AI773 (4190.773): Topics in Artificial Intelligence

Foundational Models and Physical World Models for Multimodal AI

Sangdoo Yun[1,2] Jin-Hwa Kim[1,2]
[1]NAVER AI Lab [2]Seoul National University

Fall 2026

Overview

This course is designed for graduate students who want to understand and contribute to the two frontiers that are rapidly converging in modern AI: *multimodal foundation models* that perceive, reason, and generate across vision, language, and audio, and *physical world models* that represent, simulate, and predict the 3D/4D physical world under actions. The first seven weeks of the semester are devoted to intensive lectures, given as two series: one on multimodal foundation models, covering multimodal representation and multimodal foundation models, and one on 3D/4D representation and generation together with a review of the recent world-model literature. The rest of the semester is a paper-driven seminar in which students present and critique papers selected from a topic pool proposed by the instructors, followed by an open discussion in which the whole class brainstorms possible future research directions. Prior knowledge of machine learning and deep learning is assumed.

1 Target Audience

- Graduate students with prior knowledge in machine learning and deep learning who wish to specialize in multimodal foundation models, 3D/4D generation, or world models.
- Familiarity with Python and PyTorch, and comfort with reading recent conference papers (e.g., CVPR, ICCV, NeurIPS, ICLR, ICML, etc.), is strongly recommended.

2 Course Objectives

- Understand the design principles of multimodal foundation models: architectures, pre-/post-training recipes, data, and evaluation.
- Understand the theory and practice of visual generative modeling (diffusion, flow matching, autoregressive/unified models) and its extension to 3D and video.
- Acquire the geometric background required for neural graphics (multi-view geometry, volumetric rendering, NeRF, 3D Gaussian Splatting) and its use in 3D/4D generation.
- Critically review the emerging literature on physical world models: action-conditioned video prediction, interactive and playable world models, and embodied applications. Particular attention is paid to what a world model must get right to be useful in the real world — physical and geometric fidelity, long-horizon consistency, partial observability, action controllability, inference latency, and safety — in domains such as robotic manipulation, navigation, and autonomous driving.
- Develop the ability to present a paper rigorously and to formulate concrete, novel follow-up research questions through structured discussion.

3 General Information

3.1 Class information

- Friday 1:00 PM - 3:50 PM, Sep. 4th to Dec. 18th, 2026
- Bldg. 302, Room 107 (TBC)
- Credit 3-3-0
- Enrollment cap: 24 students (3 presentation sets per session) $\sim$ 32 students (4 sets per session)
- Course website: `https://naver-ai.github.io/202602-AI773/`
- The class material is in English, while the instruction is in Korean.

3.2 Course staff and contact

- Instructors: Sangdoo Yun (sangdoo.yun@navercorp.com), Jin-Hwa Kim (j1nhwa.kim@navercorp.com)
- Teaching assistants: Juno Kim (jokim@snu.ac.kr)

3.3 Office hour

- Friday 10:00 AM, AIIS Bldg. 942, Room 208
- Reservation only

3.4 Prior offerings

- Fall 2025, Multimodal Generative AI Theories and Applications (`https://naver-ai.github.io/202502-AI773/`)
- Fall 2024, Multimodal Deep Learning Theories and Applications (`https://naver-ai.github.io/202402-AI773/`)
- Fall 2023, Multimodal Deep Learning Theories and Applications (`https://naver-ai.github.io/202302-AI773/`)
- Students without a background in multimodal learning are encouraged to skim the Fall 2025 lecture slides on multimodal representation learning, multimodal foundation models, and neural graphics before the semester begins.

4 Course Format

The 15 sessions consist of one introductory session, six lectures, and eight sessions of student presentation and discussion.

- **Introduction (1 session).** Course overview and the policy for the presentation and discussion.
- **Lectures (6 sessions).** Two lecture series of three sessions each.
  - **(3 sessions) Multimodal Foundation Models** (Sangdoo Yun): architectures and training of multimodal representations, multimodal foundation models, and a review of the recent world model literature.

  - **(3 sessions) Multimodal Generation** (Jin-Hwa Kim): foundations of 3D representation and rendering, 3D/4D generation, and a review of the recent physical AI literature.

- **Student presentations and discussion (8 sessions).** Each session runs 3–4 sets, and a set consists of a 15–25 minute presentation followed by a 15-minute discussion. Three sets fill roughly 120 minutes and four sets roughly 160 minutes of the 170-minute slot, with a short break between sets.

4.1 Presentation

The student presentations are the core of this course, not a supplement to the lectures. Each student presents once, and the presentation is expected to reach the standard of a conference-style talk rather than a summary of the abstract.

- Each student presents one paper, or a tightly coupled pair of papers, selected from the topic pool announced in Week 1. Topics are assigned by preference ranking; ties are broken on a first-come, first-served basis.
- The presentation should cover: the problem and why it matters, the background needed to follow the method, the technical core of the method, the experimental evidence and what it does not establish, and the presenter's own critique.
- Slides are submitted to the instructors no later than 2 days before the presentation date. A revised version, incorporating what came out of the discussion, is uploaded to the course repository within 3 days after the presentation.

4.2 Discussion

A 15-minute discussion collapses into silence, or into a conversation between one instructor and one student, unless it is given a structure. The following protocol is used for every set.

- **Panel.** Two students are assigned in advance as the discussion panel for each presentation. Over the semester, each student accordingly gives one presentation and serves twice on a panel. The instructors moderate and take part in the discussion.
- **Preparation.** Panelists may read the paper in advance, and are encouraged to do so when the topic is far from their own area. The discussion is nevertheless meant to build on what the presenter actually said.
- **The panel has two tasks.**
  - *Reviewer.* Interrogate the work as a reviewer would — what the paper establishes and what it does not, weaknesses in the method, missing baselines or ablations, and claims that outrun the evidence.
  - *Research-direction brainstorming.* Work with the presenter, over the remainder of the 15 minutes, on where the line of work could go next.
- **Questions from the floor** are welcome throughout, but the panel's questions take precedence. The moderator opens the floor whenever the panel's line of questioning is exhausted.
- **Example structure of the 15 minutes.** The following is one way a set may run, offered as an illustration rather than as a rule.
  - *0–3 min.* Panelist A opens as reviewer: what the paper convincingly establishes, what it does not, and two prepared questions.

  - *3–7 min.* Panelist B follows up on the weaknesses; questions from the floor if the panel is done.
  - *7–13 min.* Panel and presenter brainstorm follow-up directions together; the instructors and the floor join in.
  - *13–15 min.* An instructor names the one or two most promising directions raised and states what a first experiment testing them would look like.

- **Panel report.** Within one week of the session, each panelist submits a short report (1–2 pages) covering a summary of the presented work, the substance of the discussion, and the follow-up research directions that came out of it. Reports will be uploaded to the class webpage for the class, and any student is free to pursue any direction recorded in them, including for the final research memo and their own research.

- **Norms.** Criticism is addressed to the paper and never to the presenter. Questions that begin from a misunderstanding are welcome; in a seminar on a moving research frontier, the misunderstandings are usually shared by half the room.

5 Grade

This section is a draft and is not final. The weights and components below are a starting point for discussion between the co-instructors and will be finalized before the course registration period.

- Attendance (10%)
  - Students who miss more than one-third of the classes without a valid excuse will receive 0 points. Otherwise, full credit will be given. Exceptions can be made when the cause of absence is deemed unavoidable by the course instructor.

- Attitude (10%)
  - Asking a thoughtful question during the lecture or participating in the discussion as a floor audience earns 1–2 points.

- Presentation (50%)
  - Technical accuracy and depth, clarity of delivery and slides, and the quality of the presenter's own critique and open questions.

- Discussion and participation (30%)
  - Service on two discussion panels and the resulting panel reports, together with questions and contributions from the floor in the other sessions.

- No midterm or final examination is planned for this course.

6 Schedule

The following is a tentative schedule for the course. Two Fridays of the Fall 2026 semester fall on national holidays (Chuseok, Sep. 25 and Hangul Day, Oct. 9), so two online (pre-recorded) sessions are scheduled; their exact dates will be announced in Week 1.

Week	Date	Topics	Lecturer
1	4 Sep	Introduction	SD & JH
2	11 Sep	Multimodal Foundation Models I	SD
3	18 Sep	Multimodal Foundation Models II	SD
4	25 Sep	Multimodal Foundation Models III (Online; Chuseok)	SD
5	2 Oct	Multimodal Generation I	JH
6	9 Oct	Multimodal Generation II (Online; Hangul Day)	JH
7	16 Oct	Multimodal Generation III	JH
8	23 Oct	Paper Review and Discussion Session 1	JH
9	30 Oct	Paper Review and Discussion Session 2	JH
10	6 Nov	Paper Review and Discussion Session 3	SD & JH
11	13 Nov	Paper Review and Discussion Session 4	SD & JH
12	20 Nov	Paper Review and Discussion Session 5	SD & JH
13	27 Nov	Paper Review and Discussion Session 6	SD & JH
14	4 Dec	Paper Review and Discussion Session 7	SD & JH
15	11 Dec	No Class (NeurIPS 2026)	-
16	18 Dec	Paper Review and Discussion Session 8	SD & JH