Advanced Topics in Embodied Learning and Vision (Spring 2025)

**Instructor:** Mengye Ren ([mengye@nyu.edu](mailto:mengye@nyu.edu))

**Section Leaders:**

Chris Hoang ([ch3451@nyu.edu](mailto:ch3451@nyu.edu))

Ying Wang ([yw3076@nyu.edu](mailto:yw3076@nyu.edu))

**Recommended Prerequisites:** Machine Learning (DS-GA 1003 or CSCI-GA 2565), Computer Vision (CSCI-GA 2271), Deep Learning (DS-GA 1008), Proficiency in Python Programming

**Abstract:** This graduate level course covers advanced topics in embodied visual learning, perception and planning, with applications to robotics and self-driving. Topics include deep learning, computer vision, 3D perception, unsupervised 3D learning, self-supervised representation learning, continual learning, and LLM agents. Students are expected to know basic knowledge of machine learning, deep learning, and computer vision. Students are expected to be proficient in programming in Python. Advanced undergraduate students may enroll upon instructor approval.

**Course Website:** <https://embodied-learning-vision-course.github.io/>

**Course Google Calendar:** [Link](https://calendar.google.com/calendar/u/2?cid=NjFiOWFjOWJmYzNmMzdmYjc5YThmN2U1NzcyMzJlNDUyODE2YjhiMThkMTkwMTY4ODc4ZjNiNDg4ZGIwOTM0OEBncm91cC5jYWxlbmRhci5nb29nbGUuY29t)

**Table of Content**

[1 Goals](#_p09xyg1ta9ku)

[2 Grading](#_td47j36esggr)

[3 Lectures](#_bq6d6iz8tu1s)

[4 Tutorials](#_9e411b67vknn)

[5 Paper Reading List](#_ybbldhq7zv44)

[6 Paper Reviews (15%)](#_gpvn8sd94jm)

[7 Topic Presentation (30%)](#_qni6vwddofcr)

[7.1 Topic Presentation Schedule](#_qc4jlg9k3gxk)

[8 Course Project (45%)](#_e4j2rme4s7ga)

[8.1 Resources](#_9eygzyx7a1va)

[8.2 Presentation Schedule](#_mz81m31jxe4b)

[9 Miscellaneous](#_xri07tk808j0)

[9.1 Late Policy](#_m2txwrojzdb6)

[9.2 Collaboration Policy](#_ehq9xev54kca)

[9.3 Academic Integrity](#_6ch9n2lv8g7q)

[9.4 GenAI Policy](#_w4qvzlobkl8k)

[9.5 HPC Guide](#_6nltfncsyjbi)

[9.6 Related Educational Resources](#_kcv3ny39hktx)

# 1 Goals

* Understand and apply computer vision and deep learning in the context of embodied agent learning by applying spatial geometric priors, model inductive biases, design multi-task network architectures, and embodied foundation models.
* Learn to formulate a variety of computer vision and robotics problems using deep learning tools.
* Develop a deep understanding of supervised and self-supervised representation learning for downstream perception and planning tasks.
* Develop hands-on skill of implementing embodied learning systems in simulator environments.

# 2 Grading

* **In-Class Participation (10%):** Marks will be given on attendance, in-class participation and questions in regular classes, labs, student panel discussions, and guest lecturer presentations.
* **Paper Review (15%):** Students need to submit a one-page paper review on one of the selected readings every week.
* **Topic Presentation (30%):** Marks will be given on the prepared content, depth, analysis, presentation quality and panel discussion quality.
* **Course Project (45%):**
  + **Project Proposal (10%):** Project proposal is due in the 6th week. Student groups will need to schedule two mandatory consultations to discuss the project with the instructor and the TA. Failure to have the consultations will not get the 10% mark.
  + **Report (25%):** Project report is due in the final week.
  + **Presentation (10%):** Project presentations are conducted in the final two weeks.

# 3 Lectures

For student topic presentation see [here](#xrn2zr83cb2e).

1. [Week 1: Introduction](https://embodied-learning-vision-course.github.io/course-public/2025-spring/lectures/week01_intro.pdf)
2. [Week 2: Deep Learning for Structured Outputs](https://embodied-learning-vision-course.github.io/course-public/2025-spring/lectures/week02_structured_learning.pdf) [[Recording](https://nyu.box.com/s/ee9lm90zhzsvqqxv6307to50rmn9chf0)]
3. [Week 3: 3D Vision, Mapping](https://embodied-learning-vision-course.github.io/course-public/2025-spring/lectures/week03_3d_mapping.pdf) [[Recoding](https://nyu.box.com/s/vsla3ffhx4591sdwo9kjqcgxo0mp97a6)]
4. [Week 4: Self-Supervised Representation Learning and Object Discovery](https://embodied-learning-vision-course.github.io/course-public/2025-spring/lectures/week04_ssl.pdf) [[Recording](https://nyu.box.com/s/ff7bvi5nc8dpxjs69b4cgn1y3peg2hh6)]
5. [Week 5: World Models and Forecasting](https://embodied-learning-vision-course.github.io/course-public/2025-spring/lectures/week05_world_models.pdf) [[Recording](https://nyu.box.com/s/x0u40q9iey2e68ra3n5pgssa9kknkljq)]
6. [Week 6: End-to-End Planning and Control](https://embodied-learning-vision-course.github.io/course-public/2025-spring/lectures/week06_planning.pdf) [[Recording](https://nyu.box.com/s/h5i0ajgrac0aqkjp88tawewi68jb3xbe)]
7. [Week 7: Continual Learning and Few-Shot Learning](https://embodied-learning-vision-course.github.io/course-public/2025-spring/lectures/week07_continual_fewshot_learning.pdf) [[Recording (no sound)](https://nyu.box.com/s/xvjvyzb3x23bmcky7d981w9lbqv3tl4t)]
8. Week 8: Guest Lecture - Wei-Chiu Ma (Cornell University) + [Student Presentation](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week08_student_3d.pdf) [[Recording](https://nyu.box.com/s/rgqu30vf99ag4hs3cq5ibsfjnw38ugx1)]
9. [Week 9: Student Presentation](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week09_student_ssl_wm.pdf) [[Recording](https://nyu.box.com/s/ncmqzh33txe7yizgqdkbq3jhongd96mc)]
10. Week 10: Spring Break
11. [Week 11: Student Presentation](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week11_student_wm_e2e_planning.pdf) [[Recording](https://nyu.box.com/s/6udl36yyrmnhdeohew3wjp3ga9yortv0)]
12. Week 12: Guest Lecture - Andrei Bârsan (Waabi) + [Student Presentation](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week12_students_continual_learning.pdf) [[Recording](https://nyu.box.com/s/u701lnga46kqfwl6nh1w3qenttexr0h6)]
13. [Week 13: Student Presentation](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week13_students_few_shot_llm_agent.pdf) [[Recording](https://nyu.box.com/s/npwdz61akw59u3n3k4g4he8iqmgmxgh4)]
14. Week 14: Project Presentation I [[Recording](https://nyu.box.com/s/607hjx4wwiahh4moxxos7rd2aqhnoxi2)]
15. Week 15: Project Presentation II [[Recording](https://nyu.box.com/s/t15sanbkz76mwek25zvt0k52wcwolb2j)]

# 4 Tutorials

1. [Week 1: HPC](https://embodied-learning-vision-course.github.io/course-public/2025-spring/lab/lab1_hpc.pdf) [[notebook]](https://github.com/embodied-learning-vision-course/course-public/blob/main/2025-spring/lab/lab1_wandb_demo.ipynb)
2. [Week 2: Simulation Environments](https://embodied-learning-vision-course.github.io/course-public/2025-spring/lab/lab2_simulator.pdf) [[notebook](https://github.com/embodied-learning-vision-course/course-public/blob/main/2025-spring/lab/lab2-habitat/habitat-demo.ipynb)]
3. [Week 3: Video Learning](https://embodied-learning-vision-course.github.io/course-public/2025-spring/lab/lab3_video_learning.pdf) [[notebook](https://github.com/embodied-learning-vision-course/course-public/blob/main/2025-spring/lab/lab3-video/video-learning-demo.ipynb)]
4. [Week 4: Egocentric Videos](https://embodied-learning-vision-course.github.io/course-public/2025-spring/lab/lab4_ego4d.pdf) [[notebook](https://github.com/embodied-learning-vision-course/course-public/blob/main/2025-spring/lab/lab4_ego4d_viz.ipynb)]
5. [Week 5: Motion-Based Learning](https://embodied-learning-vision-course.github.io/course-public/2025-spring/lab/lab5_motion.pdf) [[notebook](https://github.com/embodied-learning-vision-course/course-public/blob/main/2025-spring/lab/lab5_motion.ipynb)]
6. [Week 6: LLM Agents](https://embodied-learning-vision-course.github.io/course-public/2025-spring/lab/lab6_LLMAgent.pdf)

# 5 Paper Reading List

**Week 1 Introduction**

* Turing (1950) [Computing Machinery and Intelligence](https://www.cs.ox.ac.uk/activities/ieg/e-library/sources/t_article.pdf)
* Pomerleau (1988) [ALVINN: An Autonomous Land Vehicle in a Neural Network](https://proceedings.neurips.cc/paper/1988/file/812b4ba287f5ee0bc9d43bbf5bbe87fb-Paper.pdf)
* [[Video] History Channel 1998 : Driverless Car Technology Overview at Carnegie Mellon University](https://www.youtube.com/watch?v=2KMAAmkz9go)
* Smith & Gasser (2005) [The Development of Embodied Cognition: Six Lessons from Babies](https://cogdev.sitehost.iu.edu/labwork/6_lessons.pdf)

**Week 2 Deep Learning for Structured Outputs**

Classic:

* LeCun (2006) [A Tutorial on Energy-Based Learning](https://www.cs.toronto.edu/~vnair/ciar/lecun1.pdf)
* Girshick et al. (2013) [Rich feature hierarchies for accurate object detection and semantic segmentation](https://arxiv.org/abs/1311.2524)
* Long et al. (2014) [Fully Convolutional Networks for Semantic Segmentation](https://arxiv.org/abs/1411.4038)
* Zheng et al. (2015) [Conditional Random Fields as Recurrent Neural Networks](https://arxiv.org/abs/1502.03240)
* Chen et al. (2016) [DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs](https://arxiv.org/abs/1606.00915)
* Kingma & Dhariwal (2018) [Glow: Generative Flow with Invertible 1x1 Convolutions](https://arxiv.org/abs/1807.03039)
* Ho et al. (2020) [Denoising Diffusion Probabilistic Models](https://arxiv.org/abs/2006.11239)

Recent:

* Carion et al. (2020) [End-to-End Object Detection with Transformers](https://arxiv.org/pdf/2005.12872)
* Kamath et al. (2021) [MDETR -- Modulated Detection for End-to-End Multi-Modal Understanding](https://arxiv.org/abs/2104.12763)
* Cheng et al. (2021) [Per-Pixel Classification is Not All You Need for Semantic Segmentation](https://arxiv.org/abs/2107.06278)
* Rombach et al. (2022) [High-Resolution Image Synthesis with Latent Diffusion Models](https://arxiv.org/abs/2112.10752)
* Kirillov et al. (2023) [Segment Anything](https://arxiv.org/abs/2304.02643)
* Bai et al. (2023) [Sequential Modeling Enables Scalable Learning for Large Vision Models](https://arxiv.org/abs/2312.00785)
* Chi et al. (2023) [Diffusion Policy: Visuomotor Policy Learning via Action Diffusion](https://arxiv.org/abs/2303.04137)

**Week 3 3D Vision, Mapping**

Classic:

* Moser et al. (2015) [Place Cells, Grid Cells, and Memory](https://pmc.ncbi.nlm.nih.gov/articles/PMC4315928/)
* Fischer et al. (2015) [FlowNet: Learning Optical Flow with Convolutional Networks](https://arxiv.org/abs/1504.06852)
* Godard et al. (2016) [Unsupervised Monocular Depth Estimation with Left-Right Consistency](https://arxiv.org/abs/1609.03677)
* Qi et al. (2016) [PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation](https://arxiv.org/abs/1612.00593)
* Tamar et al. (2016) [Value Iteration Networks](https://arxiv.org/abs/1602.02867)
* Parisotto et al. (2017) [Neural Map: Structured Memory for Deep Reinforcement Learning](https://arxiv.org/abs/1702.08360)
* Gupta et al. (2017) [Cognitive Mapping and Planning for Visual Navigation](https://arxiv.org/abs/1702.03920)

Recent:

* Chaplot et al. (2020) [Neural Topological SLAM for Visual Navigation](https://arxiv.org/abs/2005.12256)
* Huang et al. (2022) [FlowFormer: A Transformer Architecture for Optical Flow](https://arxiv.org/abs/2203.16194)
* Wu et al. (2023) [Policy Pre-training for Autonomous Driving via Self-supervised Geometric Modeling](https://arxiv.org/abs/2301.01006)
* Sun et al. (2023) [Dynamo-Depth: Fixing Unsupervised Depth Estimation for Dynamical Scenes](https://arxiv.org/abs/2310.18887)
* Wang et al. (2023) [DUSt3R: Geometric 3D Vision Made Easy](https://arxiv.org/abs/2312.14132)
* Kerbl et al. (2023) [3D Gaussian Splatting for Real-Time Radiance Field Rendering](https://arxiv.org/abs/2308.04079)
* Yang et al. (2024) [Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data](https://arxiv.org/abs/2401.10891)
* Wang et al. (2025) [Continuous 3D Perception Model with Persistent State](https://arxiv.org/abs/2501.12387)

**Week 4 Self-Supervised Representation Learning and Object Discovery**

Classic:

* Shi & Malik (2000) [Normalized Cuts and Image Segmentation](https://people.eecs.berkeley.edu/~malik/papers/SM-ncut.pdf)
* Sermanet et al. (2017) [Time-Contrastive Networks: Self-Supervised Learning from Video](https://arxiv.org/abs/1704.06888)
* Van den Oord et al. (2018) [Representation Learning with Contrastive Predictive Coding](https://arxiv.org/abs/1807.03748)
* Wu et al. (2018) [Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination](https://arxiv.org/abs/1805.01978)
* Chen et al. (2020) [A Simple Framework for Contrastive Learning of Visual Representations](https://arxiv.org/abs/2002.05709)
* Grill et al. (2020) [Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning](https://arxiv.org/abs/2006.07733)
* He et al. (2021) [Masked Autoencoders Are Scalable Vision Learners](https://arxiv.org/abs/2111.06377)

Recent:

* Weinzaepfel et al. (2022) [CroCo v2: Improved Cross-view Completion Pre-training for Stereo Matching and Optical Flow](https://arxiv.org/abs/2211.10408)
* Wang et al. (2022) [Self-Supervised Transformers for Unsupervised Object Discovery using Normalized Cut](https://arxiv.org/abs/2202.11539)
* Wang et al. (2023) [Cut and Learn for Unsupervised Object Detection and Instance Segmentation](https://arxiv.org/abs/2301.11320)
* Seo et al. (2022) [Masked World Models for Visual Control](https://arxiv.org/abs/2206.14244)
* Venkataramanan et al. (2023) [Is ImageNet worth 1 video? Learning strong image encoders from 1 long unlabelled video](https://arxiv.org/abs/2310.08584)
* van Steenkiste et al. (2024) [Moving Off-the-Grid: Scene-Grounded Video Representations](https://arxiv.org/abs/2411.05927)
* Cui et al. (2024) [DynaMo: In-Domain Dynamics Pretraining for Visuo-Motor Control](https://arxiv.org/abs/2409.12192)
* Wang et al. (2024) [PooDLe: Pooled and Dense Self-Supervised Learning from Naturalistic Videos](https://arxiv.org/abs/2408.11208)

**Week 5 World Models and Forecasting**

Classic:

* [Kalchbrenner et al. (2016)](https://arxiv.org/search/cs?searchtype=author&query=Kalchbrenner,+N) [Video Pixel Networks](https://arxiv.org/abs/1610.00527)
* Ha and [Schmidhuber](https://arxiv.org/search/cs?searchtype=author&query=Schmidhuber,+J) (2018) [World Models](https://arxiv.org/abs/1803.10122)
* Hafner et al. (2019) [Dream to Control: Learning Behaviors by Latent Imagination](https://arxiv.org/abs/1912.01603)

Recent:

* Liang et al. (2020) [Learning Lane Graph Representations for Motion Forecasting](https://arxiv.org/abs/2007.13732)
* Wu et al. (2022) [DayDreamer: World Models for Physical Robot Learning](https://arxiv.org/abs/2206.14176)
* Yu et al. (2022) [MAGVIT: Masked Generative Video Transformer](https://arxiv.org/abs/2212.05199)
* Hafner et al. (2023) [Mastering Diverse Domains through World Models](https://arxiv.org/abs/2301.04104)
* Hansen et al. (2023) [TD-MPC2: Scalable, Robust World Models for Continuous Control](https://arxiv.org/abs/2310.16828)
* Hu et al. (2023) [GAIA-1: A Generative World Model for Autonomous Driving](https://arxiv.org/abs/2309.17080)
* Zhang et al. (2024) [Copilot4D: Learning Unsupervised World Models for Autonomous Driving via Discrete Diffusion](https://arxiv.org/abs/2311.01017)
* Casas et al. (2024) [DeTra: A Unified Model for Object Detection and Trajectory Forecasting](https://arxiv.org/abs/2406.04426)
* Bruce et al. (2024) [Genie: Generative Interactive Environments](https://arxiv.org/abs/2402.15391)

**Week 6 End-to-End Planning and Control**

Classic:

* Ross et al. (2011) [A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning](https://arxiv.org/abs/1011.0686)
* Haarnoja et al. (2018) [Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor](https://arxiv.org/abs/1801.01290)
* Srinivas et al. (2018) [Universal Planning Networks](https://arxiv.org/abs/1804.00645)
* Sukhbaatar et al. (2018) [Intrinsic Motivation and Automatic Curricula via Asymmetric Self-Play](https://arxiv.org/abs/1703.05407)
* Amos et al. (2018) [Differentiable MPC for End-to-end Planning and Control](https://arxiv.org/abs/1810.13400)
* Zeng et al. (2019) [End-to-end Interpretable Neural Motion Planner](https://arxiv.org/abs/2101.06679)

Recent:

* Casas et al. (2021) [MP3: A Unified Model to Map, Perceive, Predict and Plan](https://arxiv.org/abs/2101.06806)
* Hu et al. (2022) [Planning-oriented Autonomous Driving](https://arxiv.org/abs/2212.10156)
* Chaplot et al. (2021) [Differentiable Spatial Planning using Transformers](https://arxiv.org/abs/2112.01010)
* Dinev et al. (2022) [Differentiable Optimal Control via Differential Dynamic Programming](https://arxiv.org/abs/2209.01117)
* Chi et al. (2023) [Diffusion Policy: Visuomotor Policy Learning via Action Diffusion](https://arxiv.org/abs/2303.04137)
* Psenka et al. (2024) [Learning a Diffusion Model Policy from Rewards via Q-Score Matching](https://arxiv.org/abs/2312.11752)

**Week 7 Continual Learning**

Classic:

* Marsland (2002) [A Self-Organising Network that Grows when Required](https://www.sciencedirect.com/science/article/pii/S0893608002000783)
* Kirkpatrick et al. (2016) [Overcoming catastrophic forgetting in neural networks](https://arxiv.org/abs/1612.00796)
* Rebuffi et al. (2016) [iCaRL: Incremental Classifier and Representation Learning](https://arxiv.org/abs/1611.07725)
* Yoon et al. (2017) [Lifelong Learning with Dynamically Expandable Networks](https://arxiv.org/abs/1708.01547)
* Nguyen et al. (2017) [Variational Continual Learning](https://arxiv.org/abs/1710.10628)
* Van de Ven et al. (2020) [Brain-Inspired Replay for Continual Learning with Artificial Neural Networks](https://www.nature.com/articles/s41467-020-17866-2)

Recent:

* Dohare et al. (2021) [Continual Backprop: Stochastic Gradient Descent with Persistent Randomness](https://arxiv.org/abs/2108.06325)
* Wang et al. (2021) [Learning to Prompt for Continual Learning](https://arxiv.org/abs/2112.08654)
* Powers et al. (2023) [Evaluating Continual Learning on a Home Robot](https://arxiv.org/abs/2306.02413)
* Zhang et al. (2023) [A Novel Visual Question Answering Continual Learning Setting](https://openaccess.thecvf.com/content/CVPR2023/papers/Zhang_VQACL_A_Novel_Visual_Question_Answering_Continual_Learning_Setting_CVPR_2023_paper.pdf)
* Lee et al. (2023) [STELLA: Continual Audio-Video Pre-training with Spatio-Temporal Localized Alignment](https://arxiv.org/abs/2310.08204)
* Majumder et al. (2023) [CLIN: A Continually Learning Language Agent for Rapid Task Adaptation and Generalization](https://arxiv.org/abs/2310.10134)

**Week 8 Few-Shot Learning, Meta-Learning**

Classic:

* Fei-Fei & Fergus (2006) [One-Shot Learning of Object Categories](http://vision.stanford.edu/documents/Fei-FeiFergusPerona2006.pdf)
* Lake et al. (2011) [One-Shot Learning of Simple Visual Concepts](https://cims.nyu.edu/~brenden/papers/LakeEtAl2011CogSci.pdf)
* Snell et al. (2017) [Prototypical Networks for Few-shot Learning](https://arxiv.org/abs/1703.05175)
* Finn et al. (2017) [Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks](https://arxiv.org/abs/1703.03400)
* James et al. (2018) [Task-Embedded Control Networks for Few-Shot Imitation Learning](https://arxiv.org/abs/1810.03237)
* Brown et al. (2020) [Language Models are Few-Shot Learners](https://arxiv.org/abs/2005.14165)
* Chen et al. (2021) [Exploring Simple Meta-Learning for Few-Shot Learning](https://openaccess.thecvf.com/content/ICCV2021/papers/Chen_Meta-Baseline_Exploring_Simple_Meta-Learning_for_Few-Shot_Learning_ICCV_2021_paper.pdf)

Recent:

* Javed & White (2019) [Meta-Learning Representations for Continual Learning](https://arxiv.org/abs/1905.12588)
* Lake (2019) [Compositional Generalization through Meta Sequence-to-Sequence Learning](https://arxiv.org/abs/1906.05381)
* Ren et al. (2021) [Wandering Within a World: Online Contextualized Few-Shot Learning](https://arxiv.org/abs/2007.04546)
* Alayrac et al. (2022) [Flamingo: a Visual Language Model for Few-Shot Learning](https://arxiv.org/abs/2204.14198)
* Song et al. (2022) [LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models](https://arxiv.org/abs/2212.04088)

**Week 9 LLM Agents**

Classic:

* Langley et al. (2009) [Cognitive architectures: Research issues and challenges](https://www.sciencedirect.com/science/article/abs/pii/S1389041708000557)
* Misra et al. (2017) [Mapping Instructions and Visual Observations to Actions with Reinforcement Learning](https://arxiv.org/abs/1704.08795)
* Andreson et al. (2018) [Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments](https://arxiv.org/abs/1711.07280)
* Andreas (2022) [Language Models as Agent Models](https://arxiv.org/abs/2212.01681)
* Sridhar et al. (2023) [Cognitive neuroscience perspective on memory: overview and summary](https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2023.1217093/full)

Recent:

* Anh et al. (2022) [Do As I Can, Not As I Say: Grounding Language in Robotic Affordances](https://arxiv.org/abs/2204.01691)
* Sumers et al. (2023) [Cognitive Architectures for Language Agents](https://arxiv.org/abs/2309.02427)
* Schick et al. (2023) [Language Models Can Teach Themselves to Use Tools](https://arxiv.org/pdf/2302.04761)
* Rana et al. (2023) [SayPlan: Grounding Large Language Models using 3D Scene Graphs for Scalable Robot Task Planning](https://arxiv.org/abs/2307.06135)
* Kim et al. (2024) [ReALFRED: An Embodied Instruction Following Benchmark in Photo-Realistic Environments](https://arxiv.org/abs/2407.18550)
* Li et al. (2024) [Embodied Agent Interface: Benchmarking LLMs for Embodied Decision Making](https://arxiv.org/abs/2410.07166)

# 6 Paper Reviews (15%)

* You need to submit a paper review every week from Week 2 - Week 9 on Gradescope.
* You may choose a recent paper (<3 years) on the designated topic. You may choose one paper from the suggested reading list. If you choose your own paper, you should choose from reputable venues such as NeurIPS, ICML, ICLR, AISTATS, CoLLAs for ML papers, CVPR, ICCV, ECCV for CV papers, and ICRA, RSS, CoRL, IROS for robotics papers. You should choose papers that have a potentially high impact.
* Paper topics:
  + Week 2: Deep Learning for Structured Outputs
  + Week 3: 3D Vision and Mapping
  + Week 4: SSL and Object Discovery
  + Week 5: World Models and Forecasting
  + Week 6: End-to-End Planning
  + Week 7: Continual Learning
  + Week 8: Few-Shot Learning and Meta-Learning
  + Week 9: LLM Agents
* Template and rubric is [here](https://docs.google.com/document/d/1dx58WGbS9zA4nn46byiuEfsKRCT-mUQqUcJ4yWCpWGA/edit?usp=sharing). You have to use the template.

# 7 Topic Presentation (30%)

* **Time:** 20 minutes
* **Sign Up: Students have to sign up for a slot by Week 3 (Feb 6).** Sign up link is [here](https://forms.gle/mfHDYn5n4zb6jBmk9). Topic is not guaranteed and will be assigned on a first-come first-serve basis.
* **Presentation Slides:** Please email me your slides the day before the presentation so that all the slides can be assembled together. You have to email me the slides after your presentation in order to receive the mark.
* **Calendar:** **Student topic presentation starts from Week 7 and lasts until Week 13.** Approximately 3-4 students present on each topic. Each student will conduct a 30-minute presentation on 1-2 designated recent papers including necessary backgrounds.
* **Panel Discussion:** After the presentation, the presenter students will form an expert panel, and lead a 30-minute discussion Q&A with the audience. The audience students are required to ask questions related to the general topic, and the presenter students will give their opinions. Participation marks will be awarded based on high quality audience questions. The panel performance will also be part of the presentation score.
* **Marking rubric** is [here](https://docs.google.com/document/d/1D0TeDU2XmJNPVRJ5FrgIBgvHXNdG_Zt3cml3vG1Umac/edit?usp=sharing).

## 7.1 Topic Presentation Schedule

| **Date** | **Topic** | **Presenters** |
| --- | --- | --- |
| Mar 6 (Week 7) | Deep Learning for Structured Outputs | [Tanishq Sardana: Segment Anything](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week07_Tanishq_Sardana.pdf)  [Qing Mu: DETR](http://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week07_Qing_Mu.pdf)  [Owais Shuja: Latent Diffusion Models](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week07_Owais_Saad_Shuja.pdf) |
| Mar 13 (Week 8) | 3D Vision and Mapping | [Sihang Li: Scene Coordinate Reconstruction](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week08_Sihang_Li.pdf)  [Kanishkha Jaisankar: NeRF](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week08_Kanishkha_Jaisankar.pdf)  [Denis Mbey Akola: DUSt3R](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week08_Denis_Mbey_Akola.pdf)  [Zijin Hu: Zero-1-to-3](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week08_Zijin_Hu.pdf) |
| Mar 20 (Week 9) | SSL and Object Discovery | [Dahye Kim: DINOv2](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week09_Dahye_Kim.pdf)  [Surbhi: IJEPA](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week09_Surbhi.pdf)  [Sal Yeung: Predictable and Robust Neural Representations by Straightening](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week09_Sal_Yeung.pdf)  [Anurup Naskar: Moving Off-the-Grid](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week09_Anurup_Naskar.pdf) |
| Mar 20 (Week 9) | World Models and Forecasting | [Andrew Deur: DayDreamer](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week09_Andrew_Deur.pdf)  [Sidhartha Reddy Potu : UniSim](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week09_Sidhartha_Reddy_Potu.pdf) |
| Apr 3 (Week 11) | World Models and Forecasting | [Sergey Sedov: DreamerV2 and](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week11_Sergey_Sedov.pdf)  [Backpropagation-based Policy](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week11_Sergey_Sedov.pdf)  [Gradients](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week11_Sergey_Sedov.pdf)  [Pratyaksh Prabhav Rao: DINO-WM](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week11_Pratyaksh_Prabhav_Rao.pdf)  [Rooholla Khorrambakht: Diffusion for World Modeling](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week11_Rooholla_Khorrambakht.pdf) |
| Apr 3 (Week 11) | End-to-End Planning | [Mrunal Sarvaiya: Differential MPC](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week11_Mrunal_Sarvaiya.pdf)  [Sushma Mareddy: MP3](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week11_Sushma_Mareddy.pdf)  [Raman Kumar Jha: UniAD](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week11_Raman_Jha.pdf)  [Jovita Gandhi: Embodied GPT](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week11_Jovita_Gandhi.pdf) |
| Apr 10 (Week 12) | Continual Learning | [Akshay Raman: DualNet](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week12_Akshay_Raman.pdf)  [Amey Joshi: LifelongRL](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week12_Amey_Joshi.pdf)  [Zifan Zhao: Loss of Plasticity](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week12_Zifan_Zhao.pdf) |
| Apr 17 (Week 13) | Few-Shot Learning and Meta-Learning | [Ellen Su: Seeing the Un-Scene](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week13_Ellen_Su.pdf)  [Xu Zhang: FSL + Diffusion](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week13_Xu_Zhang.pdf) |
| Apr 17 (Week 13) | LLM Agents | [Ravan Buddha: Gemini Robotics](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week13_Ravan_Buddha.pdf)  [Dan Zhao: Magma](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week13_Dan_Zhao.pdf)  [Solim LeGris: CoALA](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week13_Solim_LeGris.pdf)  [Sunidhi Tandel: LEO](https://embodied-learning-vision-course.github.io/course-public/2025-spring/seminars/week13_Sunidhi_Tandel.pdf) |

# 8 Course Project (45%)

The goal of the final project is to let students develop hands-on skills of implementing embodied learning systems for concrete real-world tasks such as toy games, long-form egocentric video understanding, self-driving simulation, indoor navigation, and robotic manipulation.

**Suggested Topics:**

Students are expected to follow the suggested directions below:

* **End-to-end self-supervised learning for perception and planning**

Existing end-to-end learning-based planning frameworks are mostly focused on supervised learning of labeled objects and human demonstrations. However, real-world agents need to learn to interact with the world with much less direct supervision. There are many ways towards realizing such a self-supervised system, ranging from pure reinforcement learning to more structured object-centered learning of passing discrete information. Students are encouraged to explore different design spaces and find a suitable toy environment to demonstrate their claims.

* **Enhancing foundation models for spatial intelligence**

LLMs are typically trained with discrete tokens and are less familiar with the 3D world to perform exact perception, inference and planning. Students are encouraged to explore ways to augment pretrained foundation models with the ability to perceive and plan under precision in embodied environments, and investigate whether spatial oriented learning objectives in post-training can enhance such ability.

* **Continual learning for embodied intelligence**

Machine learning models are typically trained with fixed data and are unable to adapt in deployment. In-context learning in foundation models addresses some of the adaptability requirements but is limited by its context length and can often get distracted from multiple sources of information. Here, students are encouraged to explore various ways for continuous adaptation in deployment. Potential approaches include memory design, retrieval augmentation, continuous finetuning, and incremental learning with experience/action abstraction.

**Project Consultation:**

To guide the project throughout the semester, the students are required to complete project consultations. They need to meet with the instructor and the TAs at least once each. See key dates below on when to meet with the instruction team the latest in the semester.

**Project Presentation Sign Up:**

You can choose to present on Week 14 or Week 15. **The groups who present on Week 14 will get a bonus of 2% added to their final grade**. You can sign up by filling out the form [here](https://forms.gle/4pCMUh7rBcxwNnVU8). The deadline for signing up for the presentation is Apr 10 (Week 12).

**Presentation:**

Time duration is 20 minutes + 2 min Q&A. You will be graded in terms of Project Fit (2%), Clarity (2%), Results Significance (2%), Slides Quality (2%), Q&A (2%).

**Final Report & Proposal:**

The LaTeX template is here ([pdf file](https://embodied-learning-vision-course.github.io/course-public/2025-spring/project/template.pdf)) ([zip TeX file](https://embodied-learning-vision-course.github.io/course-public/2025-spring/project/template.zip)). **You have to use the template.** You can use [Overleaf](https://www.overleaf.com) to perform LaTeX processing.

**Key Dates (unless announced otherwise, every item is due before the lecture time at 4:55pm):**

* Feb 6: Team registration of two students. Form link is [here](https://forms.gle/wZssdm6rs1Bfg5vB7). If you cannot find a partner, you still need to submit the form.
* Feb 20: Complete one consultation meeting during the office hour with the instructor and/or the TAs.
* Feb 27: Course project proposal due. Submit the proposal on Gradescope. Each group only needs to submit one copy.
* Mar 20: Complete second consultation meeting during the office hour with the instructor and/or the TAs. You need to meet with the instructor and the TA at least once each. You are expected to bring preliminary results during the second meeting.
* Apr 10: Sign up for a presentation slot [here](https://forms.gle/4pCMUh7rBcxwNnVU8).
* Apr 24/May 1: Presentation slides due the day before the presentation. Please submit your presentation slide deck [here](https://forms.gle/Anzrb7k67ERQeFxF7).
* ~~May 2~~ May 8: Course final report due. Submit the report on Gradescope. Each group only needs to submit one copy.
* ~~May 2~~ May 8: Finish mandatory peer evaluation survey. Form link is [here](https://forms.gle/Tr9r5zjTAW6B2E236).

## 8.1 Resources

**Embodied Learning Environments**

| **Name** | **Description** | **Details** |
| --- | --- | --- |
| [**Walking Tours**](https://huggingface.co/datasets/shawshankvkt/Walking_Tours) | long-duration egocentric videos captured in cities in Europe and Asia | pseudo-ground truth semantic segmentation labels (stored on Greene) |
| [**BDD100K**](https://github.com/bdd100k/bdd100k/blob/master/doc/source/download.rst) | dashcam driving videos in US cities | labels for object detection, semantic segmentation, lane markings, drivable areas, GPS/IMU |
| [**CARLA**](https://carla.org/) | autonomous driving simulator powered by Unreal Engine 5 | simulate driving, assets (maps, buildings, vehicles, etc.), traffic, weather, sensors |
| [**SMARTS**](https://github.com/huawei-noah/SMARTS) | multi-agent RL for self-driving | simulated driving world with objects |
| [**nuPlan**](https://nuplan.org/) | self-driving simulation planning benchmark | open and closed loop planning with sensor simulation |
| [**AI2-THOR**](https://ai2thor.allenai.org/) | indoor home simulation | [ALFRED challenge](https://askforalfred.com/EAI21/)  [visual semantic navigation challenge](https://ai2thor.allenai.org/robothor/cvpr-2021-challenge/)  [rearrangement challenge](https://ai2thor.allenai.org/rearrangement/) |
| [**Something-Something v2**](https://www.qualcomm.com/developer/software/something-something-v-2-dataset) | videos of humans performing basic actions with objects | action-object labels |
| [**Habitat (v1, v2, v3)**](https://aihabitat.org/) | photorealistic 3D simulator of home environments | simulate embodied agent, map layout, objects, sensors |

**Egocentric Video Datasets**

| [**Ego4d**](https://ego4d-data.org/docs/challenge/) | A large-scale egocentric video dataset on daily life activities |
| --- | --- |
| [**Ego-Exo**](https://ego-exo4d-data.org/#benchmarks) | A diverse, large-scale multi-modal, multi-view, video dataset recording skilled human activities. |
| [**EPIC-KITCHENS**](https://epic-kitchens.github.io/2025) | A large-scale egocentric video dataset on daily activities in the kitchen |
| [**Project Aria**](https://www.projectaria.com/datasets/) | Multiple real-world/sythetic egocentric video datasets, including 3D annotations |
| [**HoloAssist**](https://holoassist.github.io/) | An egocentric human interaction video dataset, where two people collaboratively complete physical manipulation tasks |

## 8.2 Presentation Schedule

| **Date** | **Group** | **Presenter** | **Topic** |
| --- | --- | --- | --- |
| Apr 24 | 1 | Pratyaksh Prabhav Rao  Mrunal Sarvaiya | Spatial Memory Augmented Reinforcement Learning |
| Apr 24 | 4 | Raman Kumar Jha  Amey Joshi | Diffusion Model Predictive Control |
| Apr 24 | 6 | Sal Yeung  Rooholla Khorrambakht  Surbhi | Exploring Data-Efficient World Modeling and Representation Learning Based On Equivariant Architectures and Foundational 3D Models |
| Apr 24 | 7 | Akshay Raman  Owais Saad Shuja | Continual Reinforcement Learning for Autonomous Robotic Tasks |
| Apr 24 | 9 | Tanishq Sardana  Sergey Sedov | Enhancing Information Retrieval of World Models by Augmenting Latents |
| Apr 24 | 11 | Kanishkha Jaisankar  Sunidhi Tandel | LLM Guided Motion Planning: Instruction-Tuned Models for Human-Aligned Autonomous Driving |
| Apr 24 | 13 | Andrew Deur  Jovita Gandhi | Finetuning Robotic MLLMs to Enhance Language Perception Capabilities |
| May 1 | 2 | Sidhartha Reddy Potu  Denis Mbey Akola | Object Tracking in Egocentric Videos |
| May 1 | 3 | Zifan Zhao  Zijin Hu | Enhancing Visual-Motor Policies with Surface Normal Estimation |
| May 1 | 5 | Anurup Naskar  Kim Da Hye | GAM: Graph-Augmented Memory for Egocentric Video Understanding |
| May 1 | 8 | Qing Mu  Dan Zhao | Adapting World and Human Action Models for Spatial Navigation |
| May 1 | 10 | Sushma Mareddy  Ravan Buddha | Self-Supervised End-to-End RL for Autonomous Driving in Simulation |
| May 1 | 12 | Swarali Borde  Xu Zhang  Sihang Li | Exo-Ego Transfer with Foundation Model on Object-Centric Videos |
| May 1 | 14 | Ellen Su  Solim LeGris | BabyGenie |

# 9 Miscellaneous

## 9.1 Late Policy

You have 4 late days in total for any deliverables. If your deliverable is a team effort then it would cost late days from each team member. You can use a maximum of 2 late days on each deliverable. If it is beyond 2 late days or you have used up your late days, then there will be a 20% penalty on the marks each day.

## 9.2 Collaboration Policy

You may not collaborate unless it is a team activity. You may talk to other students on topic selections for presentations and projects.

## 9.3 Academic Integrity

Work you submit should be your own. Please consult the [CAS academic integrity policy](http://cas.nyu.edu/page/academicintegrity) for more information. Penalties for violations of academic integrity may include failure of the course, suspension from the University, or even expulsion.

## 9.4 GenAI Policy

AI may not be used in weekly paper reviews and paper presentations. AI may be used towards coding assistance and report writing assistance in the course project. However, the use of AI can still impact the grade if the report contains poor writings and non-factual statements.

## 9.5 HPC Guide

All students are encouraged to use the NYU HPC Cloud Burst to run their experiments for the final project. The usage guide can be found [here](https://docs.google.com/document/d/13Y32fYrHLn0GNItla4YGQdi4sRu19gn1zdPCuGCwL2Y/edit?usp=sharing).

## 9.6 Related Educational Resources

* [Waabi: CVPR 2024 Self-Driving Tutorial](https://www.youtube.com/playlist?list=PLb2zgvIwtM-5JxKqS9_3CpZX2_7Uj8jG6)
* [Andreas Geiger: Autonomous Driving Course](https://uni-tuebingen.de/fakultaeten/mathematisch-naturwissenschaftliche-fakultaet/fachbereiche/informatik/lehrstuehle/autonomous-vision/lectures/self-driving-cars/)
* [Jitendra Malik: Robot That Learns](https://www.youtube.com/playlist?list=PLPaC96j0xdLcYLTSoSk9PO1Yg-1udJd-S) [Website](https://robots-that-learn.github.io/)
* [David Silver: Reinforcement Learning Course](https://www.youtube.com/watch?v=2pWv7GOvuf0&list=PLzuuYNsE1EZAXYR4FJ75jcJseBmo4KQ9-)
* [Sergey Levine: Deep Reinforcement Learning Course](https://www.youtube.com/playlist?list=PL_iWQOsE6TfX7MaC6C3HcdOf1g337dlC9) [Website](https://rail.eecs.berkeley.edu/deeprlcourse/)