AI Seminar Report

Date: 03/21/2025

Name	Vo Hoang Chuong (우황장)	**Student No.**	2024121193
School	Dongguk University	**Course Classification**	Ph.D.
Speaker Name	최종현 부교수	**Speaker Affiliation**	서울대학교
Title	Action sequence prediction by vision and language		

■ Seminar Contents

I. Foundations:

- Sequential Visual Data: Modeling sequences of visual data (video, images, sketches) + language.
- Beyond Human Vision: Neuromorphic sensors, thermal-to-photo face recognition.
- Embodied Action Heuristics: Vision + Language impact on agent actions.

II. Embodied AI: Perception to Action:

- Developmental Origins: Embodied AI principles, analogy to adolescent learning; necessity debate.
- Perception-Action Loops: Vision + language -> action sequences for tasks.
- Enabling Technologies:
  - Subtasks: Interactive QA, vision-language/visual navigation.
  - House3D: Simulator for domestic environment interactions.
  - ALFRED Benchmark: Evaluating agents on household tasks (language, vision, directives).

III. Learning and Architecture:

- Learning Paradigms:
  - Imitation Learning: Learning from expert trajectories; sample/compute efficient.
  - Reinforcement Learning: Learning via reward; high sample/compute requirements.
- Hierarchical Instructions: Inferring low-level actions from high-level instructions (2020-2022).
- Vision-Language Grounding: Architectures (LSTMs, pixel-wise interaction masks).
- Failure Cases: e.g., Optical flow in multi-task learning; information disconnects.
- Transfer Learning: Taskonomy; task suitability for knowledge transfer.
- Qualitative Analysis: Language attention; surrounding views; modularization.
- Efficient Architectures: Binary Networks.

- Continual Model Updates: Continual learning.

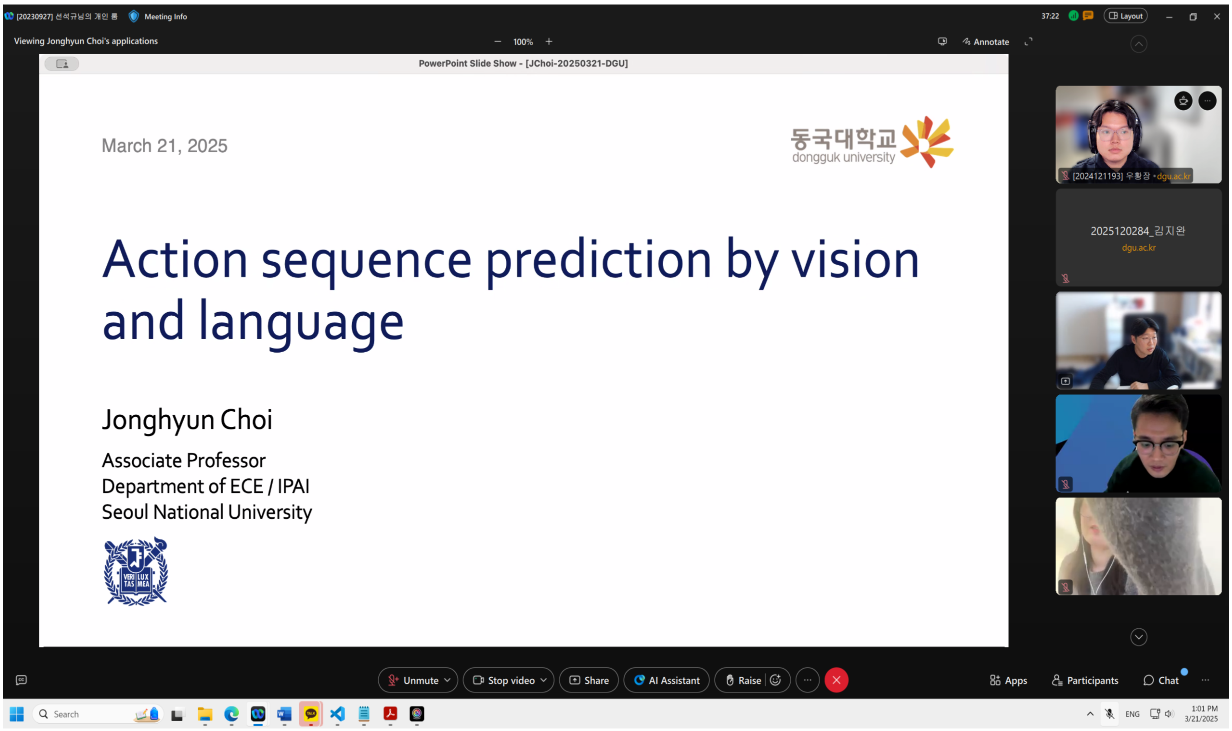
IV. Challenges and Future:

- Memory Limitations: Forgetting steps/object locations; solutions.
- Domain Gap:
  - LLM Actions: Reducing the gap between LLM-inferred actions and embodiment.
  - Virtual-Real Transfer: ReALFRED dataset.
- Continual Learning: CL-ALFRED; adaptation without forgetting.

■ What have you learned from this seminar?

- AI agents need sequential visual data combined with language for better learning.
- Embodied AI is an emerging trend for training action sequences. This approach is inspired by human nature's way of learning during adolescence through explorative physical interactions.
- Exploration learning is still debated whether it is necessary for future AI systems.
- There are benchmarks and simulations for the Embodied AI field, like House3D and ALFRED.
- There is a tradeoff between imitation learning (efficient but expert-dependent) and reinforcement learning (flexible but needs more data and computation).
- There are multiple failure cases and limitations in existing models.
- **Critical Challenges:**
- Addressing memory limitations in agents (forgetting, object relocation).
  - Bridging the "domain gap" between:
  - LLM-generated action plans and embodied execution.
  - Virtual training environments and real-world deployment.
- **Future Directions:**
  - Continual learning (e.g., CL-ALFRED) is a key to creating adaptable, lifelong learning agents.
  - There is a critical need for datasets (like ReALFRED) to facilitate real-world transfer.
  - Improving qualitative analysis methods.
  - Looking at components like attention and modularity.

Capture your Webex screen showing the start time (1:00 PM).



PowerPoint Slide Show - [JChoi-20250321-DGU]

March 21, 2025

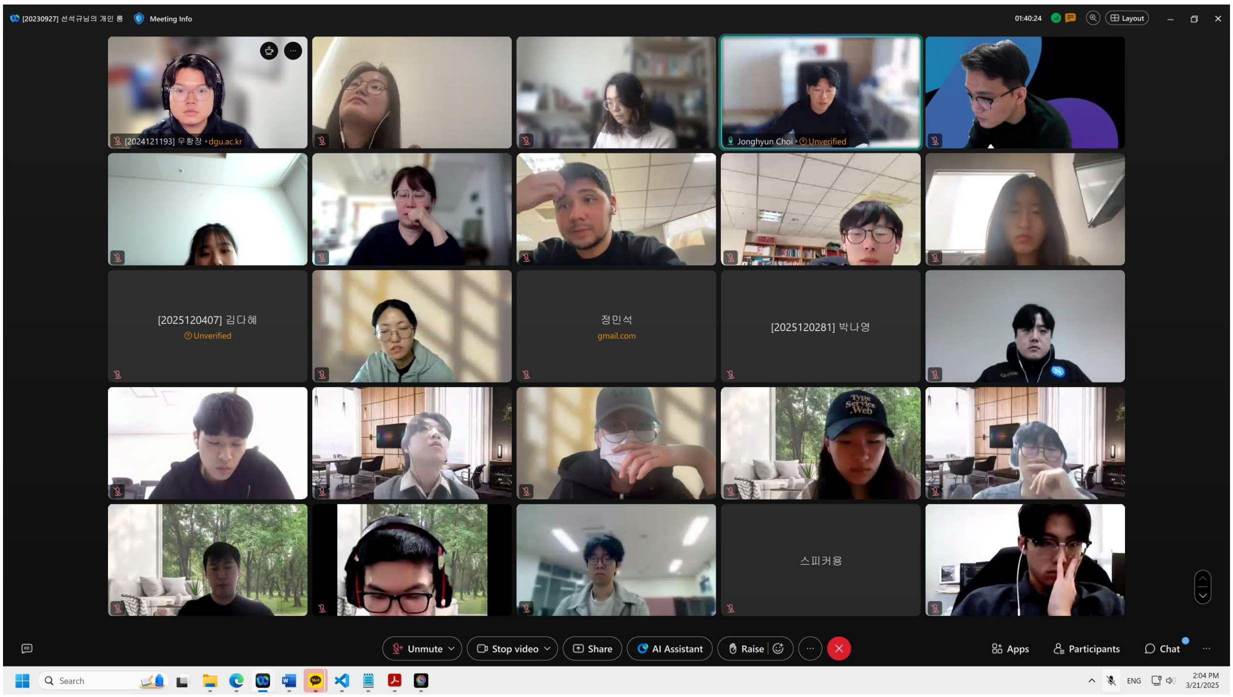
동국대학교
dongguk university

Action sequence prediction by vision and language

Jonghyun Choi

Associate Professor
Department of ECE / IPAI
Seoul National University

Capture your Webex screen showing the end time as well.



[20230927] 선석규님의 개인 룸
Meeting Info
01:40:24
Layout
[2024121193] 우황창 • dgu.ac.kr
Jonghyun Choi • Unverified
[2025120407] 김다혜
Unverified
정민석
gmail.com
[2025120281] 박나영
스피커용
Unmute
Stop video
Share
AI Assistant
Raise
Apps
Participants
Chat
Search
ENG
2:04 PM
3/21/2025