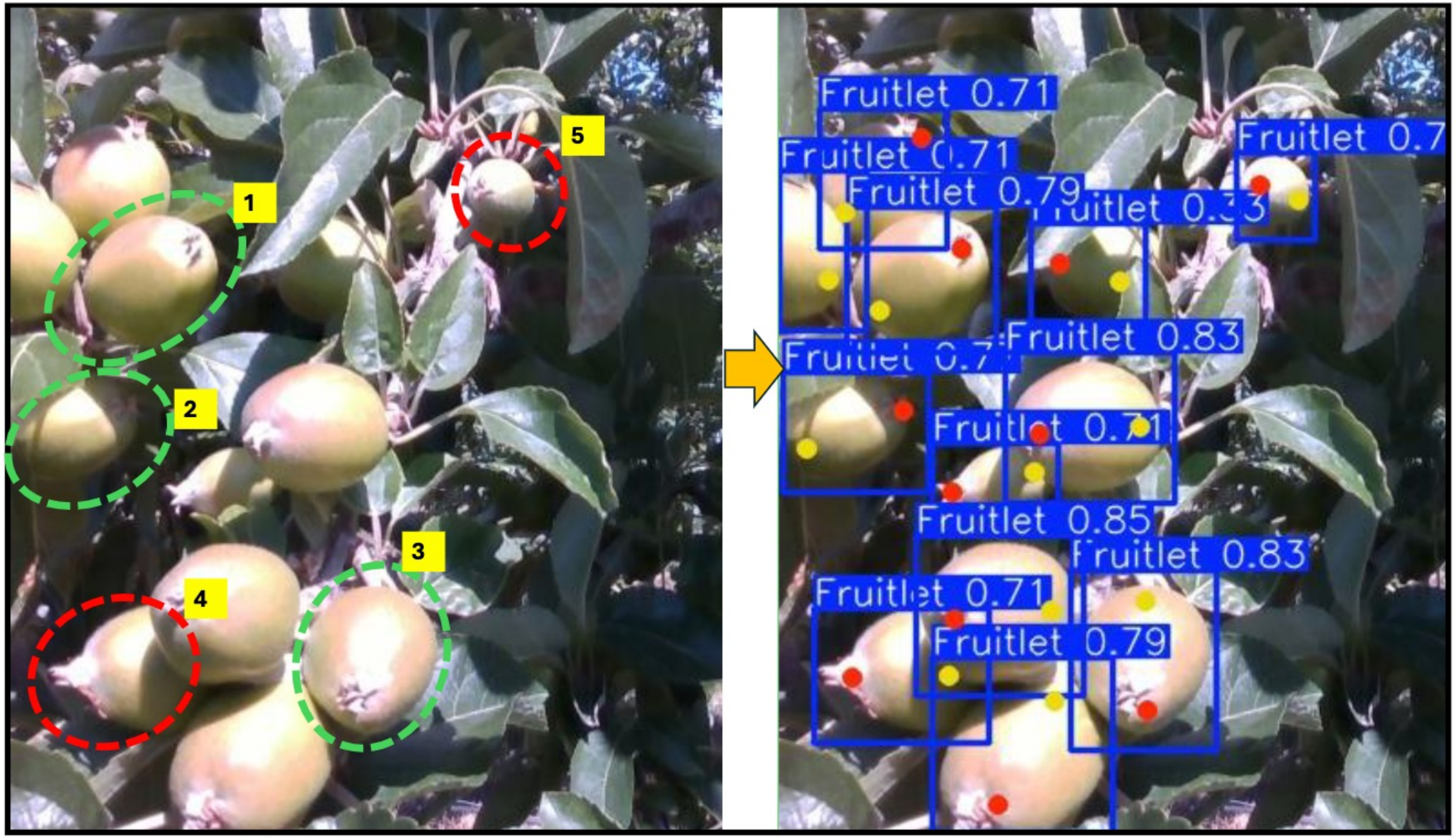
YOLO11 Architecture Diagram

for immature green fruit detection and pose estimation



Source: Sapkota, Ranjan, and Manoj Karkee. "Yolo11 and vision transformers based 3d pose estimation of immature green fruits in commercial apple orchards for robotic thinning." arXiv preprint arXiv:2410.19846 (2024)..

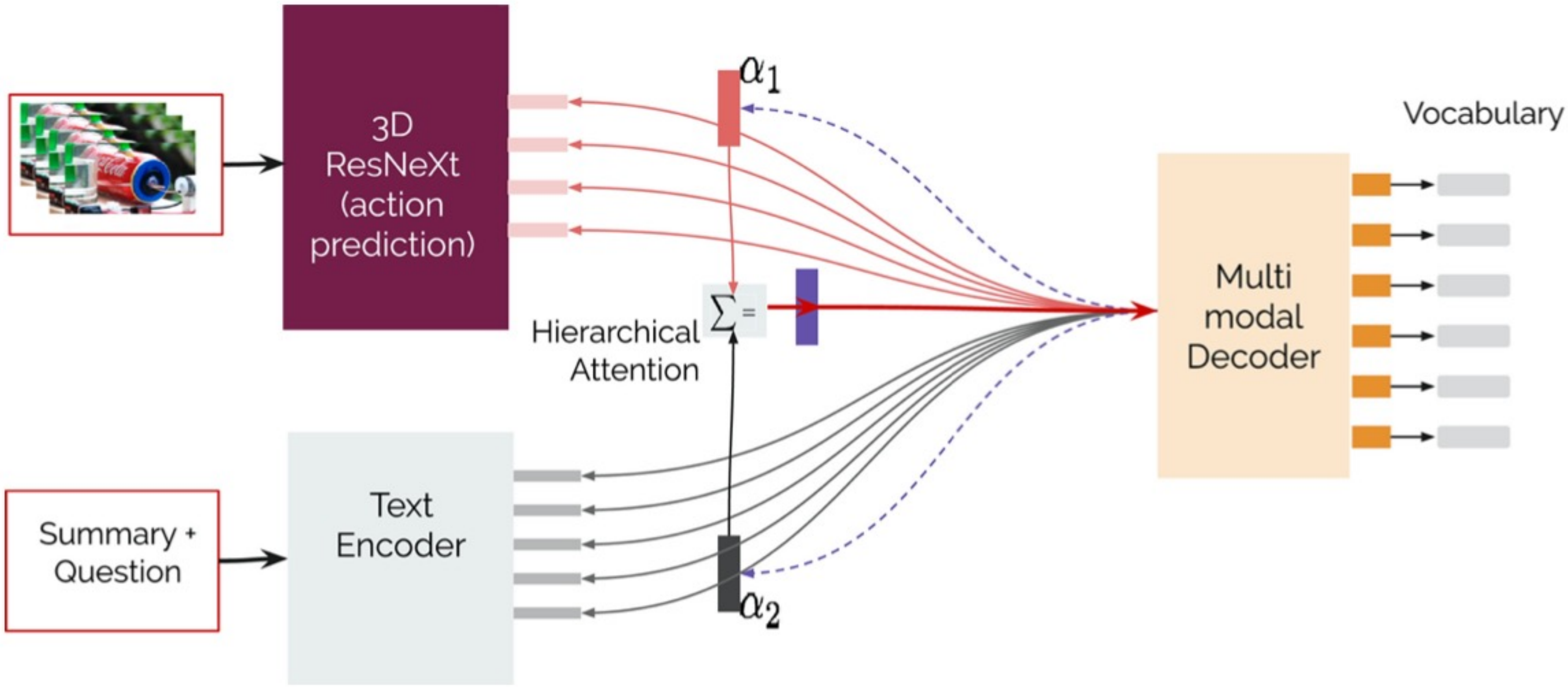
YOLO11n Detection and Pose Estimation capabilities in a commercial orchard



(a)

Source: Sapkota, Ranjan, and Manoj Karkee. "Yolo11 and vision transformers based 3d pose estimation of immature green fruits in commercial apple orchards for robotic thinning." arXiv preprint arXiv:2410.19846 (2024)..

Text-and-Video Dialog Generation Models with Hierarchical Attention



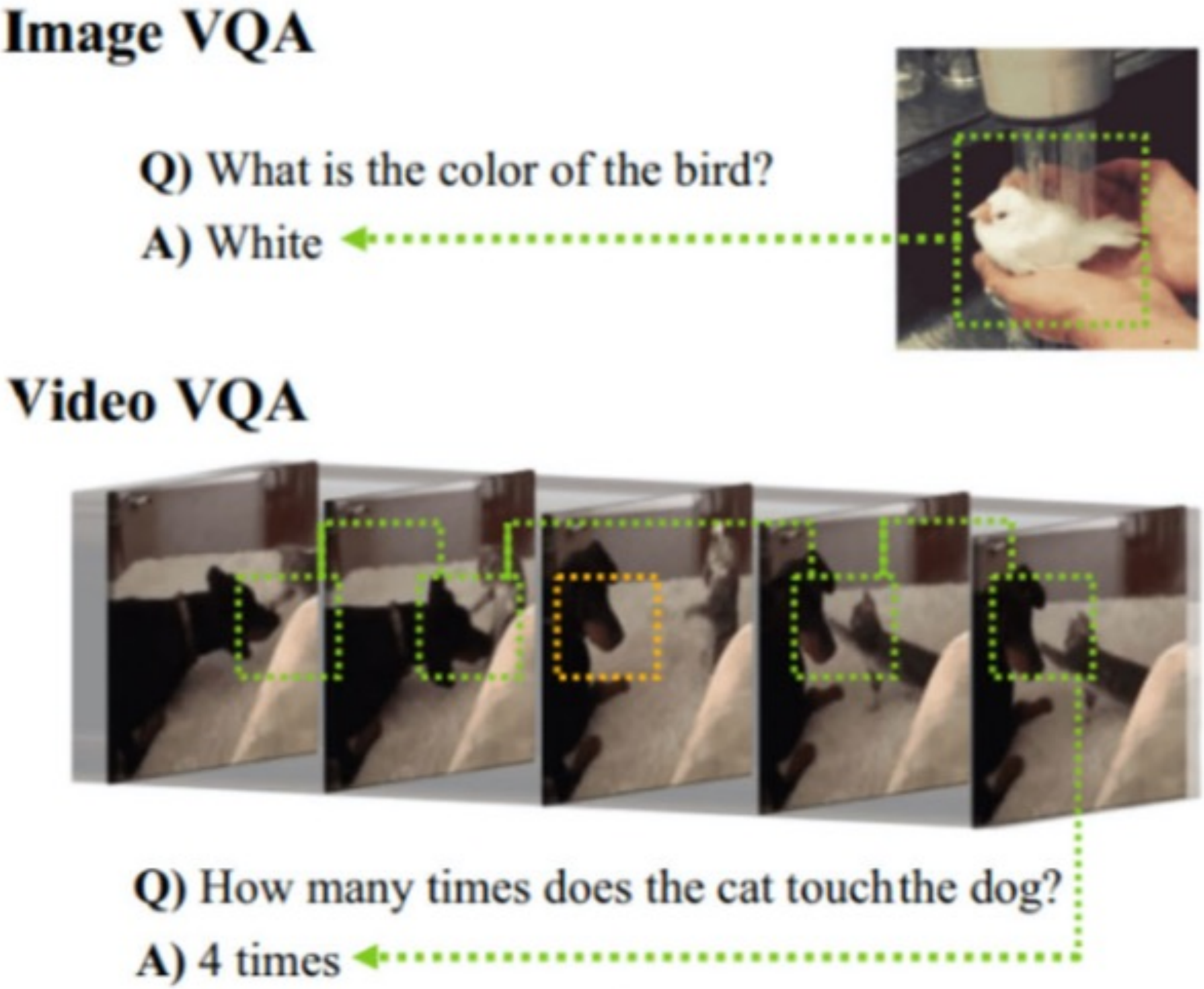
Source: Erdem, Erkut, Menekse Kuyu, Semih Yagcioglu, Anette Frank, Letitia Parcalabescu, Barbara Plank, Andrii Babii et al. "Neural Natural Language Generation: A Survey on Multilinguality, Multimodality, Controllability and Learning." Journal of Artificial Intelligence Research 73 (2022): 1131-1207.

Multimodal Few-Shot Learning with Frozen Language Models

Curated samples with about five seeds required to get past well-known language model failure modes of either repeating text for the prompt or emitting text that does not pertain to the image.
These samples demonstrate the ability to generate open-ended outputs that adapt to both images and text, and to make use of facts that it has learned during language-only pre-training.

Source: Maria Tsimpoukelli, Jacob L. Menick, Serkan Cabi, S. M. Eslami, Oriol Vinyals, and Felix Hill (2021). "Multimodal few-shot learning with frozen language models." Advances in Neural Information Processing Systems 34 (2021): 200-212.

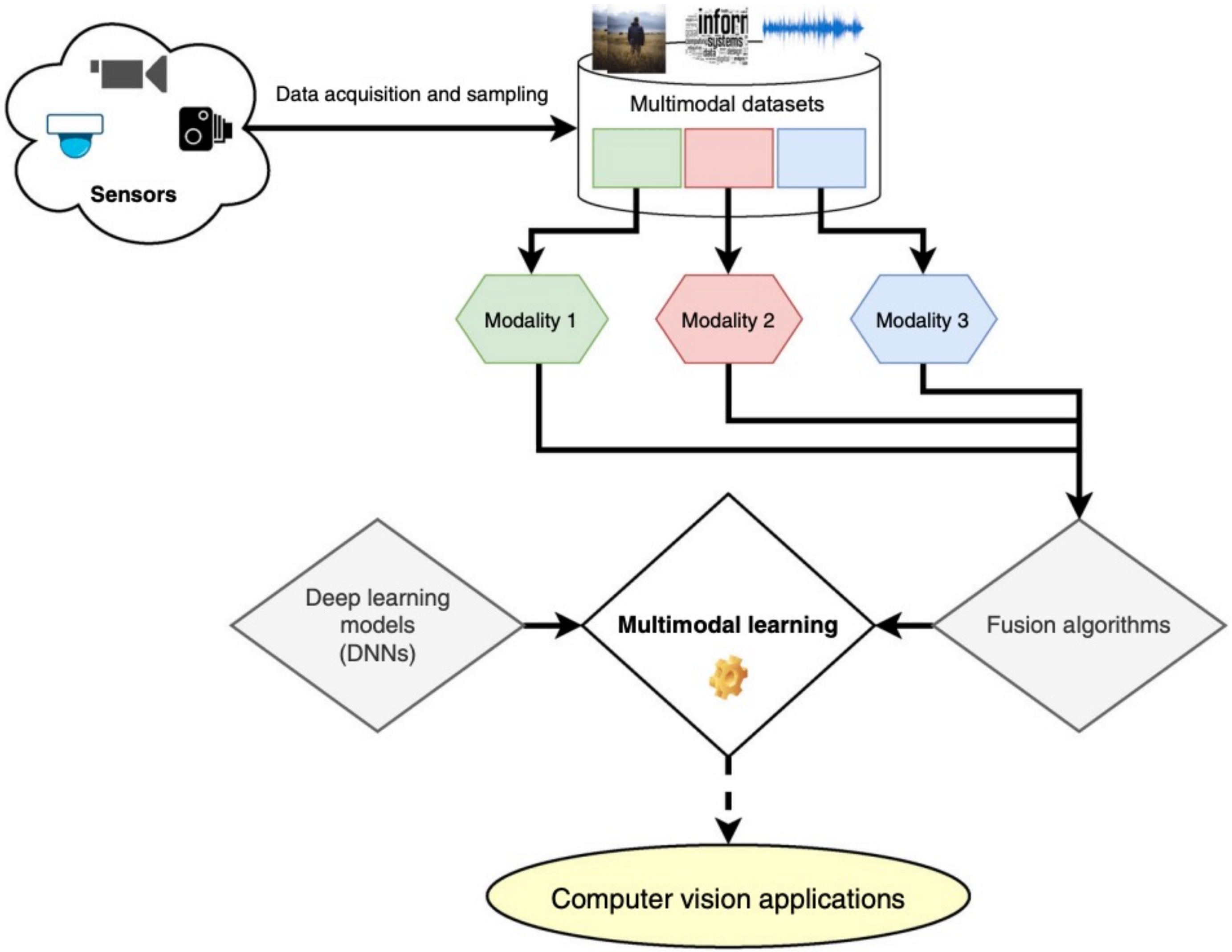
Video Question Answering (VQA)



Source: Bayoudh, Khaled, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa (2022).
"A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets." The Visual Computer 38, no. 8: 2939-2970.

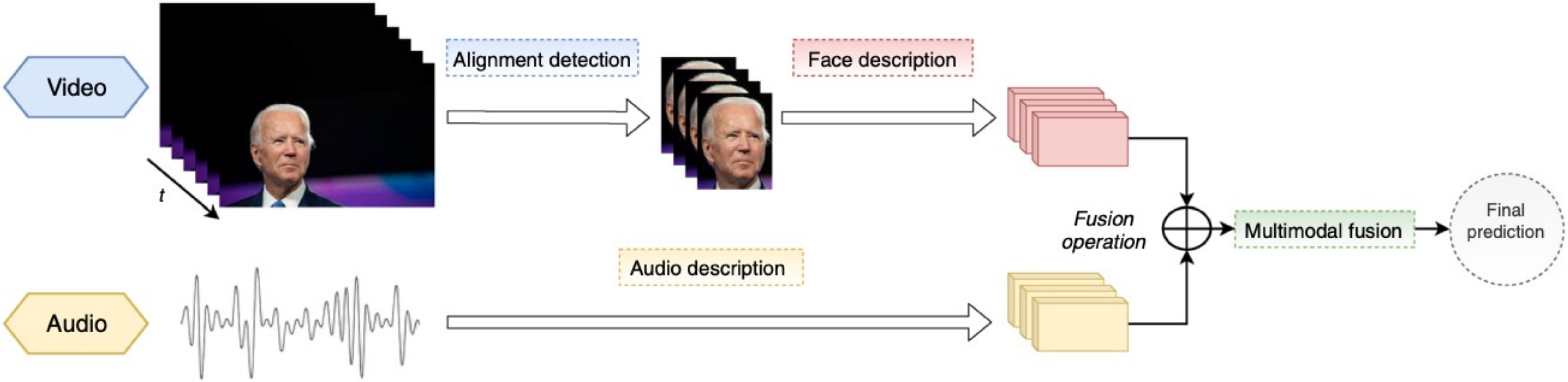
Multimodal Pipeline

that includes three different modalities (Image, Text. Audio)



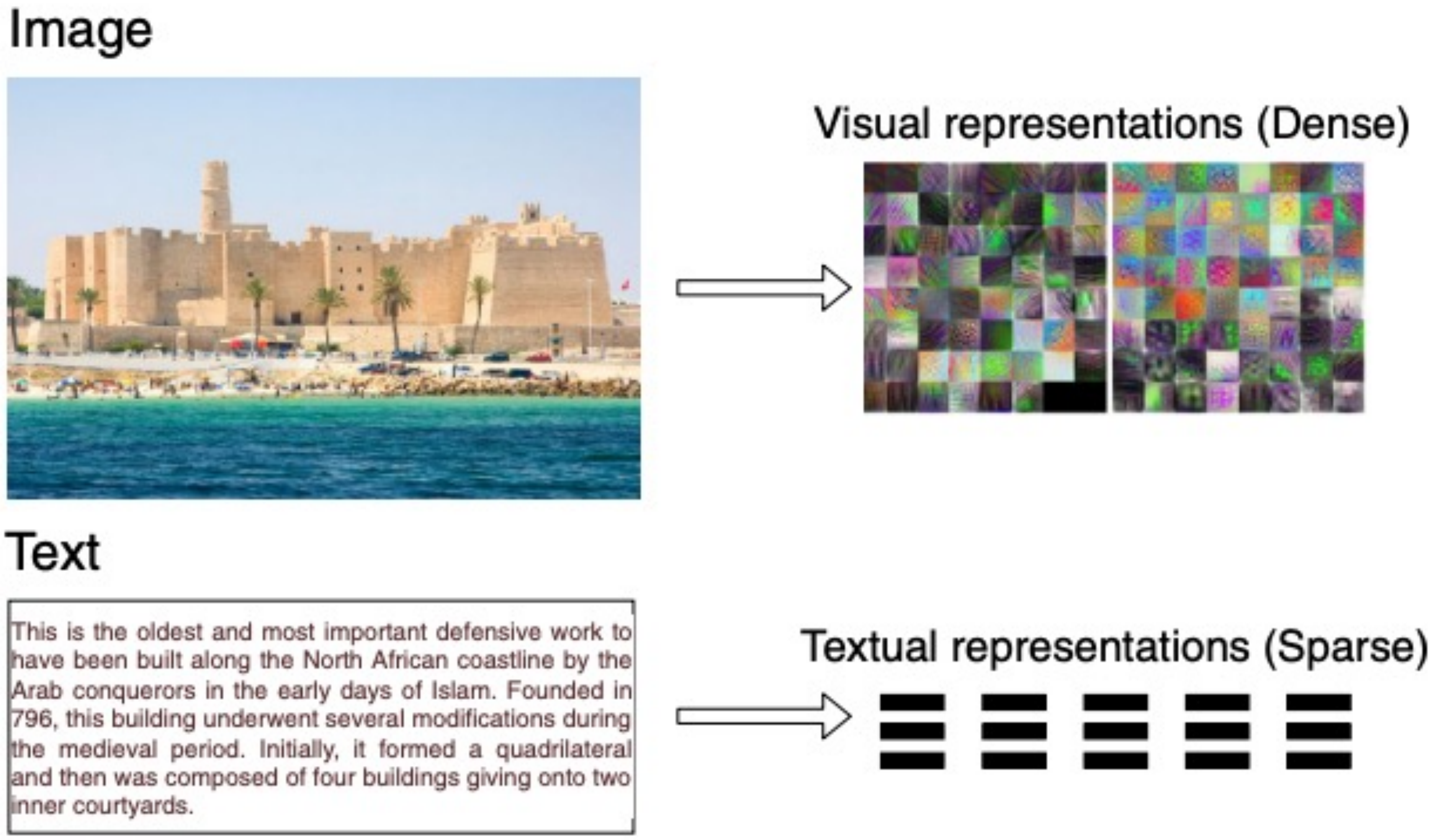
Source: Bayoudh, Khaled, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa (2022).
"A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets." The Visual Computer 38, no. 8: 2939-2970.

Video and Audio Multimodal Fusion



Source: Bayoudh, Khaled, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa (2022).
"A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets." The Visual Computer 38, no. 8: 2939-2970.

Visual and Textual Representation



Source: Bayoudh, Khaled, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa (2022). "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets." The Visual Computer 38, no. 8: 2939-2970.

Hybrid Multimodal Data Fusion

Text

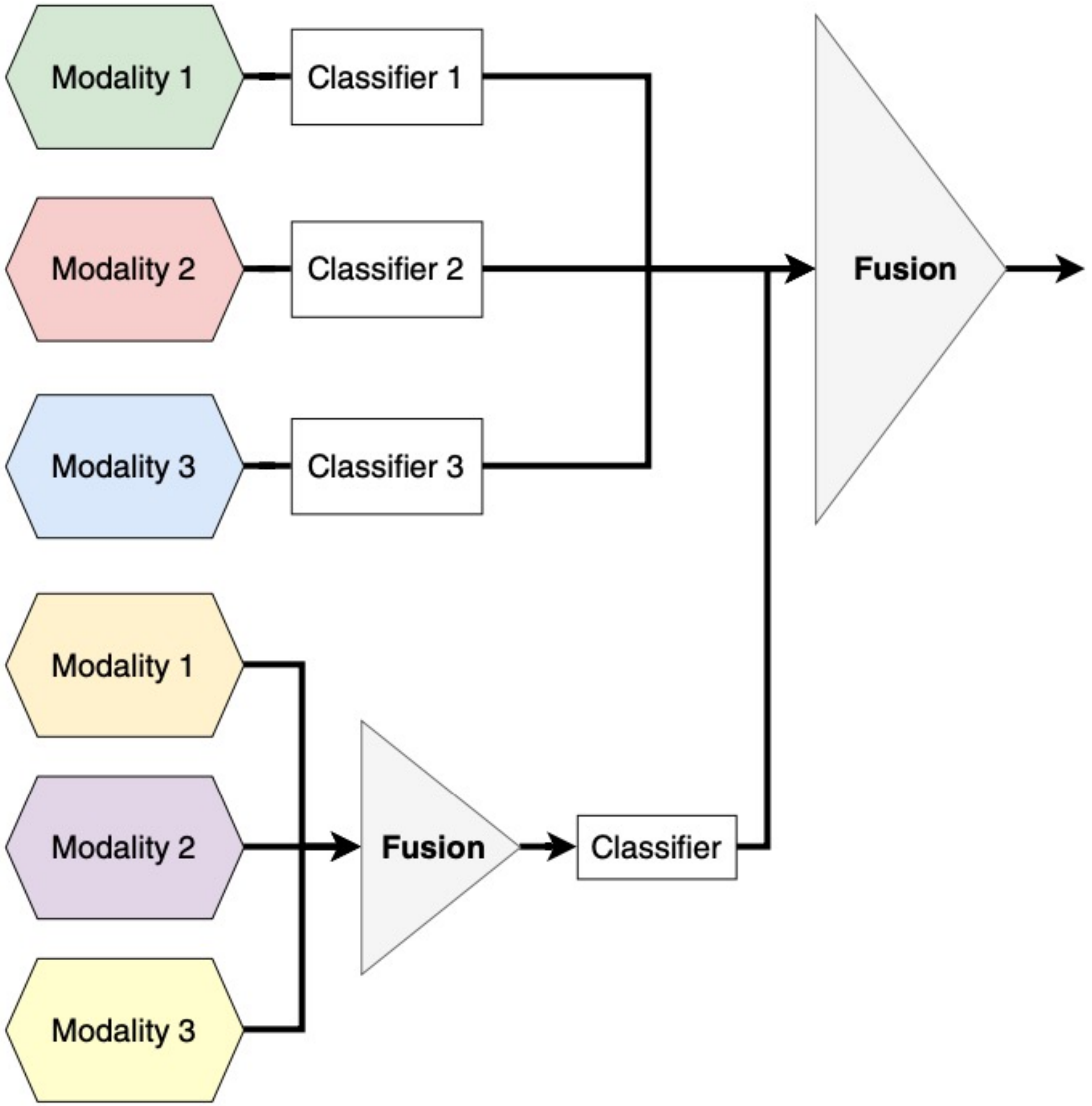
Audio

Image

Text

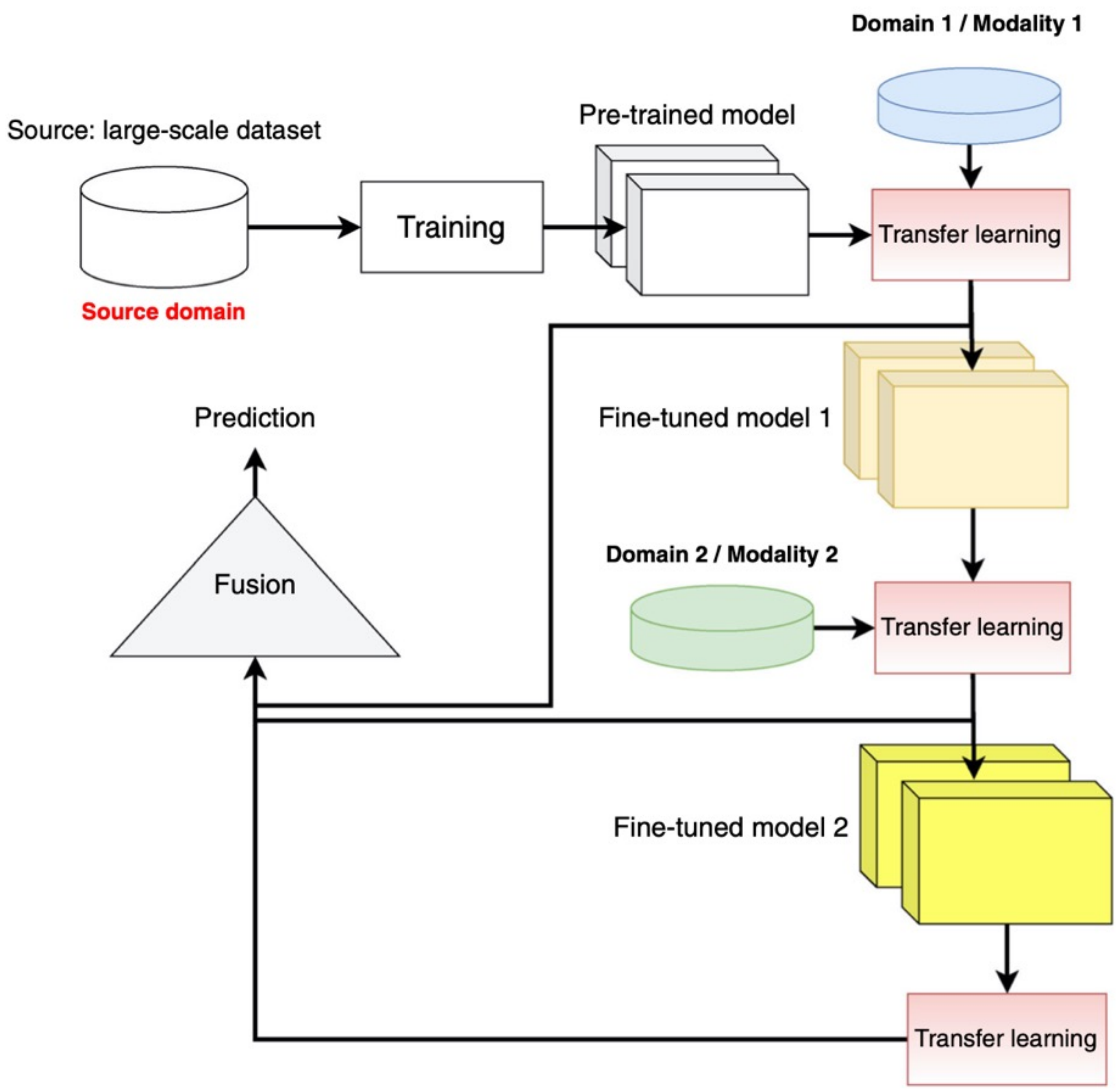
Speech

Video



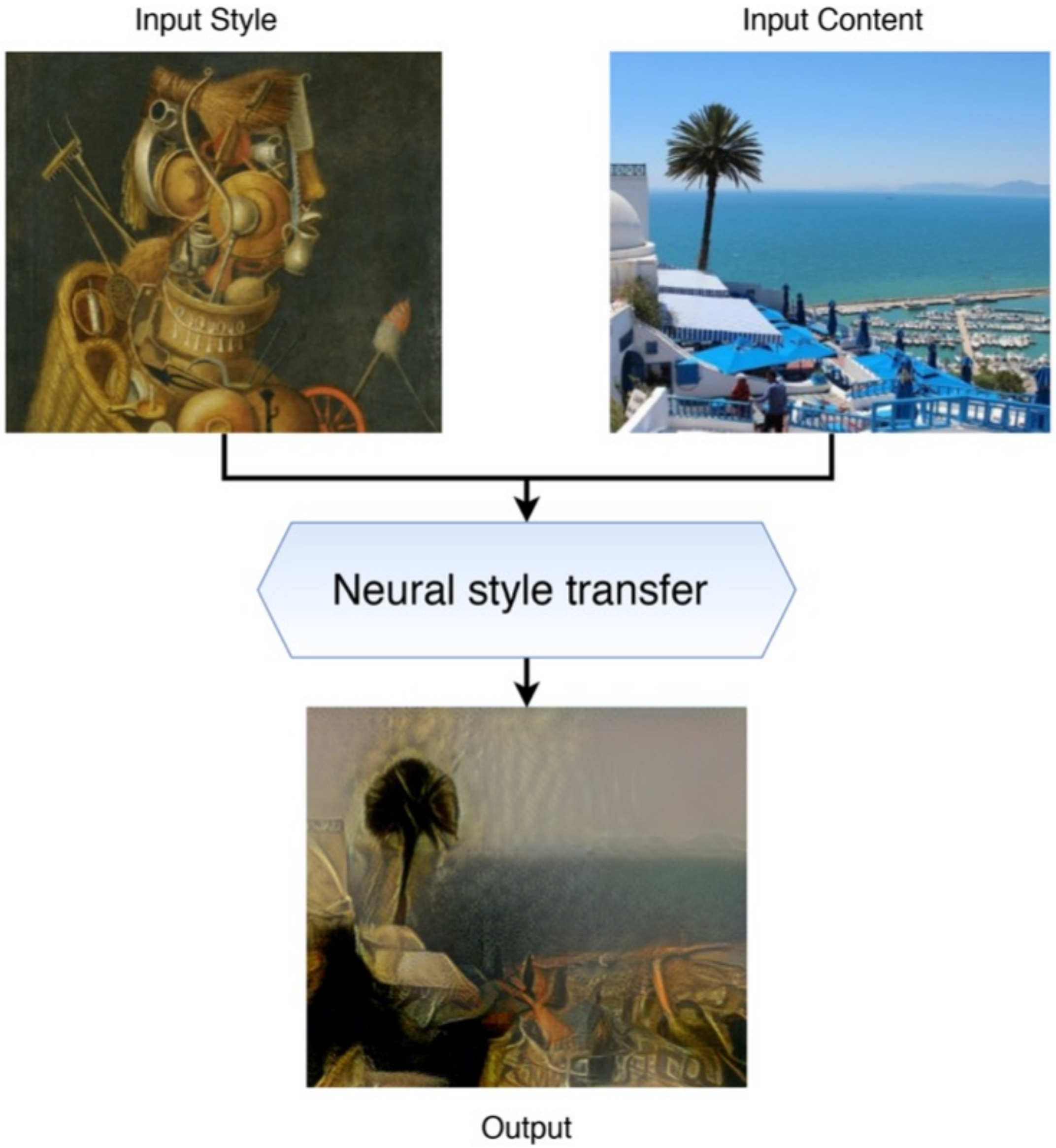
Source: Bayoudh, Khaled, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa (2022).
"A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets." The Visual Computer 38, no. 8: 2939-2970.

Multimodal Transfer Learning



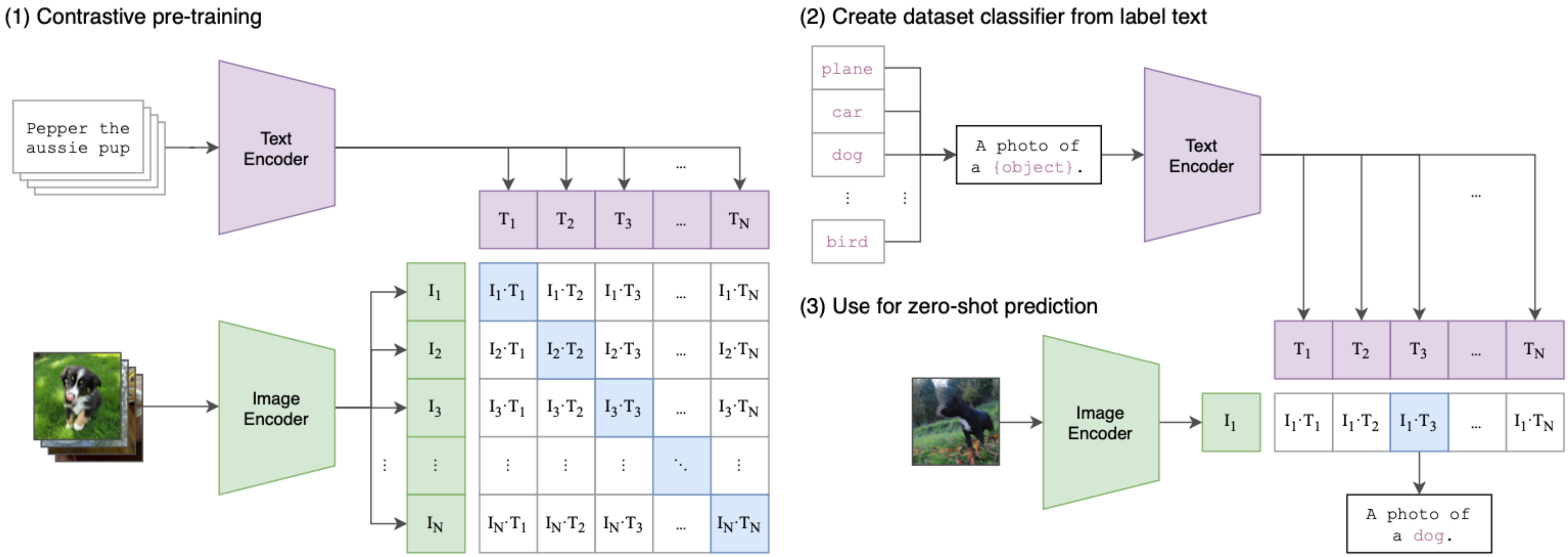
Source: Bayoudh, Khaled, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa (2022).
"A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets." The Visual Computer 38, no. 8: 2939-2970.

Neural Style Transfer (NST)



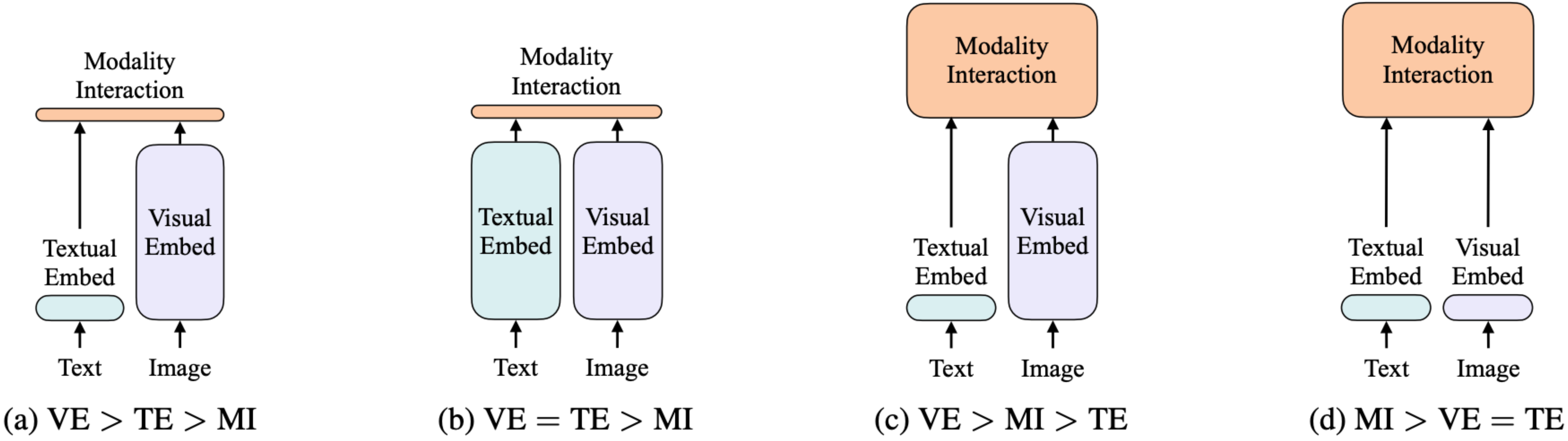
Source: Bayoudh, Khaled, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa (2022).
"A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets." The Visual Computer 38, no. 8: 2939-2970.

CLIP: Learning Transferable Visual Models From Natural Language Supervision



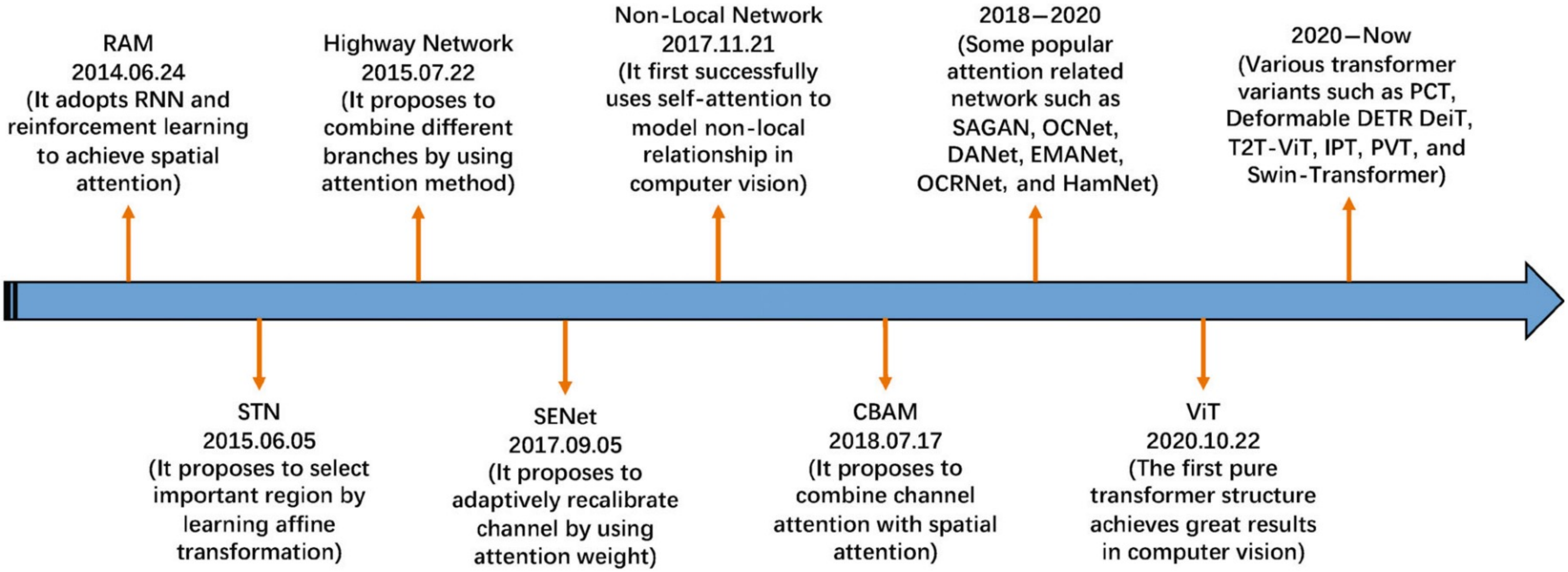
Source: Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry et al. (2021) "Learning transferable visual models from natural language supervision." In International Conference on Machine Learning, pp. 8748-8763. PMLR.

ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision



Source: Kim, Wonjae, Bokyung Son, and Ildoo Kim (2021). "Vilt: Vision-and-language transformer without convolution or region supervision." In International Conference on Machine Learning, pp. 5583-5594. PMLR.

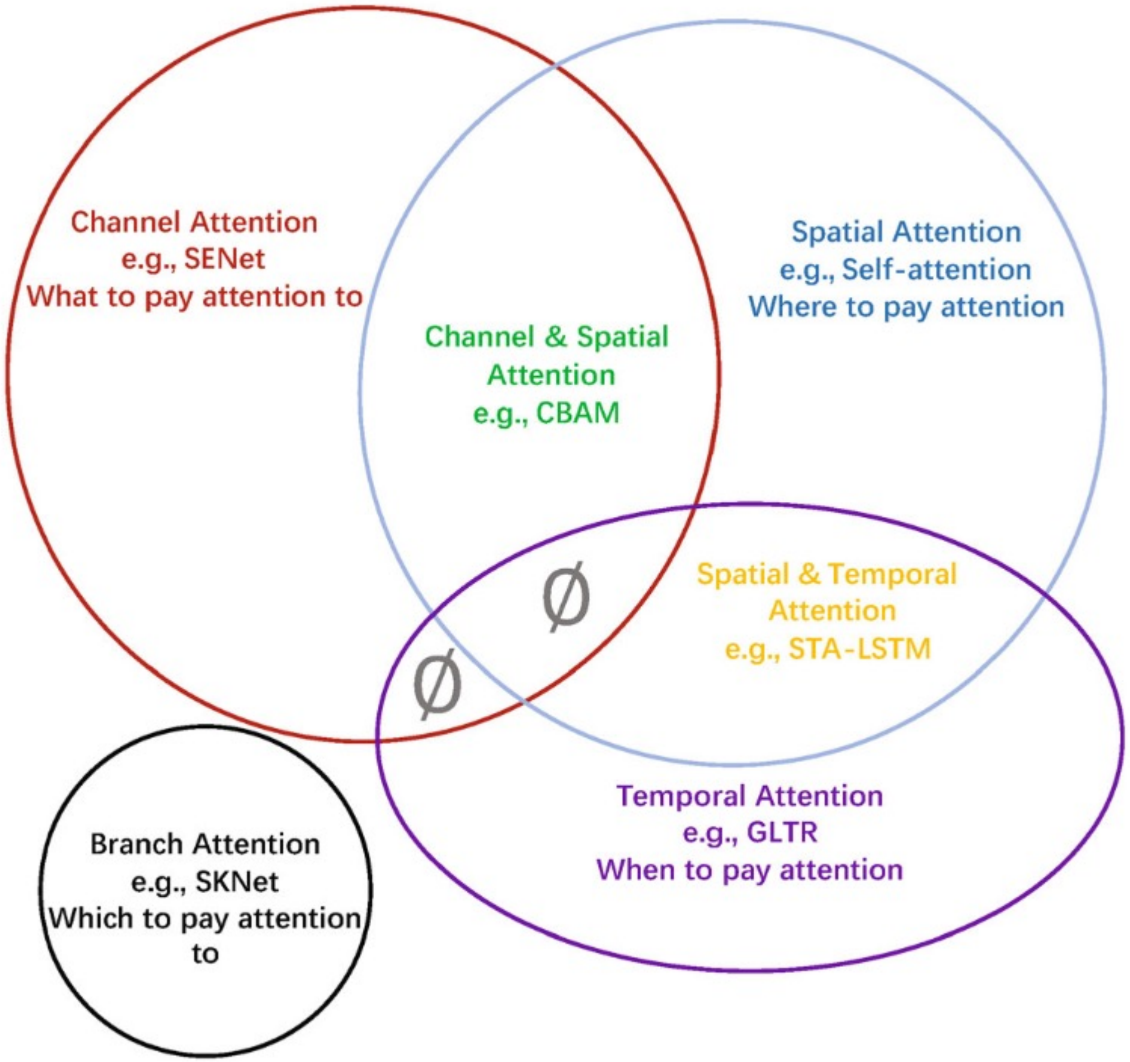
Attention Mechanisms in Computer Vision: A survey



Source: Guo, Meng-Hao, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R. Martin, Ming-Ming Cheng, and Shi-Min Hu. (2022) "Attention mechanisms in computer vision: A survey." Computational Visual Media ,:1-38.

Attention Mechanisms in Computer Vision:

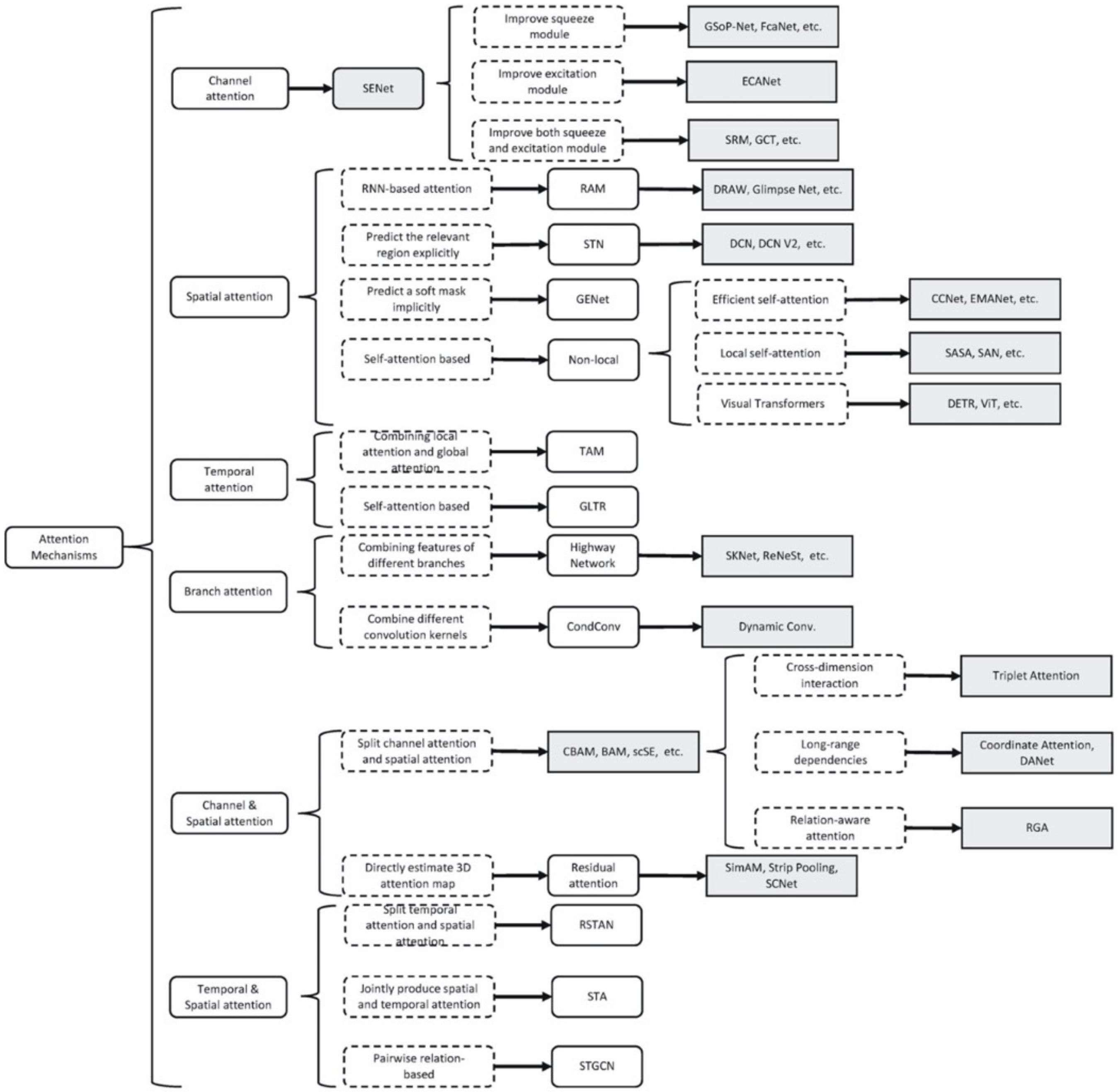
Data domain



Source: Guo, Meng-Hao, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R. Martin, Ming-Ming Cheng, and Shi-Min Hu. (2022) "Attention mechanisms in computer vision: A survey." Computational Visual Media ,:1-38.

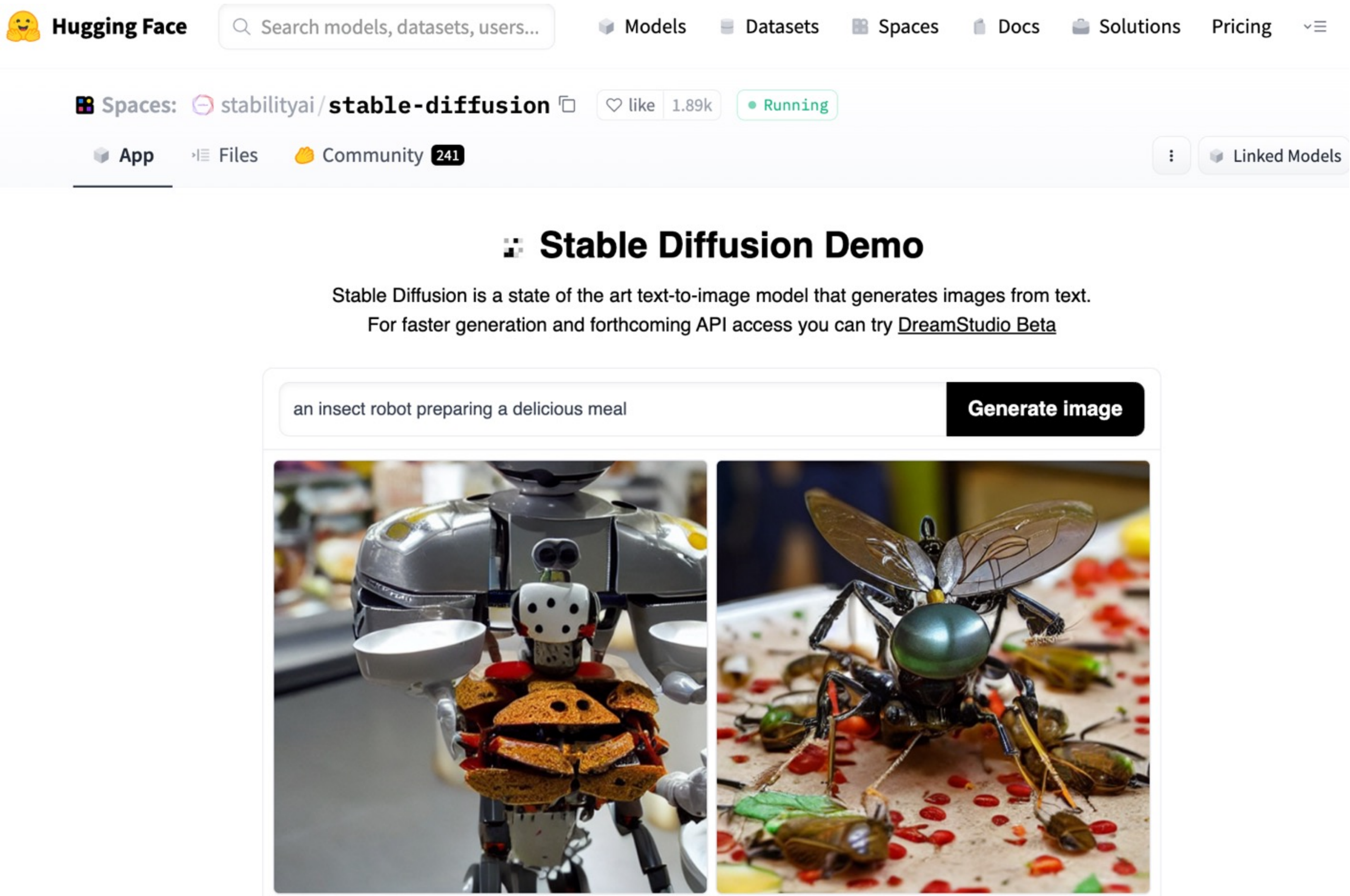
Attention Mechanisms in Computer Vision:

Developmental context of visual attention



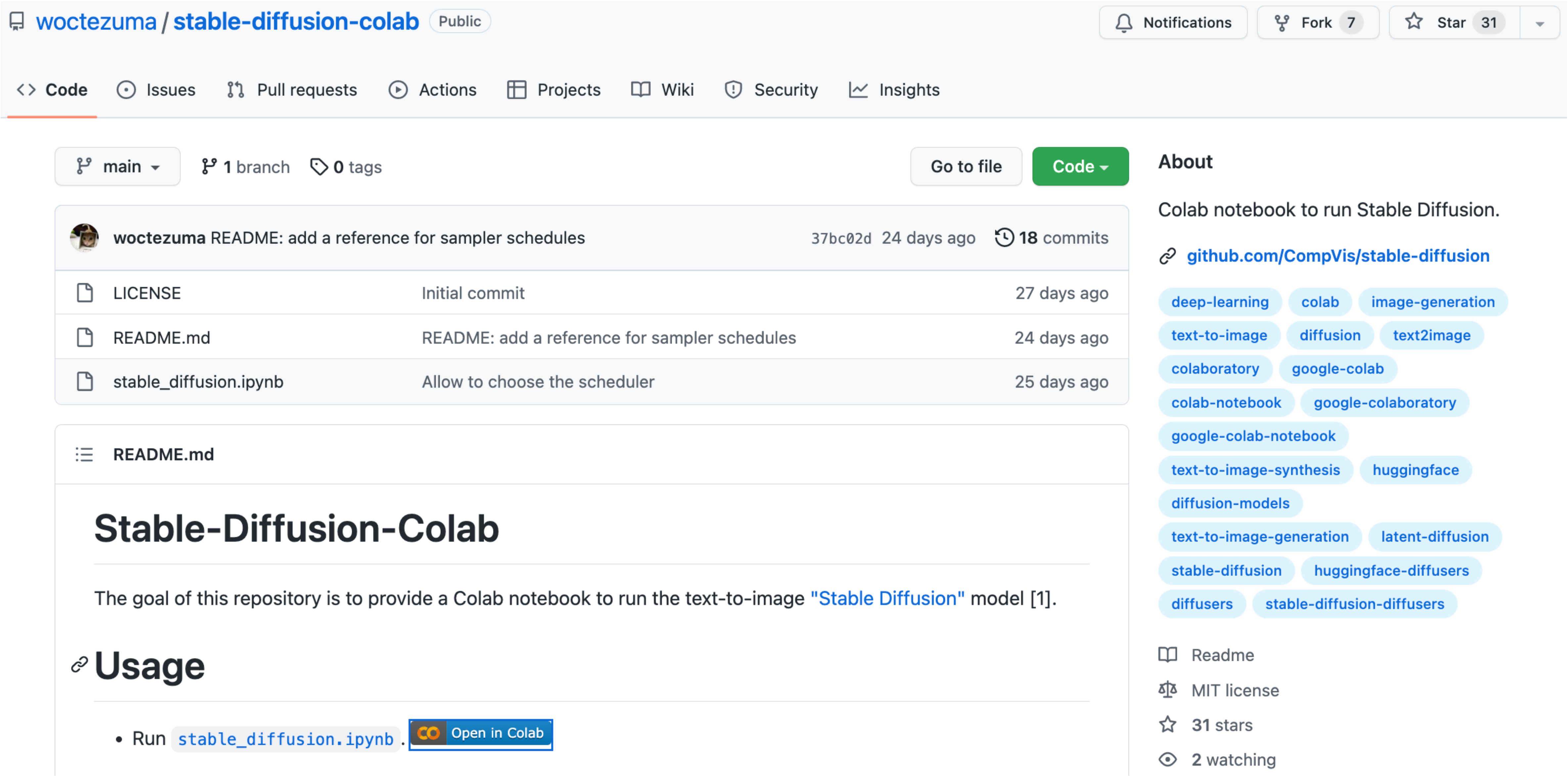
Source: Guo, Meng-Hao, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R. Martin, Ming-Ming Cheng, and Shi-Min Hu. (2022) "Attention mechanisms in computer vision: A survey." Computational Visual Media ,:1-38.

Stable Diffusion



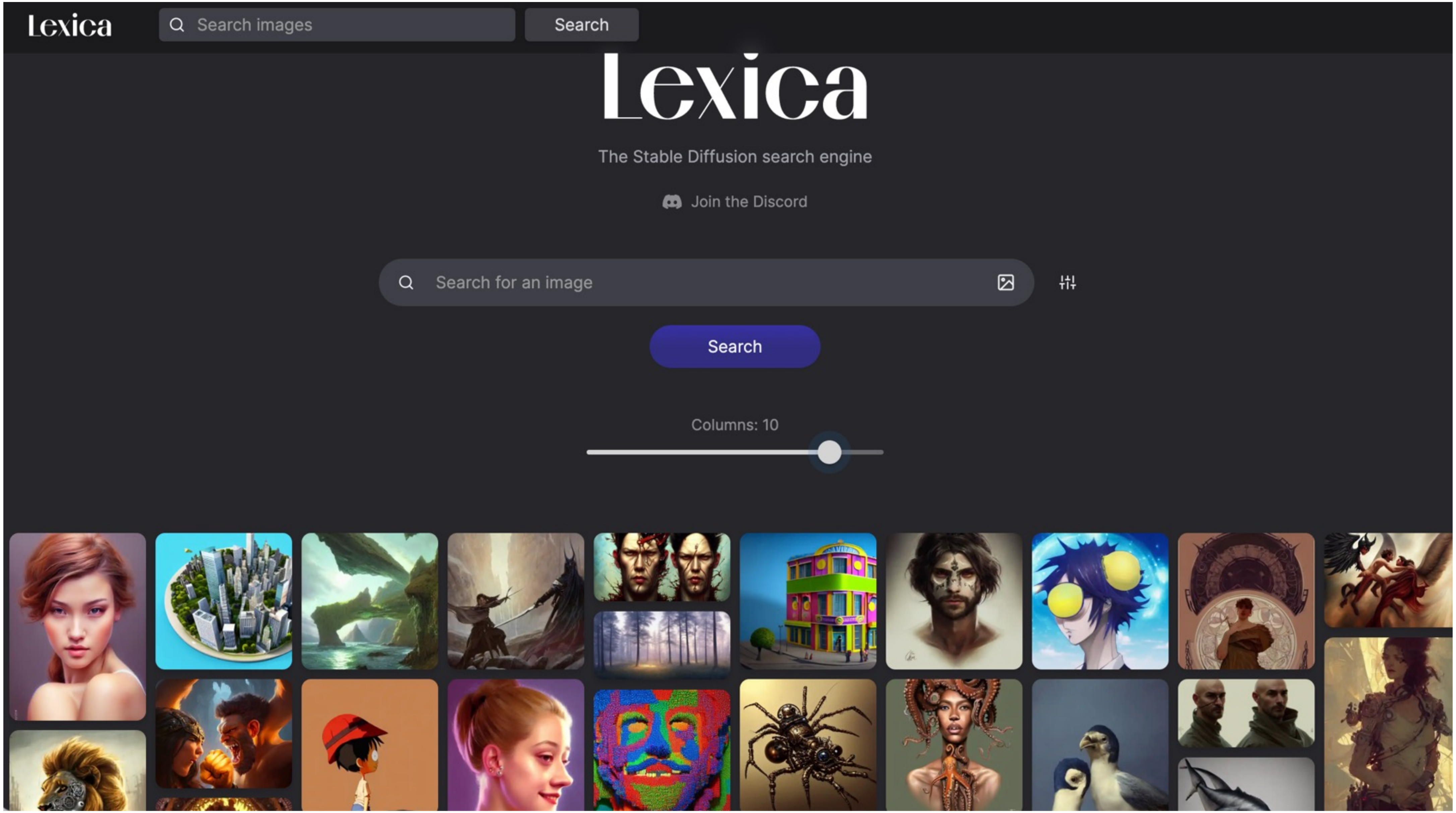
https://huggingface.co/spaces/stabilityai/stable-diffusion

Stable Diffusion Colab



https://github.com/woctezuma/stable-diffusion-colab

Lexica Art: Search Stable Diffusion images and prompts



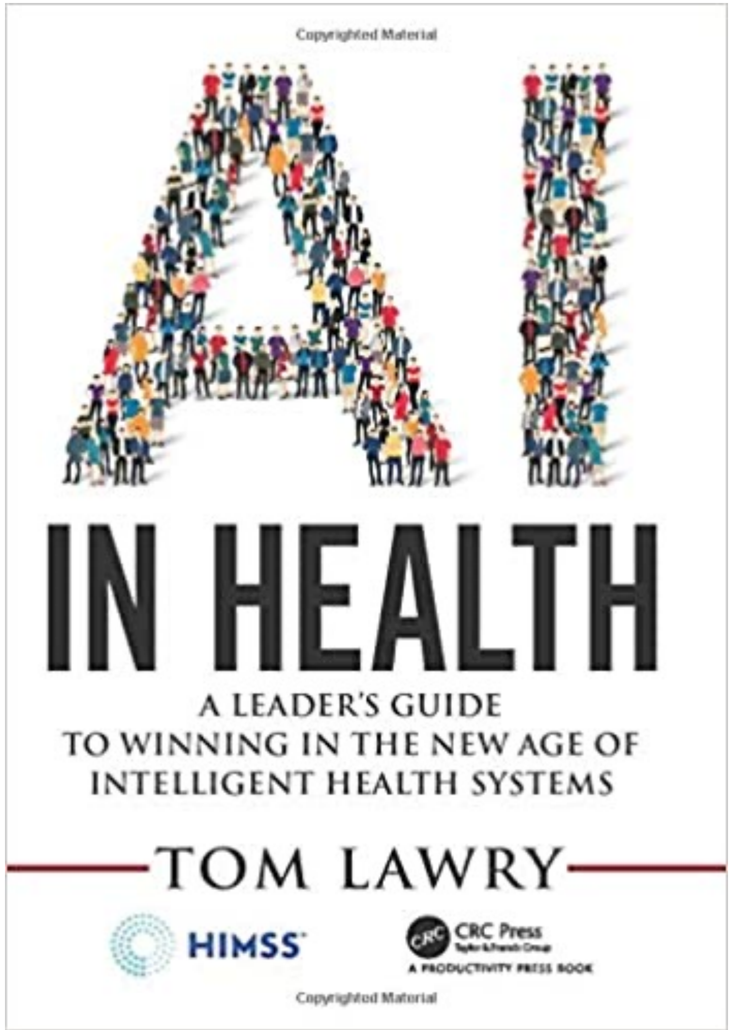
https://lexica.art/

Tom Lawry (2020),

AI in Health:

A Leader's Guide to Winning in the New Age of Intelligent Health Systems,

HIMSS Publishing



Source: Tom Lawry (2020), AI in Health: A Leader's Guide to Winning in the New Age of Intelligent Health Systems, HIMSS Publishing

https://www.amazon.com/Health-HIMSS-Book-Tom-Lawry/dp/0367333716/

AI in Healthcare

Source: Secinaro, Silvana, Davide Calandra, Aurelio Secinaro, Vivek Muthurangu, and Paolo Biancone. "The role of artificial intelligence in healthcare: a structured literature review." BMC Medical Informatics and Decision Making 21, no. 1 (2021): 1-23.

Multimodal Fall Detection

18398 IEEE SENSORS JOURNAL, VOL. 21, NO. 17, SEPTEMBER 1, 2021

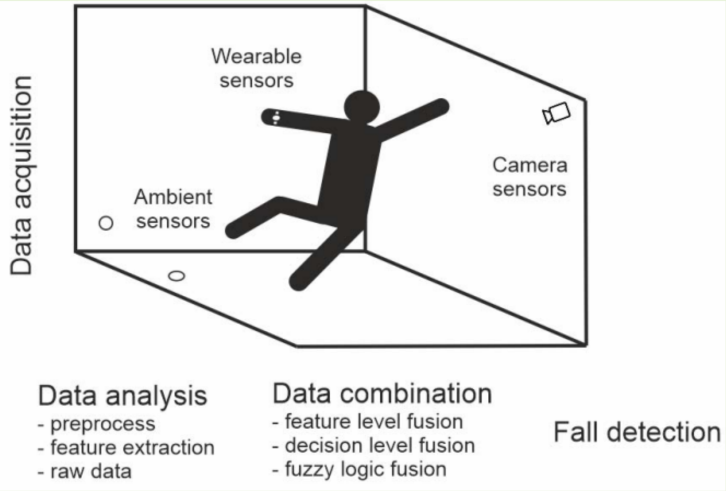
IEEE Sensors Council

Performance, Challenges, and Limitations in Multimodal Fall Detection Systems: A Review

Vasileios-Rafail Xefteris, Athina Tsanousa, Georgios Meditskos, Stefanos Vrochidis, and Ioannis Kompatsiaris

Ambient Assisted Living (AAL)

***Abstract*—Fall events among older adults are a serious concern, having an impact on their health and well-being. The development of the Internet of Things (IoT) over the last years has led to the emergence of systems able to track abnormal body movements and falls, thus facilitating fall detection and in some cases prevention. Fusing information from multiple unrelated sources is one of the recent trends in healthcare systems. This work aims to provide a survey of recent methods and trends of multisensor data fusion in fall detection systems and discuss their performance, challenges, and limitations. The paper highlights the benefits of developing multimodal systems for fall detection compared to single-sensor approaches, categorizes the different methods applied to this field, and discusses issues and trends for future work.**



***Index Terms*—Data fusion, fall detection, multisensor fusion, non-wearable sensors, wearable sensors.**

Source: Xefteris, Vasileios-Rafail, Athina Tsanousa, Georgios Meditskos, Stefanos Vrochidis, and Ioannis Kompatsiaris. "Performance, challenges, and limitations in multimodal fall detection systems: a review." IEEE Sensors Journal (2021).

Multimodal Fall Detection

Ambient Assisted Living (AAL)

Sensor modalities	Intrusion	ROI specific	Accuracy	Power needs	Computational needs	Environment affected
Wearable	Obtrusive	No	Scenario dependent	High	Low/dependent	No
Ambient	No	Yes	Scenario dependent	Low	Low/dependent	Yes
Camera	Privacy	Yes	High	Low	High	Yes

Source: Xefteris, Vasileios-Rafail, Athina Tsanousa, Georgios Meditskos, Stefanos Vrochidis, and Ioannis Kompatsiaris. "Performance, challenges, and limitations in multimodal fall detection systems: a review." IEEE Sensors Journal (2021).

Challenges of Multimodal Fall Detection

Modalities combined	Performance	Response time	Power consumption	Unaddressed issues	Other advantages
Wearable	Reasonable accuracy.	Reasonably low time.	Up to 62 days.	Obtrusiveness.	Offer to other healthcare applications, continuous monitoring.
Non-wearable	High accuracy.	Reasonably low response time.	No action needed.	ROI restriction.	No recharge power needs.
Wearable and non-wearable	High accuracy.	Low response time.	No evidence.	Complexity.	Takes advantage of both modalities, no ROI restriction.

Source: Xefteris, Vasileios-Rafail, Athina Tsanousa, Georgios Meditskos, Stefanos Vrochidis, and Ioannis Kompatsiaris. "Performance, challenges, and limitations in multimodal fall detection systems: a review." IEEE Sensors Journal (2021).

Fall Detection
Non-Wearable Sensors Fusion

Reference	Year	Sensors	Method	Evaluation	Performance
[46]	2013	PIR and PM sensors.	Graph-theoretical concepts to track user and rule-based algorithm to detect falls.	Falls and ADLs from 5 healthy young subjects.	Accuracy: 82.86%
[47]	2014	Doppler radar sensor and PIR motion sensors.	SVM classifier on Doppler radar features, rule-based algorithm to correct false alarms using PIR data.	A week of continuous data monitoring of a volunteer.	Reduced false alarms by 63% with 100% detection rate.
[48]	2018	IR sensor and an ultrasonic distance sensor.	Thermal IR and ultrasonic features, SVM classifier.	180 falls and ADLs from 3 healthy young subjects, 6 continuous recordings.	Accuracy: 96.7% (discrete test), 90.3% (continuous test).
[52]	2018	Doppler radar sensor and RGB camera.	Multiple CNN, movement classification from radar, aspect ratio sequence from camera, max voting fusion.	1 type of fall and 3 types of ADLs from 3 subjects.	Accuracy: 99.85%
[53]	2019	Doppler radar and depth camera.	Joints' coordinates from depth camera, feature extraction from joints' coordinates and radar data, Linear Discriminant Classifier.	3 different datasets.	Sensitivity: 100% (FD).

Source: Xefteris, Vasileios-Rafail, Athina Tsanousa, Georgios Meditskos, Stefanos Vrochidis, and Ioannis Kompatsiaris. "Performance, challenges, and limitations in multimodal fall detection systems: a review." IEEE Sensors Journal (2021).

Fall Detection Datasets

Datasets	Posture samples	Subject					Type sensor	year
		Number	Height(cm)	Weight(kg)	Age(year)	Gender(M/F)		
Fall detection [4]	380	4	159-182	48-85	24-31	3M-1F	RGB camera	2007
Fall detection[5]	72	2	N/A	N/A	N/A	2M	RGB camera	2008
Multicam Fall [6]	24	1	N/A	N/A	N/A	M	8 RGB camera	2010
Le2i[7]	249	10	N/A	N/A	N/A	N/A	RGB camera	2013
Thermal simulated fall [8]	35	10	N/A	N/A	N/A	N/A	Thermal camera	2016
SisFall[9]	154	45	149-183	42-102	19-75	23M-21F	RGB camera, 2 accelerometers, 1 gyroscope	2016
UR Fall Detection[10]	70	5	N/A	N/A	N/A	5M	2 Kinect camera, accelerometer	2016
NTU RGB+D Action Recognition [11]	56880	302	N/A	N/A	N/A	N/A	Kinect camera v2	2016
UMA Fall [12]	531	17	155-195	50-93	18-55	10M-7F	Mobility sensors (smartphone)	2017
CMD Fall [13]	20	50	N/A	N/A	21-40	30M-20F	Kinect camera, accelerometer	2018
TST Fall Detection Dataset V2[8]	264	11	N/A	N/A	N/A	N/A	Microsoft Kinect v2, accelerometer	2018
UP-Fall[14]	561	17	N/A	N/A	22-58	N/A	Infrared ,inertial measurement	2019

Note: N/A_Not Available; M_Male; F_Femal

Source: Oumaima, Guendoul, Ait Abdelali Hamd, Tabii Youness, Oulad Haj Thami Rachid, and Bourja Omar. "Vision-based fall detection and prevention for the elderly people: A review & ongoing research." In 2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS), pp. 1-6. IEEE, 2021.

Human Action Recognition (HAR)

Human Action Recognition from Various Data Modalities: A Review

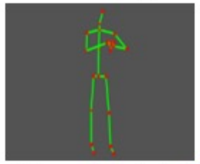
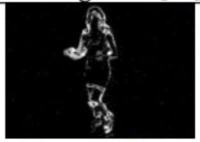
Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu

Abstract—Human Action Recognition (HAR) aims to understand human behavior and assign a label to each action. It has a wide range of applications, and therefore has been attracting increasing attention in the field of computer vision. Human actions can be represented using various data modalities, such as RGB, skeleton, depth, infrared, point cloud, event stream, audio, acceleration, radar, and WiFi signal, which encode different sources of useful yet distinct information and have various advantages depending on the application scenarios. Consequently, lots of existing works have attempted to investigate different types of approaches for HAR using various modalities. In this paper, we present a comprehensive survey of recent progress in deep learning methods for HAR based on the type of input data modality. Specifically, we review the current mainstream deep learning methods for single data modalities and multiple data modalities, including the fusion-based and the co-learning-based frameworks. We also present comparative results on several benchmark datasets for HAR, together with insightful observations and inspiring future research directions.

Index Terms—Human Action Recognition, Deep Learning, Data Modality, Single Modality, Multi-modality.

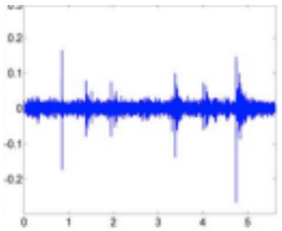
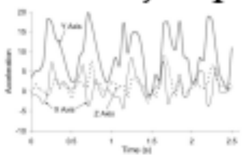
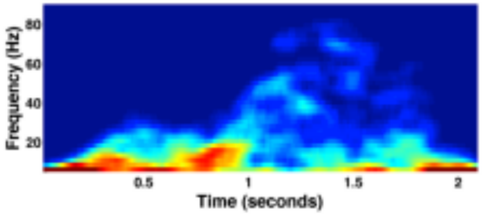
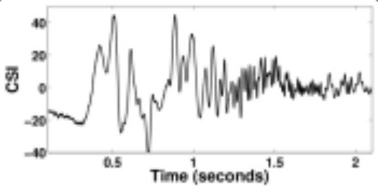
Source: Sun, Zehua, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. "Human action recognition from various data modalities: A review." IEEE transactions on pattern analysis and machine intelligence (2022).

Human Action Recognition (HAR)
Modality

Modality		Example	Pros	Cons
Visual Modality	RGB	 Hand-waving [27]	· Provide rich appearance information · Easy to obtain and operate · Wide range of applications	· Sensitive to viewpoint · Sensitive to background · Sensitive to illumination
	3D Skeleton	 Looking at watch [28]	· Provide 3D structural information of subject pose · Simple yet informative · Insensitive to viewpoint · Insensitive to background	· Lack of appearance information · Lack of detailed shape information · Noisy
	Depth	 Mopping floor [29]	· Provide 3D structural information · Provide geometric shape information	· Lack of color and texture information · Limited workable distance
	Infrared Sequence	 Pushing [30]	· Workable in dark environments	· Lack of color and texture information · Susceptible to sunlight
	Point Cloud	 Bending over [31]	· Provide 3D information · Provide geometric shape information · Insensitive to viewpoint	· Lack of color and texture information · High computational complexity
	Event Stream	 Running [32]	· Avoid much visual redundancy · High dynamic range · No motion blur	· Asynchronous output · Spatio-temporally sparse · Capturing device is relatively expensive

Source: Sun, Zehua, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. "Human action recognition from various data modalities: A review." IEEE transactions on pattern analysis and machine intelligence (2022).

Human Action Recognition (HAR) Modality

Non-visual Modality	Audio	 Audio wave of jumping [33]	· Easy to locate actions in temporal sequence	· Lack of appearance information
	Acceleration	 Acceleration measurements of walking [34]	· Can be used for fine-grained HAR · Privacy protecting · Low cost	· Lack of appearance information · Capturing device needs to be carried by subject
	Radar	 Spectrogram of falling [35]	· Can be used for through-wall HAR · Insensitive to illumination · Insensitive to weather · Privacy protecting	· Lack of appearance information · Capturing device is relatively expensive
	WiFi	 CSI waveform of falling [35]	· Simple and convenient · Privacy protecting · Low cost	· Lack of appearance information · Sensitive to environments · Noisy

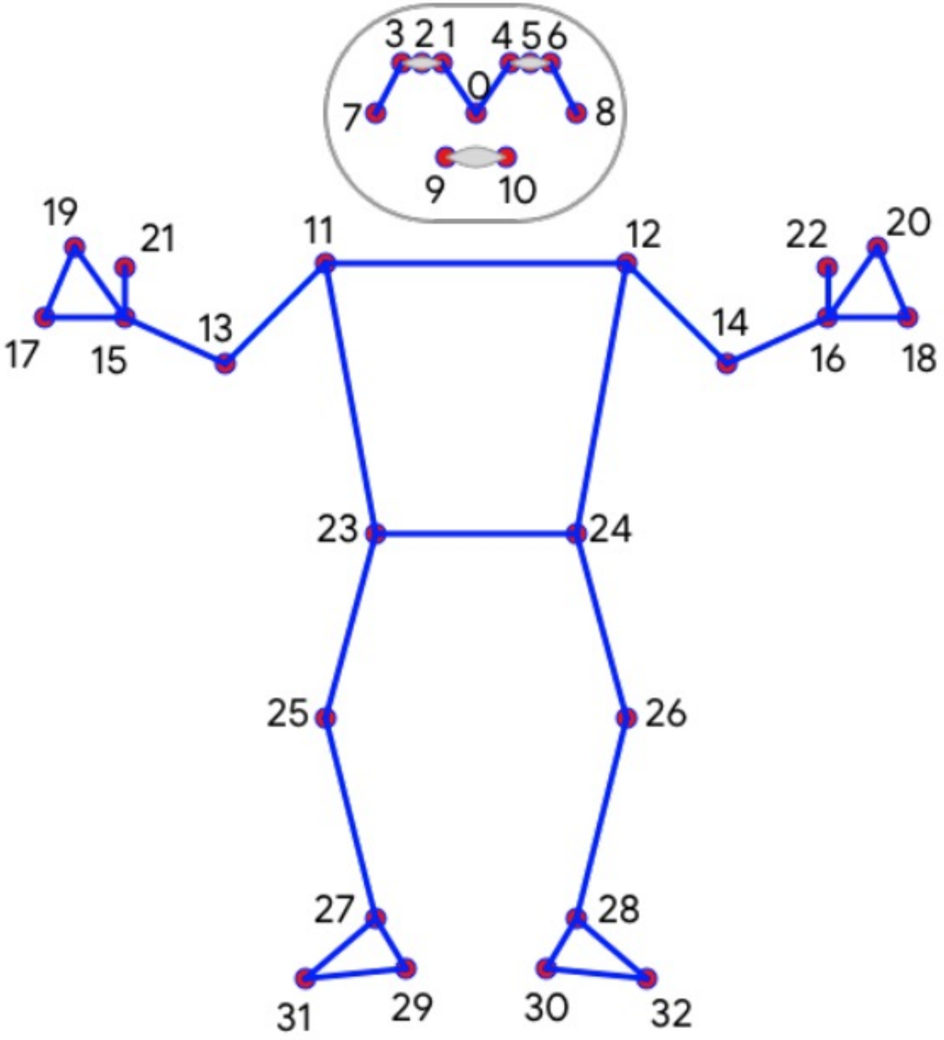
Source: Sun, Zehua, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. "Human action recognition from various data modalities: A review." IEEE transactions on pattern analysis and machine intelligence (2022).

Fall Detection

Source: https://www.ankecare.com/article/1929-2022-05-09-08-29-47

BlazePose:

On-device Real-time Body Pose tracking



BlazePose 33 Keypoint topology

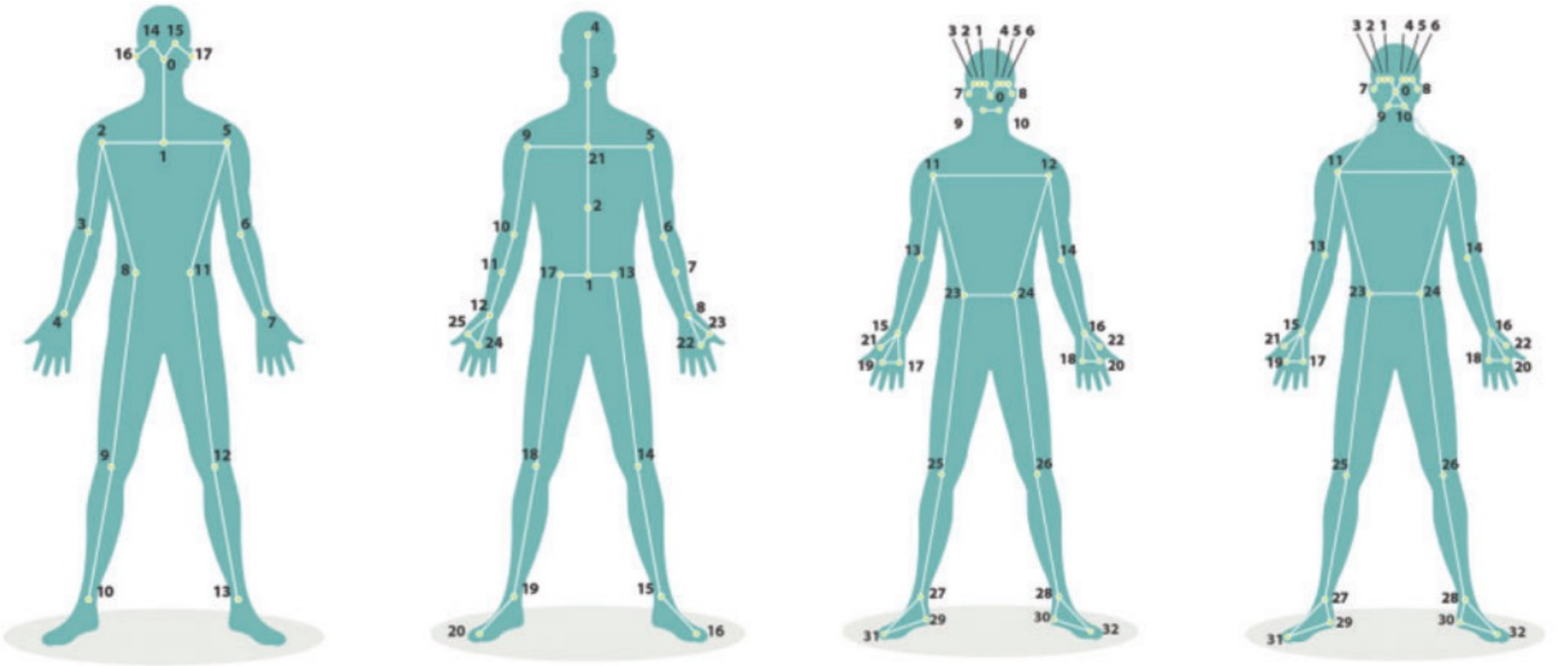
0. Nose
1. Left eye inner
2. Left eye
3. Left eye outer
4. Right eye inner
5. Right eye
6. Right eye outer
7. Left ear
8. Right ear
9. Mouth left
10. Mouth right
11. Left shoulder
12. Right shoulder
13. Left elbow
14. Right elbow
15. Left wrist
16. Right wrist
17. Left pinky #1 knuckle
18. Right pinky #1 knuckle
19. Left index #1 knuckle
20. Right index #1 knuckle
21. Left thumb #2 knuckle
22. Right thumb #2 knuckle
23. Left hip
24. Right hip
25. Left knee
26. Right knee
27. Left ankle
28. Right ankle
29. Left heel
30. Right heel
31. Left foot index
32. Right foot index

SourceBazarevsky, Valentin, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. "Blazepose: On-device real-time body pose tracking." arXiv preprint arXiv:2006.10204 (2020).

BlazePose results on yoga and fitness poses

SourceBazarevsky, Valentin, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. "Blazepose: On-device real-time body pose tracking." arXiv preprint arXiv:2006.10204 (2020).

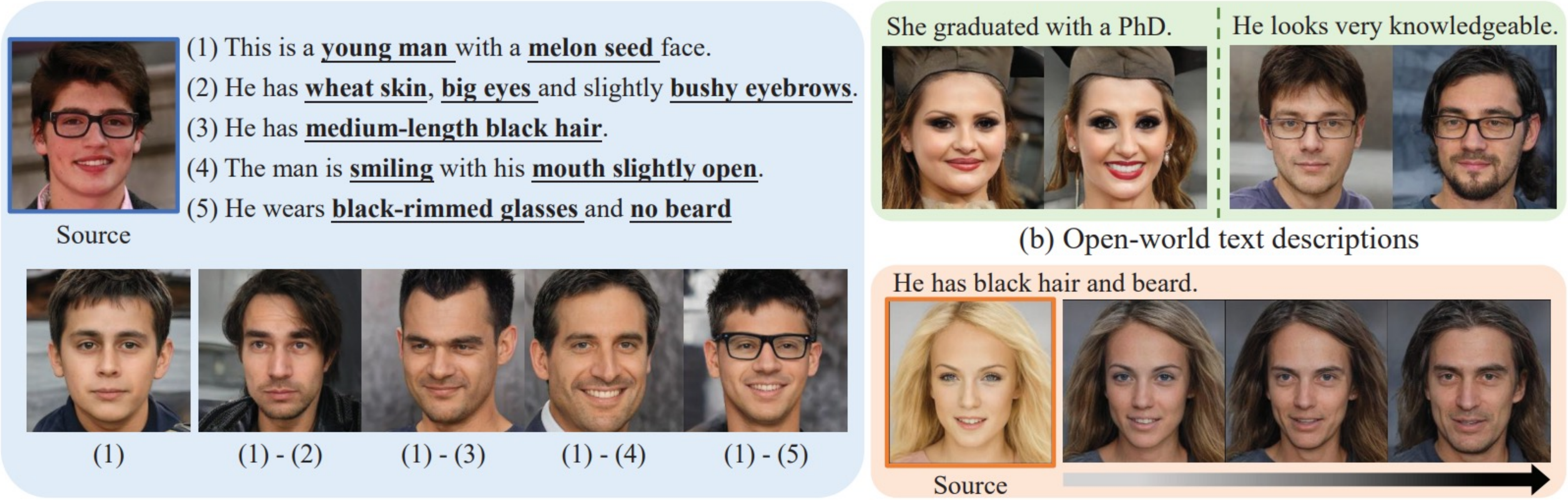
OpenPose vs. BlazePose



a) OpenPose COCO   b) NTU-RGB+D   c) BlazePose   d) Enhanced-BlazePose

Source: Alsawadi, Motasem S., El-Sayed M. El-Kenawy, and Miguel Rio. "Using BlazePose on Spatial Temporal Graph Convolutional Networks for Action Recognition." Computers, Materials and Continua 74, no. 1 (2022): 19-36.

AnyFace: Free-style Text-to-Face Synthesis and Manipulation



(a) One caption *vs* Multi-caption

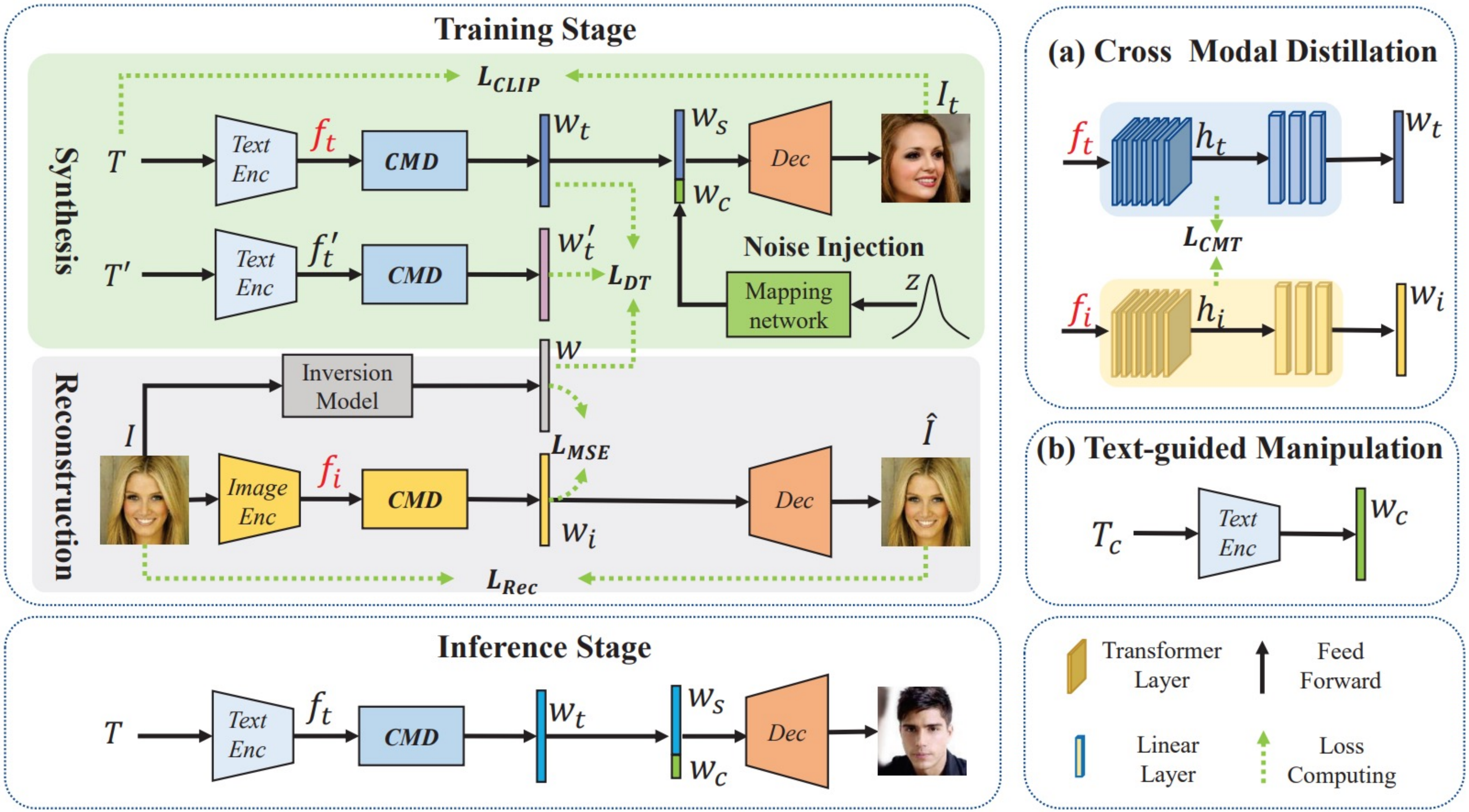
(b) Open-world text descriptions

(c) Face manipulation

Methods	AttnGAN [31]	DFGAN [25]	RiFeGAN [1]	SEA-T2F [24]	CIGAN [28]	TediGAN-B [30]	**AnyFace**
Single Model	✓	✓	✓	✓	✓	-	✓
One Generator	-	✓	-	-	✓	✓	✓
Multi-caption	-	-	✓	✓	-	-	✓
High Resolution	-	-	-	-	✓	✓	✓
Manipulation	-	-	-	-	✓	✓	✓
Open-world	-	-	-	-	-	✓	✓

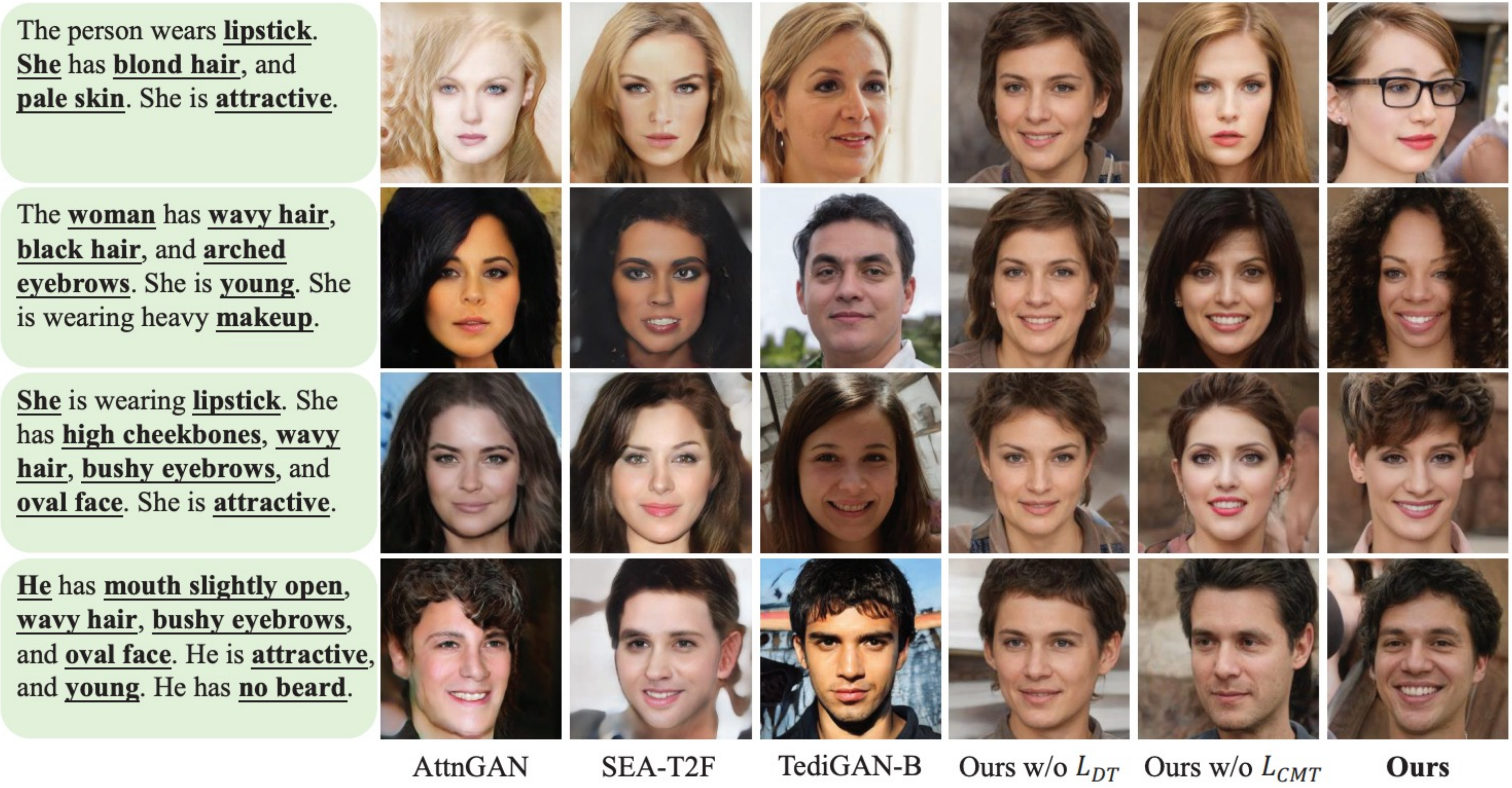
Source: Sun, Jianxin, Qiyao Deng, Qi Li, Muyi Sun, Min Ren, and Zhenan Sun. (2022) "AnyFace: Free-style Text-to-Face Synthesis and Manipulation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18687-18696.

AnyFace: Free-style Text-to-Face Synthesis and Manipulation



Source: Sun, Jianxin, Qiyao Deng, Qi Li, Muyi Sun, Min Ren, and Zhenan Sun. (2022) "AnyFace: Free-style Text-to-Face Synthesis and Manipulation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18687-18696.

AnyFace: Free-style Text-to-Face Synthesis and Manipulation



Source: Sun, Jianxin, Qiyao Deng, Qi Li, Muyi Sun, Min Ren, and Zhenan Sun. (2022) "AnyFace: Free-style Text-to-Face Synthesis and Manipulation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18687-18696.

AnyFace: Free-style Text-to-Face Synthesis and Manipulation

Source: Sun, Jianxin, Qiyao Deng, Qi Li, Muyi Sun, Min Ren, and Zhenan Sun. (2022)
"AnyFace: Free-style Text-to-Face Synthesis and Manipulation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18687-18696.

AnyFace: Free-style Text-to-Face Synthesis and Manipulation

Text-guided Face Manipulation

Source: Sun, Jianxin, Qiyao Deng, Qi Li, Muyi Sun, Min Ren, and Zhenan Sun. (2022) "AnyFace: Free-style Text-to-Face Synthesis and Manipulation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18687-18696.

Papers with Code
State-of-the-Art (SOTA)

Computer Vision

Task	Benchmarks	Papers with code
Semantic Segmentation	185 benchmarks	3397 papers with code
Image Classification	390 benchmarks	2778 papers with code
Object Detection	269 benchmarks	2559 papers with code
Contrastive Learning	2 benchmarks	1119 papers with code
Image Generation	208 benchmarks	1097 papers with code

▸ **See all 1415 tasks**

Natural Language Processing

Task	Benchmarks	Papers with code
Language Modelling	458 benchmarks	2248 papers with code
Question Answering	181 benchmarks	1818 papers with code
Machine Translation	78 benchmarks	1721 papers with code
Sentiment Analysis	87 benchmarks	1040 papers with code
Text Generation	242 benchmarks	931 papers with code

▸ **See all 664 tasks**

https://paperswithcode.com/sota

Papers with Code
State-of-the-Art (SOTA)

Computer Vision

- 3425 benchmarks
- 1088 tasks
- 2320 datasets
- 29741 papers with code

https://paperswithcode.com/area/computer-vision

Computer Vision: State-of-the-Art (SOTA)

Image Classification

Image Classification
390 benchmarks
2780 papers with code

Knowledge Distillation
3 benchmarks
724 papers with code

OOD Detection
166 papers with code

Few-Shot Image Classification
95 benchmarks
156 papers with code

Fine-Grained Image Classification
35 benchmarks
130 papers with code

▸ See all 26 tasks

Object Detection

Object Detection
268 benchmarks
2562 papers with code

3D Object Detection
61 benchmarks
342 papers with code

RGB Salient Object Detection
33 benchmarks
90 papers with code

Real-Time Object Detection
9 benchmarks
85 papers with code

Few-Shot Object Detection
6 benchmarks
52 papers with code

▸ See all 34 tasks

https://paperswithcode.com/area/computer-vision

Computer Vision: State-of-the-Art (SOTA)

Image Classification

Task	Benchmarks	Papers with code
Image Classification	390 benchmarks	2780 papers with code
Knowledge Distillation	3 benchmarks	724 papers with code
OOD Detection		166 papers with code
Few-Shot Image Classification	95 benchmarks	156 papers with code
Fine-Grained Image Classification	35 benchmarks	130 papers with code

▸ See all 26 tasks

Object Detection

Task	Benchmarks	Papers with code
Object Detection	268 benchmarks	2562 papers with code
3D Object Detection	61 benchmarks	342 papers with code
RGB Salient Object Detection	33 benchmarks	90 papers with code
Real-Time Object Detection	9 benchmarks	85 papers with code
Few-Shot Object Detection	6 benchmarks	52 papers with code

▸ See all 34 tasks

https://paperswithcode.com/area/computer-vision

Computer Vision: State-of-the-Art (SOTA)

Image Generation

Image Generation
208 benchmarks
1097 papers with code

Image-to-Image Translation
57 benchmarks
388 papers with code

Image Inpainting
18 benchmarks
198 papers with code

Conditional Image Generation
10 benchmarks
105 papers with code

Face Generation
11 benchmarks
88 papers with code

▸ **See all 18 tasks**

Pose Estimation

Pose Estimation
482 benchmarks
968 papers with code

3D Human Pose Estimation
380 benchmarks
215 papers with code

Keypoint Detection
7 benchmarks
114 papers with code

3D Pose Estimation
6 benchmarks
106 papers with code

6D Pose Estimation
4 benchmarks
73 papers with code

▸ **See all 18 tasks**

https://paperswithcode.com/area/computer-vision

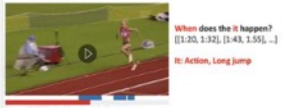
Computer Vision: Video

State-of-the-Art (SOTA)

Object Tracking

55 benchmarks

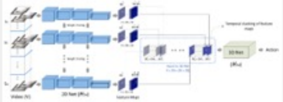
389 papers with code



Temporal Action Localization

273 benchmarks

332 papers with code



Video Understanding

2 benchmarks

186 papers with code

Action Classification

49 benchmarks

184 papers with code

Video Object Segmentation

47 benchmarks

171 papers with code

Video Retrieval

17 benchmarks

151 papers with code

Video Classification

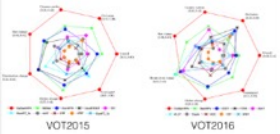
143 benchmarks

138 papers with code

Video Prediction

15 benchmarks

138 papers with code



Visual Object Tracking

20 benchmarks

115 papers with code

Video Generation

15 benchmarks

109 papers with code

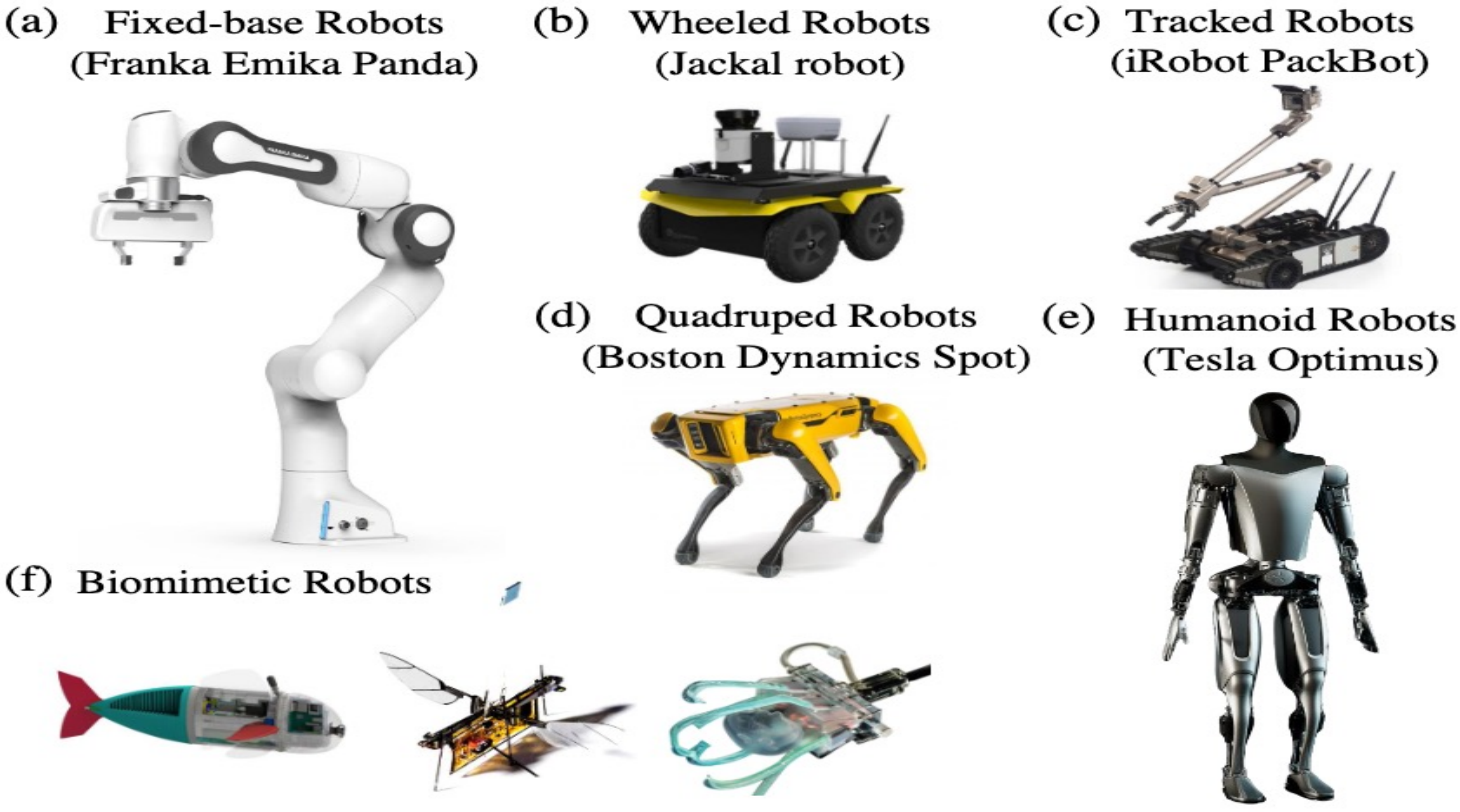
https://paperswithcode.com/area/computer-vision/video

Robotics

Artificial Intelligence: Robotics

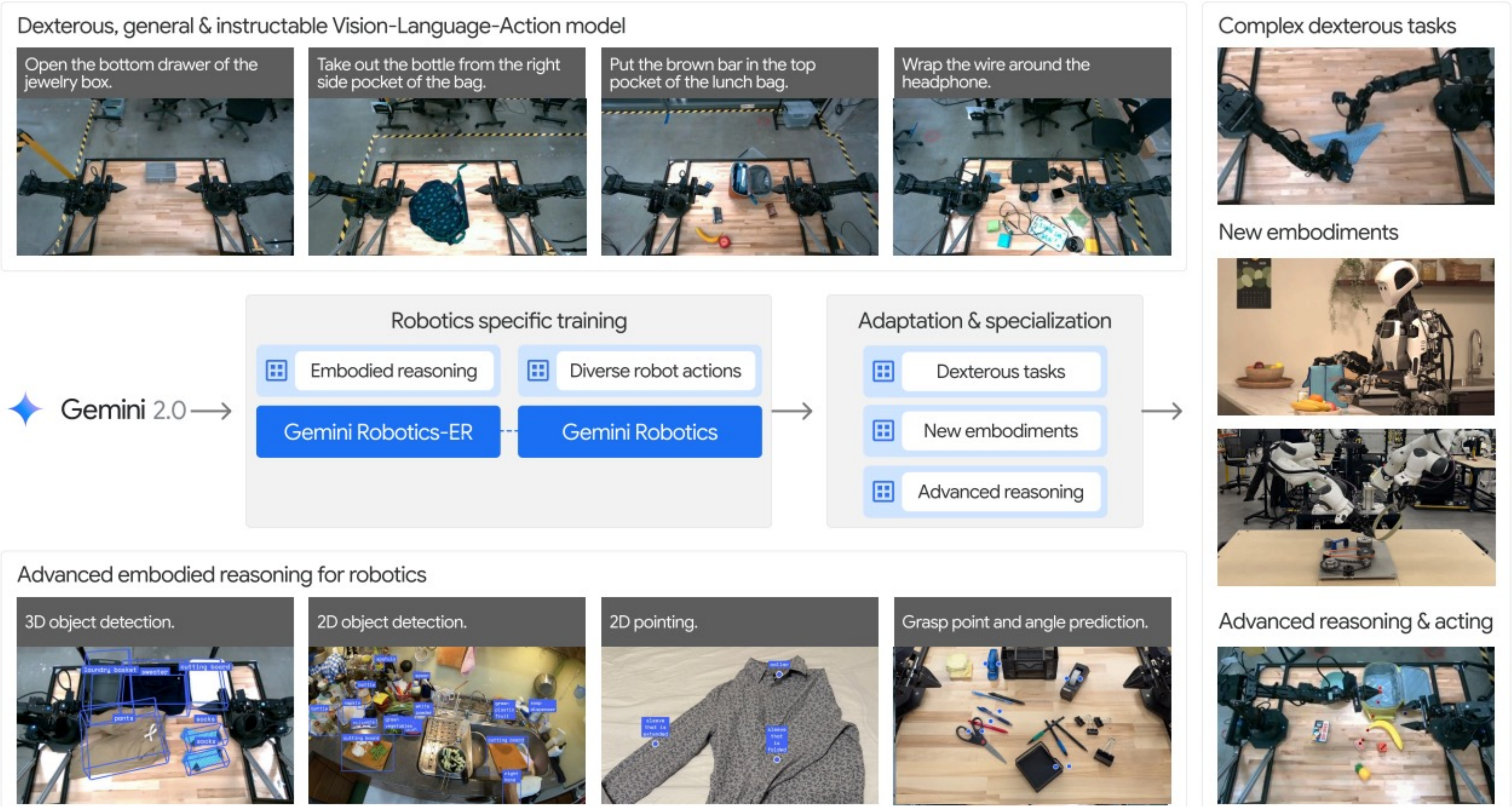
- **Agents** are endowed with **sensors** and **physical effectors** with which to move about and make mischief in the real world.

Embodied Robots



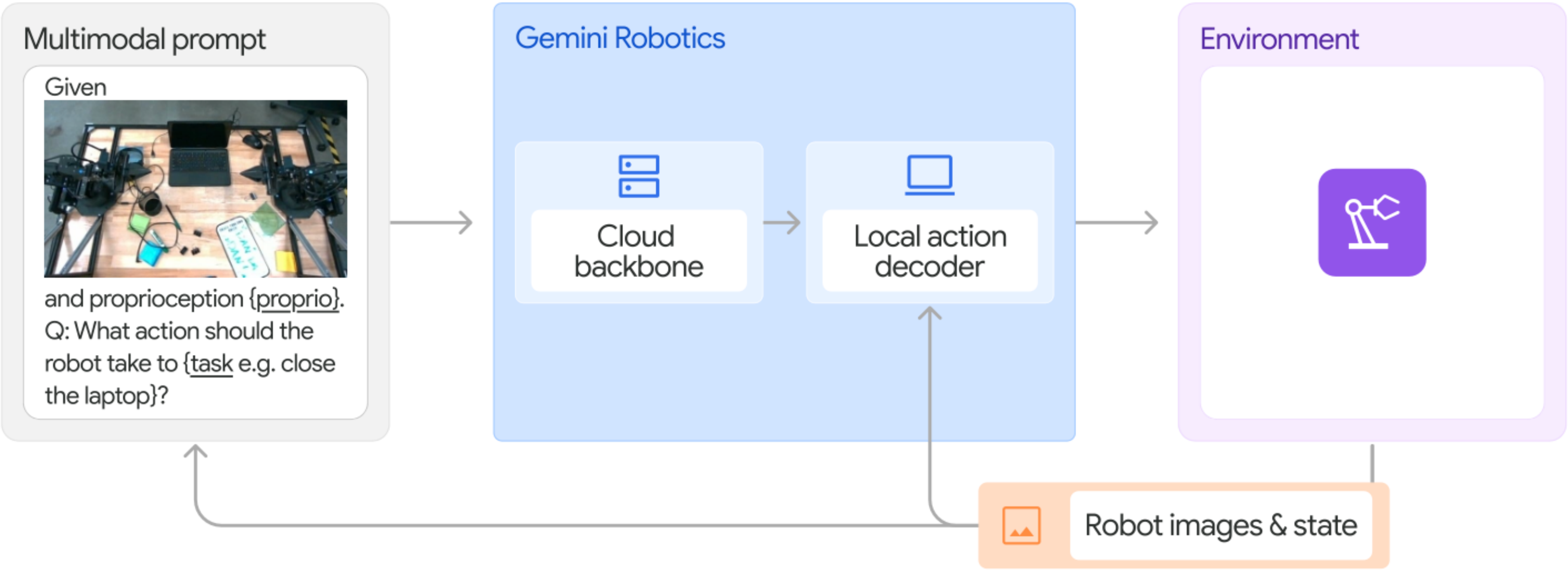
Source: Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. (2024) "Aligning cyber space with physical world: A comprehensive survey on embodied ai." arXiv preprint arXiv:2407.06886.

Gemini Robotics: Bringing AI into the Physical World



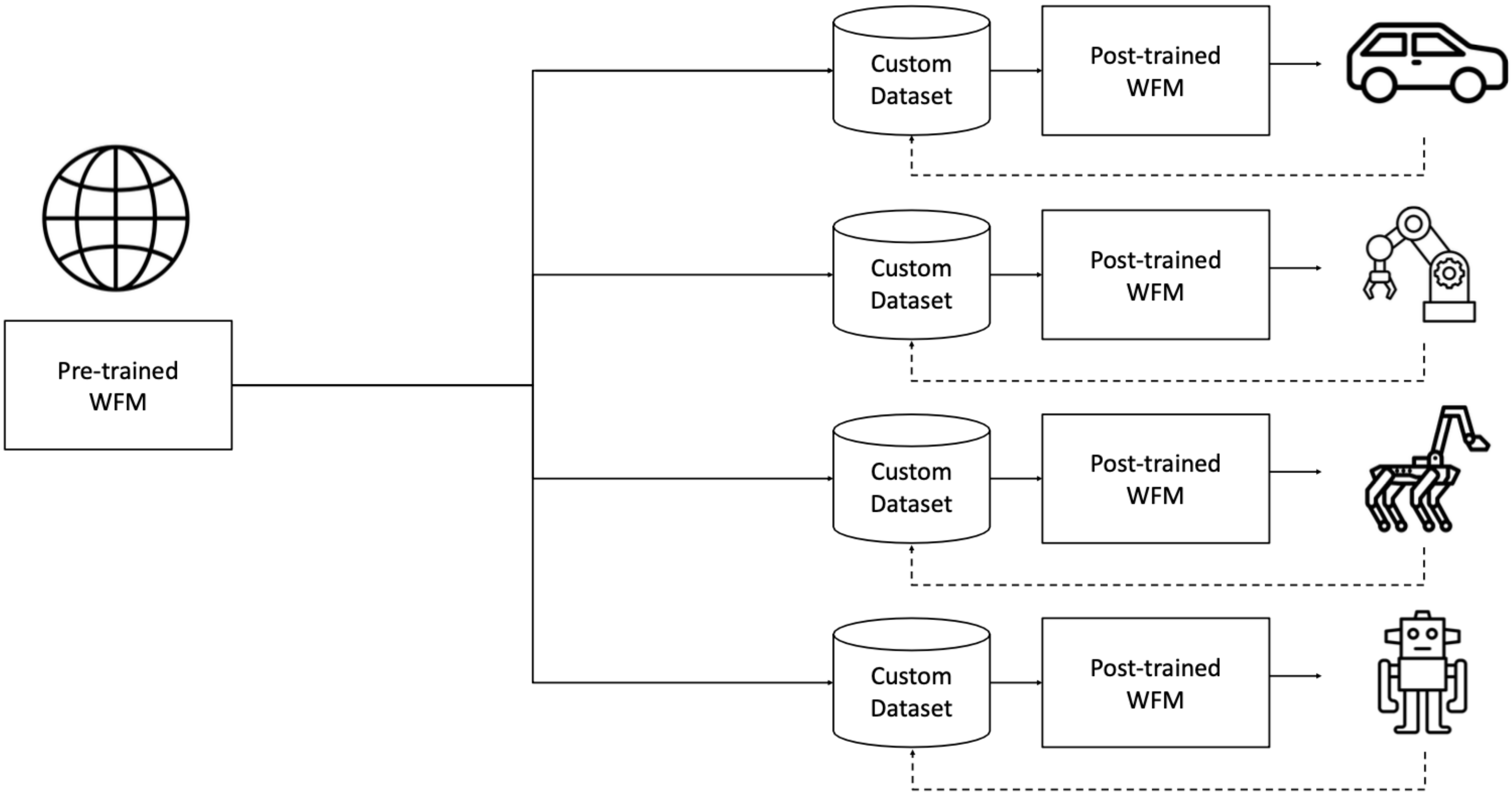
Source: Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna et al.(2025) "Gemini robotics: Bringing ai into the physical world." arXiv preprint arXiv:2503.20020 (2025).

Gemini Robotics Models: Architecture, Input and Output



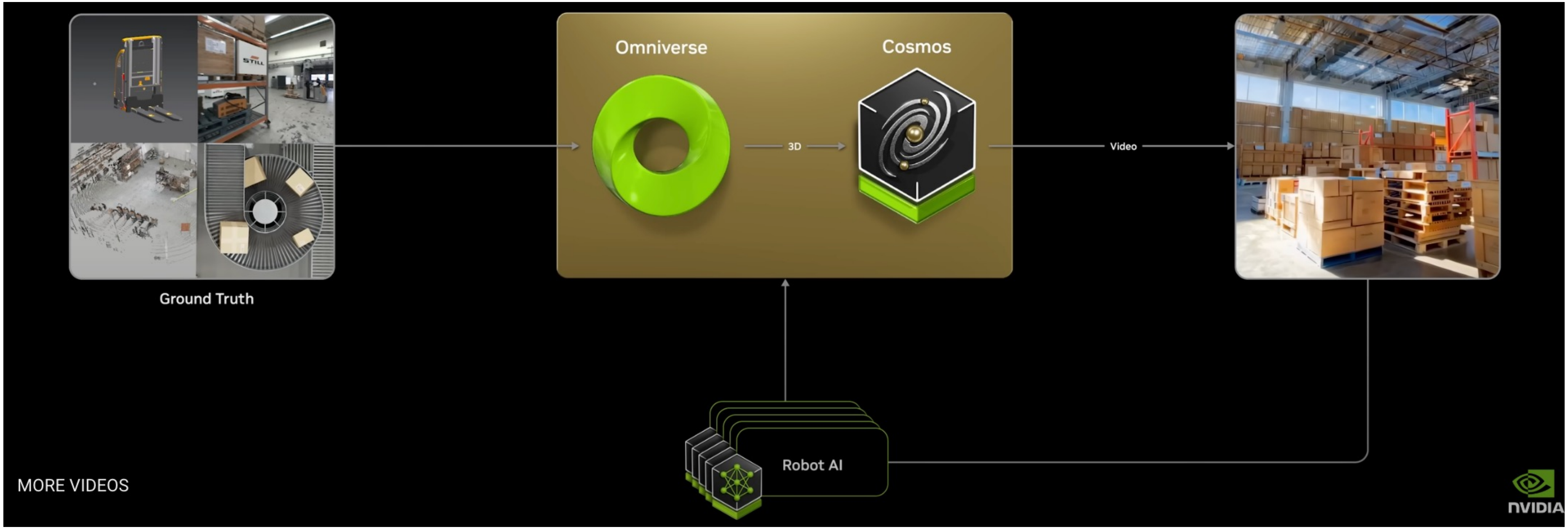
Source: Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna et al.(2025) "Gemini robotics: Bringing ai into the physical world." arXiv preprint arXiv:2503.20020 (2025).

World Foundation Model Platform for Physical AI



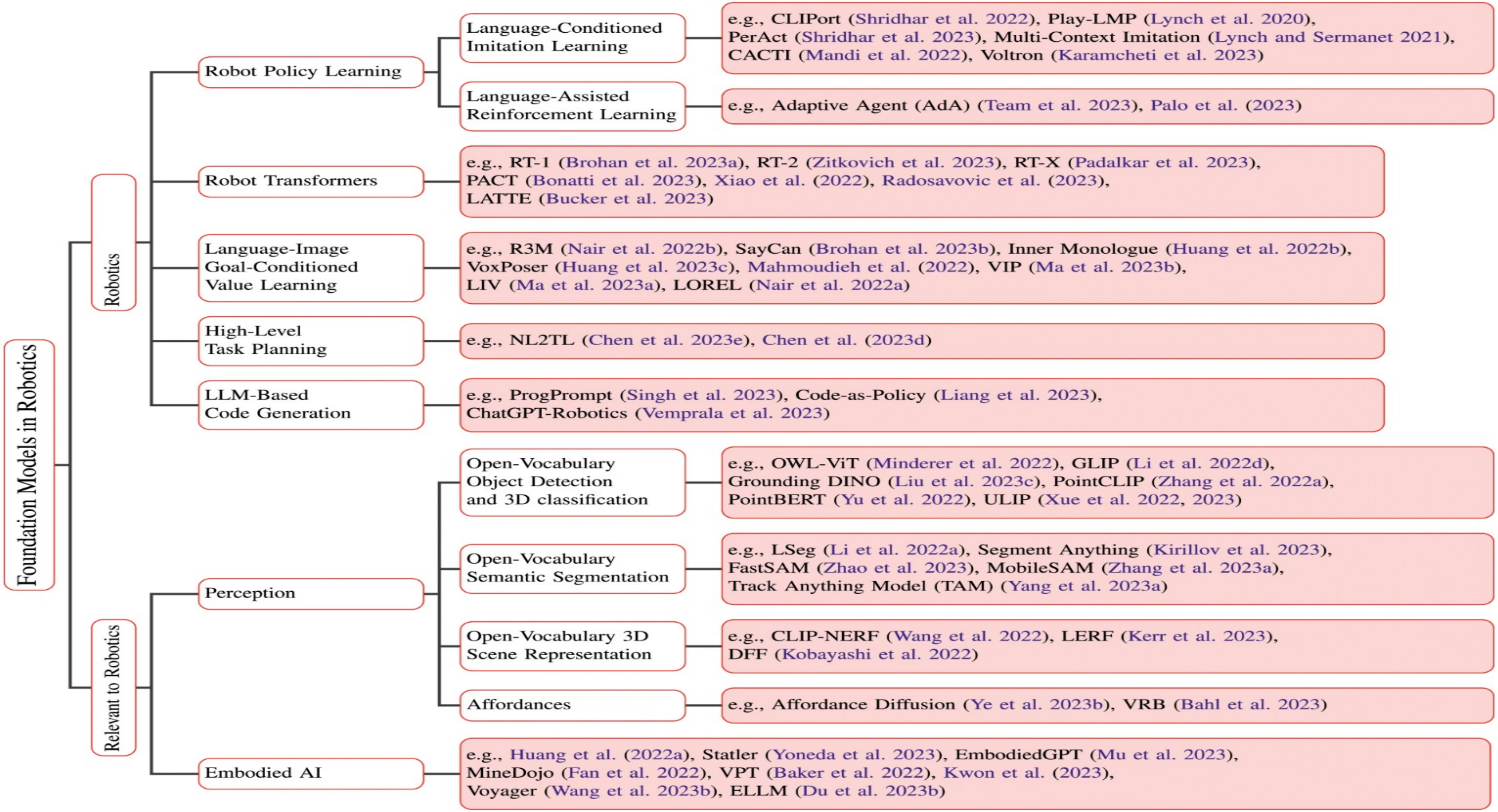
Source: Agarwal, Niket, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay et al. "Cosmos world foundation model platform for physical ai." arXiv preprint arXiv:2501.03575 (2025).

NVIDIA Cosmos
World Foundation Model Platform for Physical AI



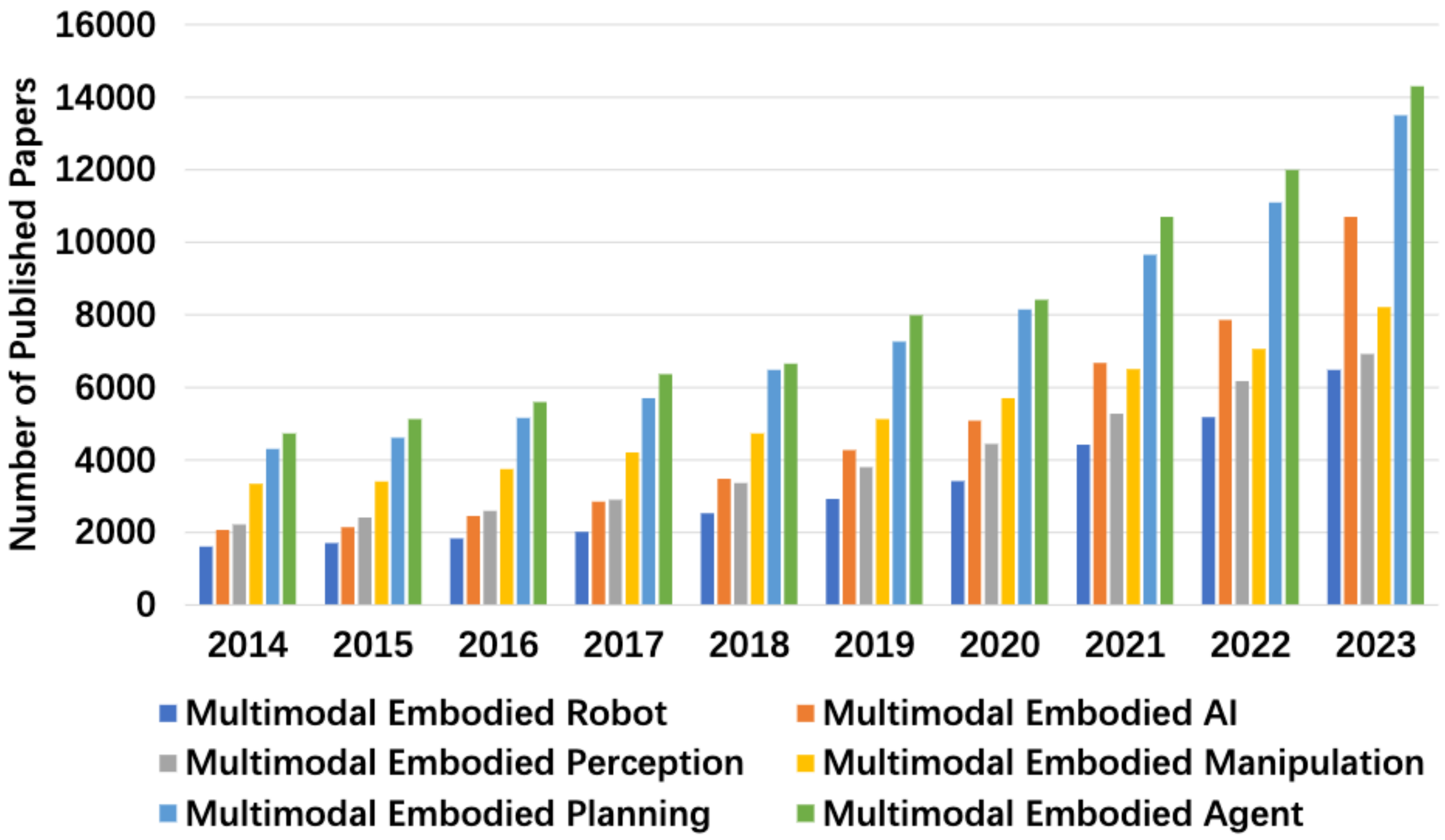
Source: Agarwal, Niket, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay et al. "Cosmos world foundation model platform for physical ai." arXiv preprint arXiv:2501.03575 (2025)., https://research.nvidia.com/labs/dir/cosmos-predict1/

Foundation Models in Robotics



Source: Firoozi, Roya, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu et al. "Foundation models in robotics: Applications, challenges, and the future." The International Journal of Robotics Research 44, no. 5 (2025): 701-739.

Robotics and Embodied AI



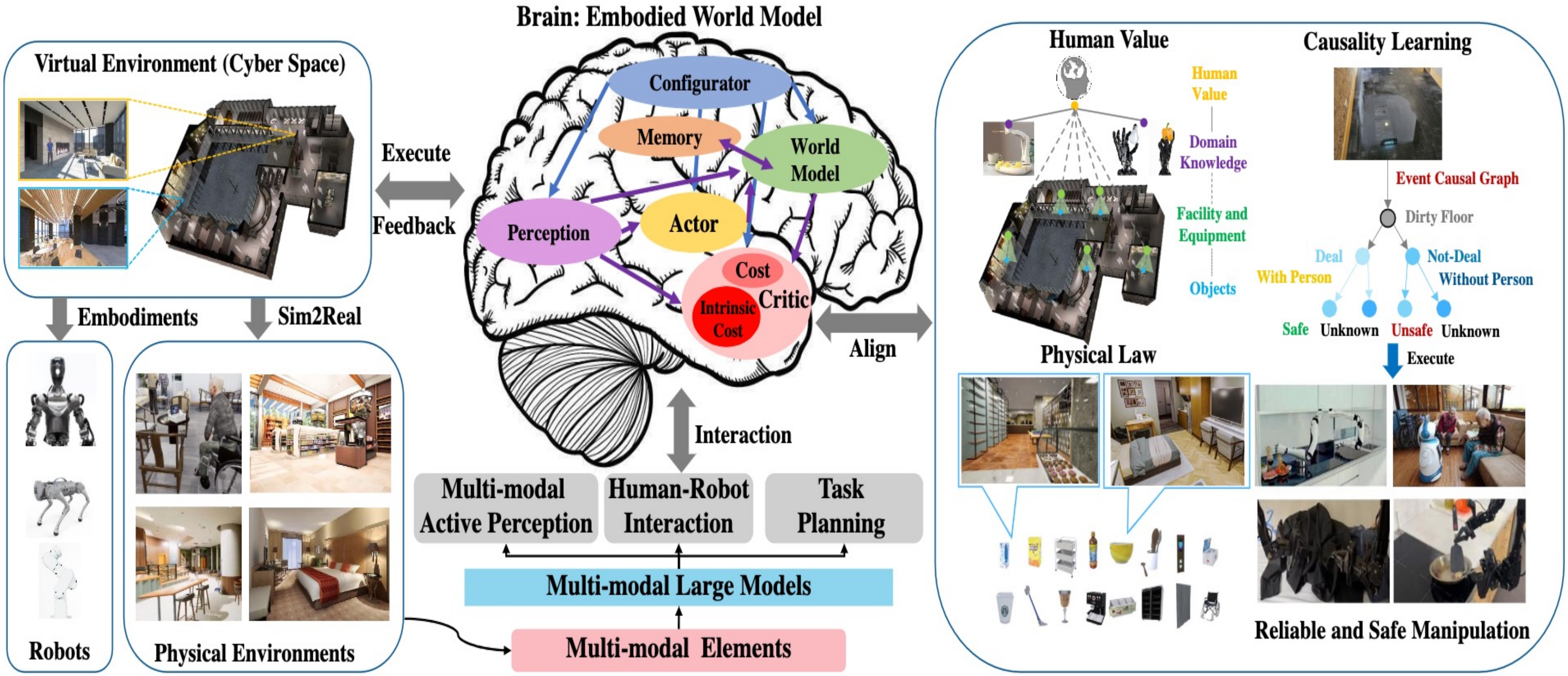
Source: Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. (2024) "Aligning cyber space with physical world: A comprehensive survey on embodied ai." arXiv preprint arXiv:2407.06886.

Embodied AI and Disembodied AI

Type	Environment	Physical Entities	Description	Representative Agents
Disembodied AI	Cyber Space	No	Cognition and physical entities are disentangled	ChatGPT [9], RoboGPT [10]
Embodied AI	Physical Space	Robots, Cars, Other devices	Cognition is integrated into physical entities	RT-1 [11], RT-2 [3], RT-H [4]

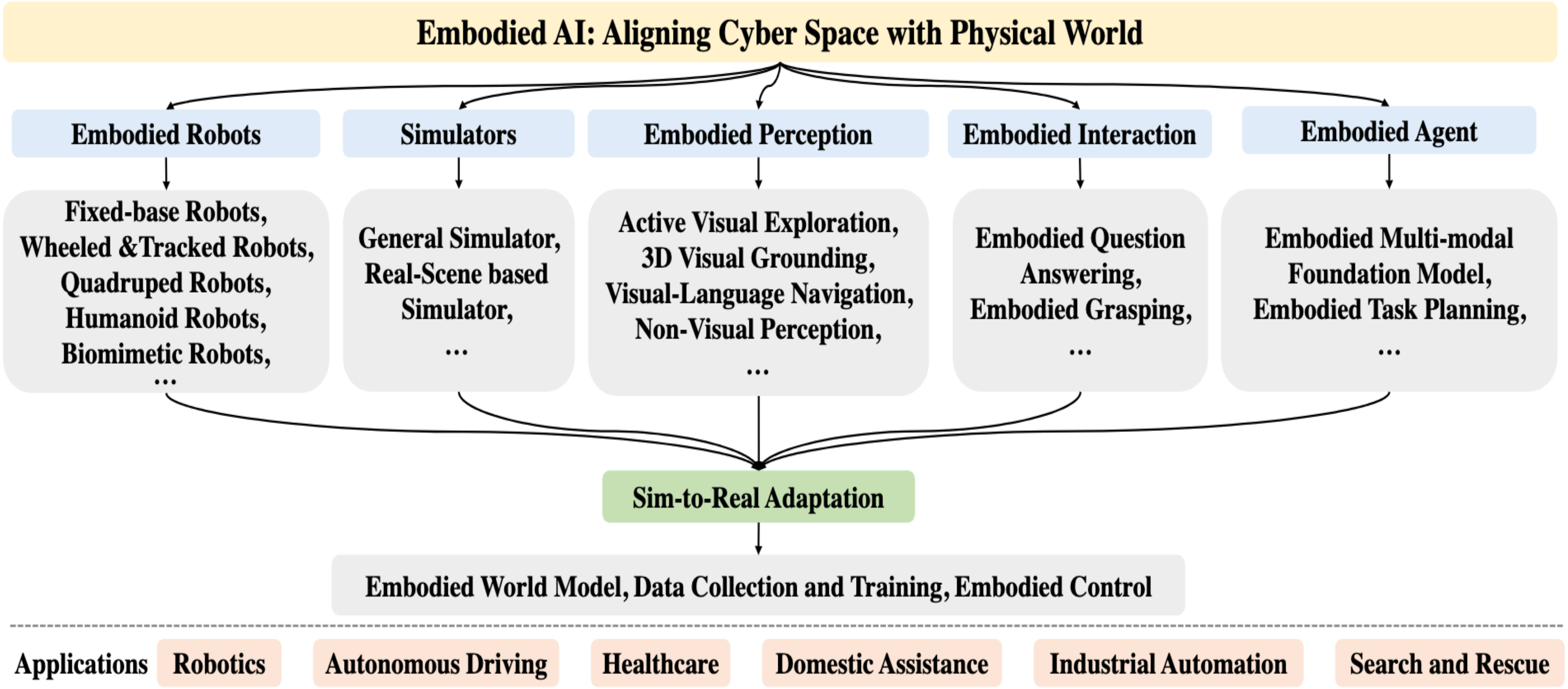
Source: Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. (2024) "Aligning cyber space with physical world: A comprehensive survey on embodied ai." arXiv preprint arXiv:2407.06886.

Framework of the Embodied Agent based on MLMs and WMs



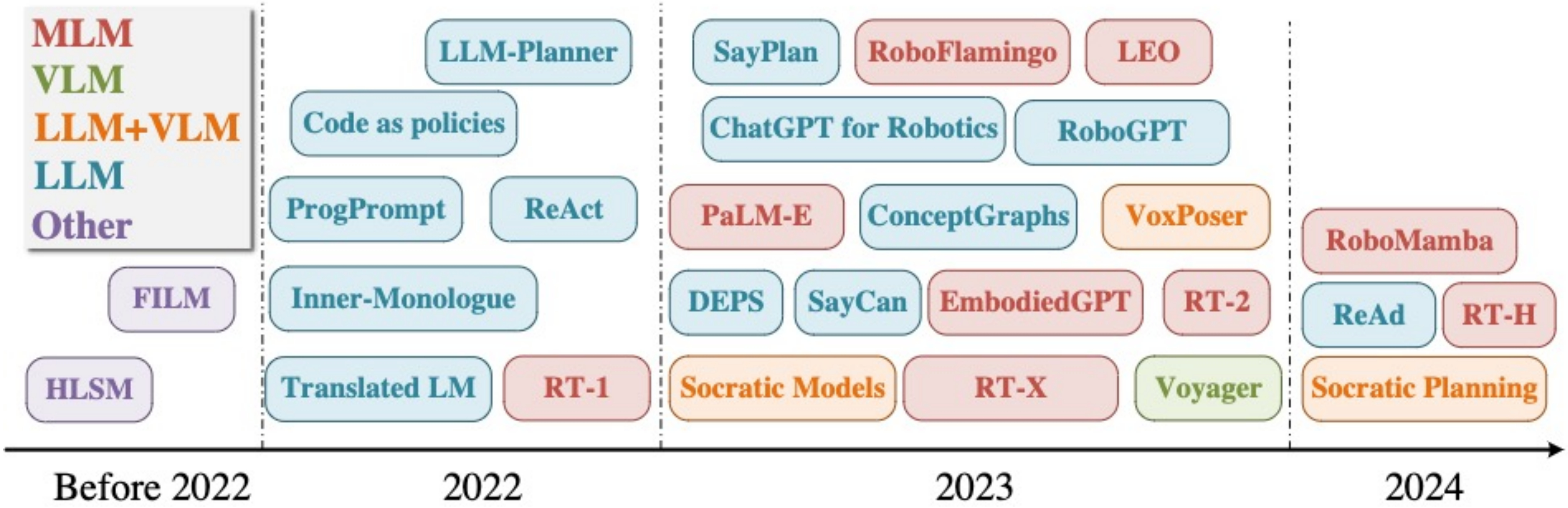
Source: Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. (2024) "Aligning cyber space with physical world: A comprehensive survey on embodied ai." arXiv preprint arXiv:2407.06886.

Embodied AI



Source: Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. (2024) "Aligning cyber space with physical world: A comprehensive survey on embodied ai." arXiv preprint arXiv:2407.06886.

Embodied Agents



MLM: Multimodal Language Model, which directly perceive the world and control the embodiment
VLM: Visual-Language Model with the outer policy models
LLM + VLM: LLM-based agent that perceives the world utilizing the VLM, and LLM means the Large-Language Model with visual context and outer policy models.

Source: Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. (2024) "Aligning cyber space with physical world: A comprehensive survey on embodied ai." arXiv preprint arXiv:2407.06886.

Boston Dynamics: Spot

Automate sensing and inspection, capture limitless data, and explore without boundaries.

Source: https://www.bostondynamics.com/spot

Boston Dynamics: Atlas

The world’s most dynamic humanoid robot

Atlas is a research platform designed to push the limits of whole-body mobility

Source: https://www.bostondynamics.com/atlas

Boston Dynamics: Atlas Goes Hands On

Atlas uses a machine learning (ML) vision model to detect and localize the environment fixtures and individual bins.

The robot uses a specialized grasping policy and continuously estimates the state of manipulated objects to achieve the task.

Source: https://bostondynamics.com/video/atlas-goes-hands-on/

Boston Dynamics: Atlas

#13 ON TRENDING

What's new, Atlas?

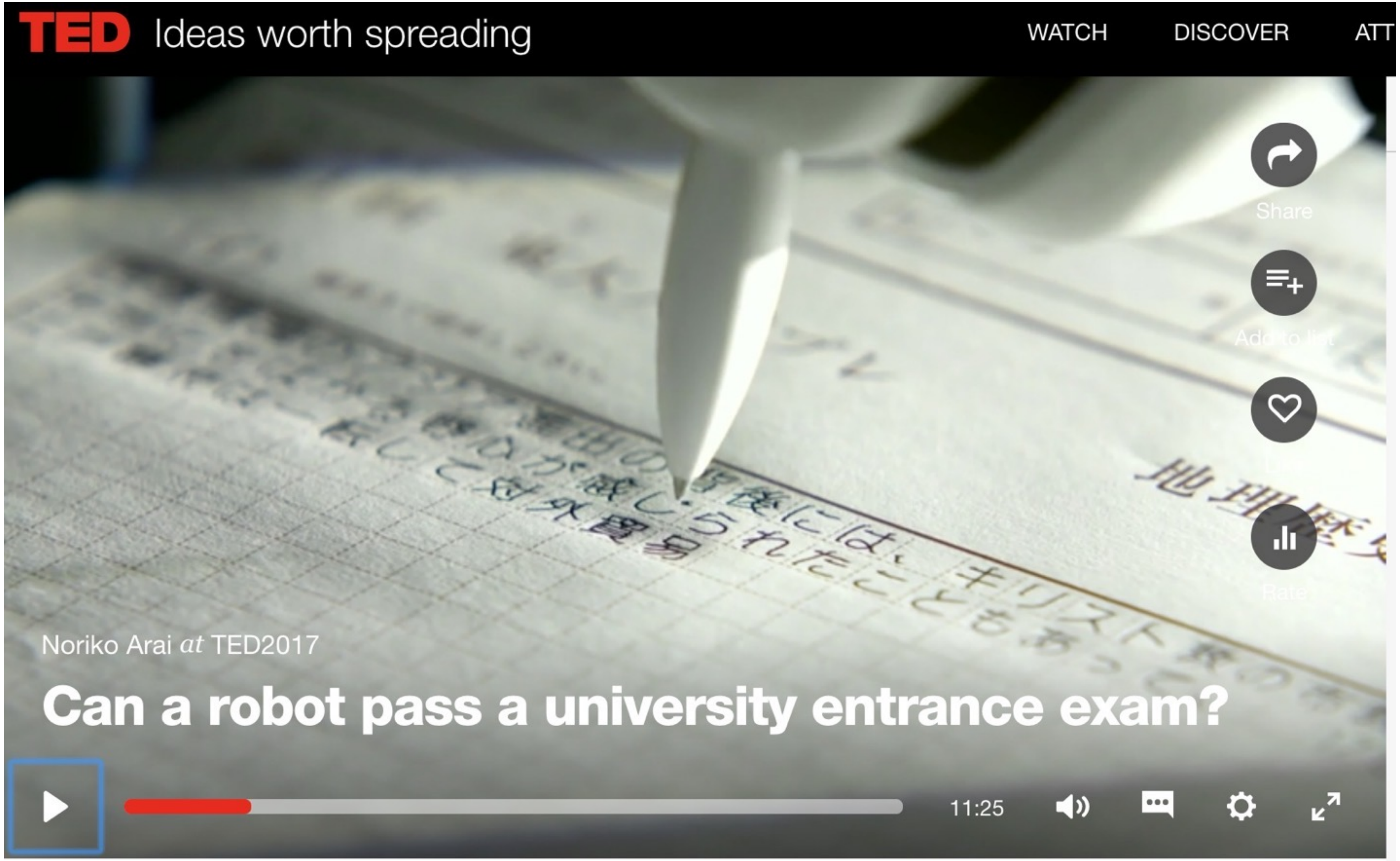
https://www.youtube.com/watch?v=fRj34o4hN4I

Humanoid Robot: Sophia

https://www.youtube.com/watch?v=S5t6K9iwcdw

Can a robot pass a university entrance exam?

Noriko Arai at TED2017



https://www.ted.com/talks/noriko_arai_can_a_robot_pass_a_university_entrance_exam

https://www.youtube.com/watch?v=XQZjkPyJ8KU

Robots

- Robots are **physical agents** that perform tasks by manipulating the physical world.
  - To do so, they are equipped with **effectors** such as **legs**, **wheels**, **joints**, and **grippers**.
- **Effectors** are designed to assert physical forces on the environment.

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Robots and Effectors

- When they do this, a few things may happen:
  - the robot’s state might change
  - the state of the environment might change
  - the state of the people around the robot might change

Robots

- The most common types of robots are manipulators (robot arms) and mobile robots.
- They have sensors for perceiving the world and actuators that produce motion, which then affects the world via effectors.

Robotics Problem

- The general robotics problem involves
  - stochasticity
    (which can be handled by MDPs)
  - partial observability
    (which can be handled by POMDPs)
  - acting with and around other agents
    (which can be handled with game theory)

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Robotic Perception

- We typically separate **perception (estimation)** from **action (motion generation)**.
- **Perception** in robotics involves **computer vision** to recognize the surroundings through cameras, but also localization and mapping.

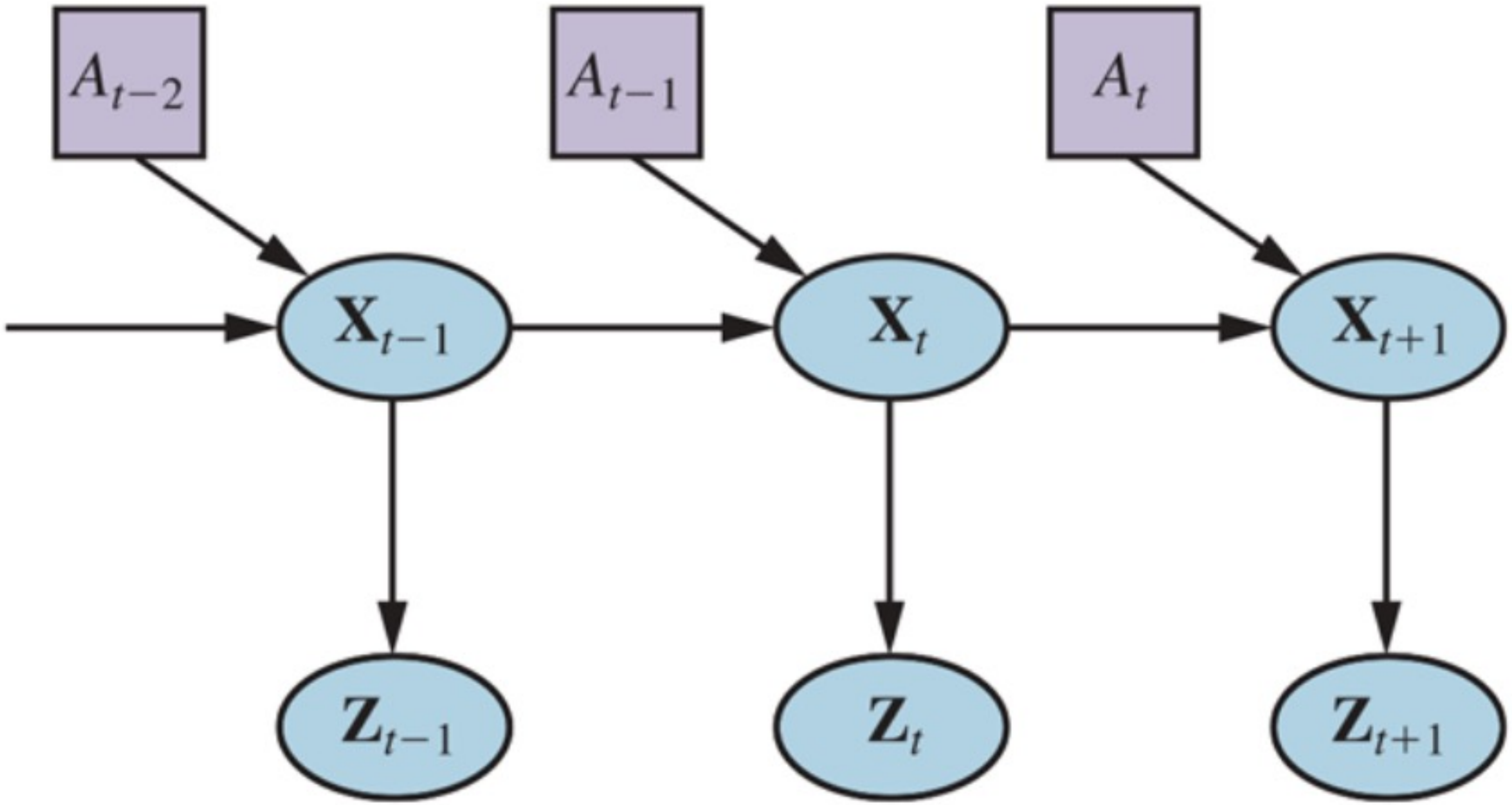
Robotic Perception

- Robotic perception concerns itself with estimating decision-relevant quantities from sensor data.
  - To do so, we need an internal representation and a method for updating this internal representation over time.

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Robot Perception

can be viewed as temporal inference from sequences of actions and measurements



Dynamic Decision network

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Probabilistic Filtering Algorithms

- **Probabilistic filtering algorithms** such as particle filters and Kalman filters are useful for robot perception.
  - These techniques maintain the belief state, a posterior distribution over state variables.

Configuration Spaces

- For generating motion, we use **configuration spaces**, where a point specifies everything we need to know to locate every **body point** on the robot.
  - For instance, for a robot arm with two joints, a configuration consists of the two joint angles.

Motion Generation

- We typically decouple the motion generation problem into
  - **motion planning**, concerned with producing a plan, and
  - **trajectory tracking control**, concerned with producing a policy for control inputs (actuator commands) that results in executing the plan.

Motion Planning

- **Motion planning can be solved via graph search**
  - **using cell decomposition**
  - **using randomized motion planning algorithms, which sample milestones in the continuous configuration space**
  - **using trajectory optimization, which can iteratively push a straight-line path out of collision by leveraging a signed distance field.**

Planning and Control

- **Optimal control** unites **motion planning** and **trajectory tracking** by computing an optimal trajectory directly over control inputs.

Planning Uncertain Movements

- **Planning under uncertainty** unites perception and action by
  - **online replanning** (such as model predictive control) and
  - **information gathering** actions that aid perception.

Reinforcement learning in robotics

- **Reinforcement learning is applied in robotics, with techniques striving to reduce the required number of interactions with the real world.**
- **Such techniques tend to exploit models, be it estimating models and using them to plan, or training policies that are robust with respect to different possible model parameters.**

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Humans and Robots

- Interaction with humans requires the ability to coordinate the robot’s actions with theirs, which can be formulated as a game.
- We usually decompose the solution into prediction, in which we use the person’s ongoing actions to estimate what they will do in the future, and action, in which we use the predictions to compute the optimal motion for the robot.

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Humans and Robots

- Helping humans also requires the ability to learn or infer what they want.
- Robots can approach this by learning the desired cost function they should optimize from human input, such as demonstrations, corrections, or instruction in natural language.
- Alternatively, robots can imitate human behavior, and use reinforcement learning to help tackle the challenge of generalization to new states.

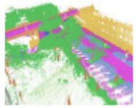
Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Papers with Code
State-of-the-Art (SOTA)

Robots

Motion Planning

130 papers with code



3D Semantic Segmentation

11 benchmarks

111 papers with code

Robot Navigation

5 benchmarks

84 papers with code

Visual Odometry

5 benchmarks

83 papers with code

Visual Navigation

5 benchmarks

72 papers with code

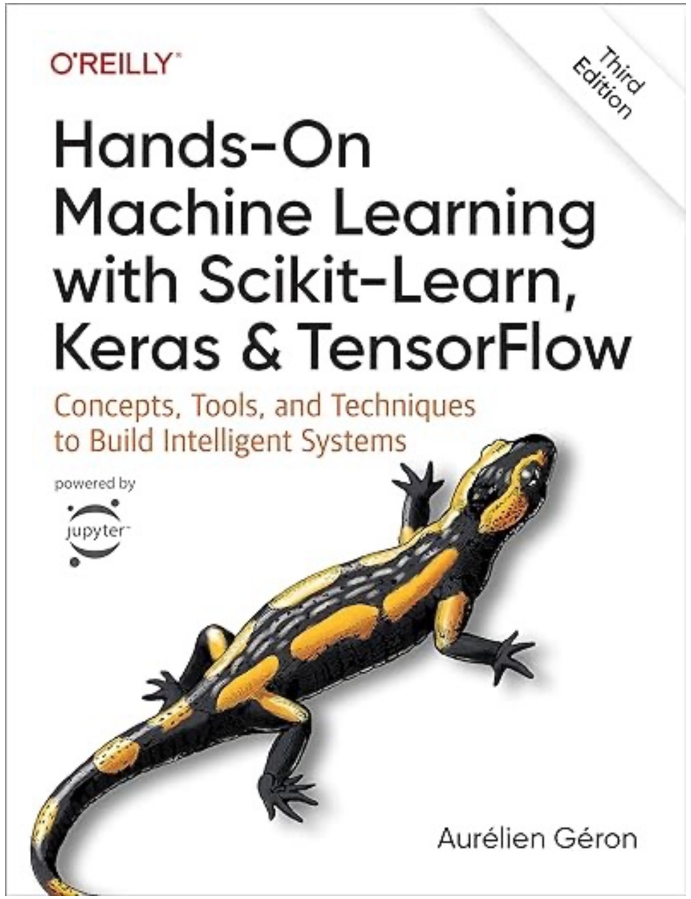
▸ **See all 54 tasks**

https://paperswithcode.com/sota

Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow

Notebooks

1. The Machine Learning landscape
2. End-to-end Machine Learning project
3. Classification
4. Training Models
5. Support Vector Machines
6. Decision Trees
7. Ensemble Learning and Random Forests
8. Dimensionality Reduction
9. Unsupervised Learning Techniques
10. Artificial Neural Nets with Keras
11. Training Deep Neural Networks
12. Custom Models and Training with TensorFlow
13. Loading and Preprocessing Data
14. **Deep Computer Vision Using Convolutional Neural Networks**
15. **Processing Sequences Using RNNs and CNNs**
16. Natural Language Processing with RNNs and Attention
17. **Autoencoders, GANs, and Diffusion Models**
18. Reinforcement Learning
19. Training and Deploying TensorFlow Models at Scale



https://github.com/ageron/handson-ml3

Summary

- Computer Vision
  - Classifying Images
  - Detecting Objects
  - The 3D World
- Robotics
  - Robotic Perception
  - Planning and Control
  - Planning Uncertain Movements
  - Reinforcement Learning in Robotics

References

- Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson.
- Denis Rothman (2024), Transformers for Natural Language Processing and Computer Vision - Third Edition: Explore Generative AI and Large Language Models with Hugging Face, ChatGPT, GPT-4V, and DALL-E 3, 3rd ed. Edition, Packt Publishing
- Aurélien Géron (2022), Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 3rd Edition, O'Reilly Media.
- Steven D'Ascoli (2022), Artificial Intelligence and Deep Learning with Python: Every Line of Code Explained For Readers New to AI and New to Python, Independently published.
- Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. (2022) "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors." arXiv preprint arXiv:2207.02696.
- Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. (2025) "Yolov9: Learning what you want to learn using programmable gradient information." In European Conference on Computer Vision, pp. 1-21. Springer, Cham.
- Nidhal Jegham, Chan Young Koh, Marwan Abdelatti, and Abdeltawab Hendawi. (2024) "Evaluating the Evolution of YOLO (You Only Look Once) Models: A Comprehensive Benchmark Study of YOLO11 and Its Predecessors." arXiv preprint arXiv:2411.00201.
- Ranjan Sapkota, and Manoj Karkee. (2024) "Yolo11 and vision transformers based 3d pose estimation of immature green fruits in commercial apple orchards for robotic thinning." arXiv preprint arXiv:2410.19846.
- Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. (2024) "Aligning cyber space with physical world: A comprehensive survey on embodied ai." arXiv preprint arXiv:2407.06886.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry et al. (2021) "Learning transferable visual models from natural language supervision." In International Conference on Machine Learning, pp. 8748-8763. PMLR.
- Wonjae Kim, Bokyung Son, and Ildoo Kim. (2021) "Vilt: Vision-and-language transformer without convolution or region supervision." In International Conference on Machine Learning, pp. 5583-5594. PMLR.
- Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R. Martin, Ming-Ming Cheng, and Shi-Min Hu. (2022) "Attention mechanisms in computer vision: A survey." Computational Visual Media ,:1-38.
- Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann.(2020) "Blazepose: On-device real-time body pose tracking." arXiv preprint arXiv:2006.10204.
- Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna et al.(2025) "Gemini robotics: Bringing ai into the physical world." arXiv preprint arXiv:2503.20020 (2025).
- Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay et al. (2025) "Cosmos world foundation model platform for physical ai." arXiv preprint arXiv:2501.03575 (2025).
- Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu et al. (2025) "Foundation models in robotics: Applications, challenges, and the future." The International Journal of Robotics Research 44, no. 5 (2025): 701-739.
- Min-Yuh Day (2025), Python 101, https://tinyurl.com/aintpupython101