



UNIVERSIDAD
POLITÉCNICA
DE MADRID



LEARNING GUIDE

INTRODUCTION TO MULTIMODAL ARTIFICIAL INTELLIGENCE

MÁSTER EN INGENIERÍA DE TELECOMUNICACIÓN

ACADEMIC YEAR	CODE	COURSE	CREDITS	TYPE	SEMESTER
2026-27	93001387	2	5 ECTS	OPTIONAL/ ELECTIVE	SECOND SEMESTER
LANGUAGE	ASSIGNED CENTER				
ENGLISH	E.T.S. DE INGENIEROS DE TELECOMUNICACION				
ASSIGNED DEPARTMENT					
SEÑALES, SISTEMAS Y RADIOCOMUNICACIONES					

COORDINATOR
LUIS ALFONSO HERNÁNDEZ GÓMEZ - luisalfonso.hernandez@upm.es

REQUIRED PREREQUISITES

REQUIRED COURSES

—

RECOMMENDED PRIOR KNOWLEDGE

RECOMMENDED COURSES

- (93001369)

OTHER RECOMMENDATIONS

- Elementary background in linear algebra, probability and statistics.
- Elementary background in multimedia signal processing (image, video, speech, audio, music,...).
- Basic programming experience in Python, familiarity with the use of Jupyter / Google Colab notebook.
- No prior knowledge of Machine Learning, Deep Learning, Generative AI or AI Agents is assumed.



UNIVERSIDAD
POLITÉCNICA
DE MADRID



RESULTS

RD-822/2021

Knowledge

KG1 Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación

Skills

HG2 Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio

HE1 Capacidad para aplicar métodos de la teoría de la información, la modulación adaptativa y codificación de canal, así como técnicas avanzadas de procesamiento digital de señal a los sistemas de comunicaciones y audiovisuales

Competences

CG4 Que los estudiantes sepan comunicar sus conclusiones ¿y los conocimientos y razones últimas que las sustentan¿ a públicos especializados y no especializados de un modo claro y sin ambigüedades

CG5 Que los estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo

CT4 Capacidad para trabajar de forma efectiva como individuo, organizando y planificando su propio trabajo, de forma independiente o como miembro de un equipo

CT3 Capacidad para adoptar soluciones creativas que satisfagan adecuadamente las diferentes necesidades planteadas

DESCRIPTION

IMAN (Introduction to Multimodal Artificial Intelligence) is an introductory course on the use of Artificial Intelligence in Multimedia. The goal is to give students a clear view of the field, from the basics of Machine Learning and Deep Learning to the latest Generative AI, Multimodal Large Language Models and Multimodal AI Agents, covering applications in image, video, speech, audio, music and text.

The course will follow an essentially practical methodology. Lectures will introduce the main concepts and technologies, while practical activities will allow students to collaborate in the design, experimentation and

evaluation of different AI technologies applied to multimodal applications. Through course presentations and final reports, students will integrate the theoretical and practical work completed throughout the course.

Due to the rapid evolution of AI, the technological examples included in this guide reflect the state of the field at the time of writing and will be updated every academic year.

TOPICS

1. Multimodal AI

- 1.1. AI applications in multimedia: extracting information, generating content, enabling autonomous behaviour
- 1.2. The three learning paradigms (supervised, unsupervised / self-supervised, reinforcement) and Generative AI: the conceptual map of the course
- 1.3. Interaction in natural language: multimodal alignment
- 1.4. Ethics from day one: deepfakes, voice cloning and manipulation. AI, AGI and Superintelligence as an open questions

2. From Classical ML to Multimodal Models: an Overview Case Study (audio information extraction)

- 2.1. Feature engineering and classical ML.
- 2.2. End-to-end Deep Learning view vs Classical ML
- 2.3. Representation learning, Self-supervised multimodal representations.
- 2.4. Native multimodal models

3. Multimodal AI Design, Train, Evaluation and Benchmarking

- 3.1. Understanding the Multimodal AI Development Pipeline
- 3.2. Train / validation / test, cross-validation, data leakage
- 3.3. Classification and regression metrics
- 3.4. Benchmarks and leaderboards for Multimodal AI Models

4. Deep Learning Architectures for Multimedia

4.1. 1D CNNs for time-varying signals: speech, music, audio

4.2. 2D CNNs for images and video frames

4.3. Building blocks: convolution, pooling, normalization, residual connections; the limits of locality

5. Sequence Models and the Transformer

5.1. Recurrent networks (LSTMs) and long-range dependencies

5.2. Attention, Residual Stream and the Transformer

5.3. Tokenization across modalities: word pieces, image patches (ViT), spectrogram patches (AST)

6. Representation Learning and Multimodal Alignment

6.1. Self-supervised learning: representations without manual labels

6.2. Contrastive learning and shared embedding spaces

6.3. Cross-modal retrieval, zero and few-shot

7. Signal Processing and Deep Learning

7.1. Differentiable DSP (DDSP): trainable filters, oscillators, reverb models and vocoders

7.2. Audio applications: virtual analog modeling, room acoustics, source separation, pathological voice analysis

8. Generative Models for Audio, Speech and Music

8.1. Introduction to VAEs, GANs, Normalizing Flows, Diffusion models

8.2. Applications such as voice cloning and music generation

8.3. Evaluation of generated audio, speech, music,...

9. Generative Models for Image and Video

9.1. Text-to-image generation with diffusion models

9.2. Video generation: temporal consistency, motion coherence, audio-video synchrony

9.3. Evaluation assessing quality, realism, diversity,...



UNIVERSIDAD
POLITÉCNICA
DE MADRID



10. Multimodal Large Language Models

- 10.1. From text-only LLMs to MLLMs: modality connectors and multimodal token spaces
- 10.2. Current proprietary and open-weights reference systems
- 10.3. Multimodality for World Models
- 10.4. Limits: hallucination, sensitivity, benchmarks.

11. Multimodal AI Agents

- 11.1. The agent loop: perceive, plan, act, observe
- 11.2. Components: foundation model, tools, memory, harnessing
- 11.3. Agents for multimedia production; real-world capability and real-world responsibility

12. Ethics

- 12.1. A framework for multimedia AI ethics: consent, provenance, bias and fairness, safety
- 12.2. AI, AGI and Superintelligence: the session-1 question revisited
- 12.3. Open research problems for future AI-in-Multimedia

ACTIVITIES

☒ Assessment activity ☐ Learning activity

Lectures Topics 1 to 12

WEEKS

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

E

FORMAT	TOTAL DURATION	LOCATION	GROUP
In-person	20h 00m	Classroom (scheduled)	Normal

TOPICS

T1

T2

T3

T12

T11

T10

T9

T8

T7

T4

T5

T6

METHODOLOGY

Lecture



UNIVERSIDAD
POLITÉCNICA
DE MADRID



	Practical activities, small projects and demonstrations			WEEKS 123456789101112131415E															
FORMAT In-person	TOTAL DURATION 20h 00m	LOCATION Classroom (scheduled)	GROUP Normal																
TOPICS																			
T1T11T8T9T10T7T4T5T6T2T3T12																			
METHODOLOGY																			
Laboratory assignmentsProject-based learning																			

	Class presentations for monitoring and feedback on theory and practical activities			WEEKS 123456789101112131415E																
FORMAT In-person	TOTAL DURATION 08h 00m	LOCATION Classroom (scheduled)	ASSESSMENT Progressive	WEIGHT 10%	MINIMUM GRADE 5															
TOPICS																				
T1T2T3T4T5T6T7T8T9T10T11T12																				
METHODOLOGY																				
Group presentation																				
ASSESSED OUTCOMES																				
KG1HG2HE1CG4CG5CT3CT4																				

	Final report on AI in Multimedia			WEEKS 123456789101112131415E																
FORMAT Not in person, without teacher	TOTAL DURATION 00h 00m	ASSESSMENT Progressive, Global, Retake (Extraordinary)			WEIGHT 20%	MINIMUM GRADE 5														
TOPICS																				
T1T2T3T4T5T6T7T8T9T10T11T12																				
METHODOLOGY																				
Work in groupIndividual work																				
ASSESSED OUTCOMES																				
KG1HG2HE1CG4CG5CT3CT4																				



UNIVERSIDAD
POLITÉCNICA
DE MADRID



Individual written assessment based on all course contents

WEEKS

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

E

FORMAT

TOTAL DURATION

LOCATION

ASSESSMENT

WEIGHT

MINIMUM GRADE

In-person

02h 00m

Classroom (scheduled)

Progressive, Global, Retake (Extraordinary)

50%

3.5

TOPICS

T1

T12

T11

T10

T9

T8

T7

T2

T3

T4

T5

T6

METHODOLOGY

Written exam

ASSESSED OUTCOMES

KG1

HG2

HE1

CG5

CT3

Course Presentations and Final Report Defense (extraordinary)

WEEKS

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

E

FORMAT

TOTAL DURATION

LOCATION

ASSESSMENT

WEIGHT

MINIMUM GRADE

In-person

00h 30m

Exam Classroom

Retake (Extraordinary)

30%

5

TOPICS

T1

T2

T3

T4

T5

T6

T7

T8

T9

T10

T11

T12

METHODOLOGY

Group presentation

Individual presentation

ASSESSED OUTCOMES

KG1

HG2

HE1

CG4

CG5

CT3

CT4

Personal work to study theoretical concepts and complete practical exercises and small projects..

WEEKS

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

E

FORMAT

TOTAL DURATION

GROUP

Not in person, without teacher

80h 00m

Normal

TOPICS

T1

T2

T3

T4

T5

T6

T7

T8

T9

T10

T11

T12

METHODOLOGY

Project-based learning

Cooperative activities

Laboratory assignments



UNIVERSIDAD
POLITÉCNICA
DE MADRID



Course Report oral presentation and defense

WEEKS

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

E

FORMAT

TOTAL DURATION

ASSESSMENT

WEIGHT

MINIMUM GRADE

In-person

00h 15m

Progressive

20%

5

TOPICS

T1

T2

T3

T4

T5

T6

T7

T8

T9

T10

T11

T12

METHODOLOGY

Group presentation

Individual presentation

ASSESSED OUTCOMES

KG1

HG2

HE1

CG4

CG5

CT3

CT4

Course Presentations and Final Report Defense (global)

WEEKS

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

E

FORMAT

TOTAL DURATION

LOCATION

ASSESSMENT

WEIGHT

MINIMUM GRADE

In-person

00h 25m

Exam Classroom

Global

30%

5

TOPICS

T1

T12

T11

T10

T9

T8

T7

T6

T5

T4

T3

T2

METHODOLOGY

Group presentation

Individual presentation

ASSESSED OUTCOMES

KG1

HG2

HE1

CG4

CG5

CT3

CT4

WEEKLY SCHEDULE

Assessment activity Learning activity Week with activity

E

 Exam period

Activity	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	E
Lectures Topics 1 to 12	.	.	.	.	.	.	.	.	.	.	.	.	.			
Practical activities, small projects and demonstrations	.	.	.	.	.	.	.	.	.	.	.	.	.			
Class presentations for monitoring and feedback on theory and practical activities			.	.	.	.	.	.	.	.	.	.	.			
Final report on AI in Multimedia																.

Activity	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	E
 Individual written assessment based on all course contents																.
 Course Presentations and Final Report Defense (extraordinary)																.
 Personal work to study theoretical concepts and complete practical exercises and small projects..	.	.	.	.	.	.	.	.	.	.	.	.	.			
 Course Report oral presentation and defense													.			.
 Course Presentations and Final Report Defense (global)												.	.			.

ASSESSMENT CRITERIA

General criteria

Assessment will verify that students have acquired the competences of the course. The extraordinary call will therefore use the same types of assessment techniques as the ordinary call (EX, TG and PG/PI).

Assessment will be based on three complementary activities:

EX – Individual Written Examination (50%)

Students will complete an individual written examination covering all the contents of the course. The examination will assess their understanding of the main concepts, technologies and applications studied throughout the course. A minimum grade of **3.5 out of 10** is required.

TG – Final Course Report and Oral Defence on AI in Multimedia (40%)

The course expects students to make responsible and transparent use of Artificial Intelligence tools in their academic work. As these tools may support the preparation of reports and practical activities, the assessment will include an oral presentation and defence. This will allow lecturers to evaluate each student's understanding of the theoretical concepts, experimental developments, methodological decisions and results presented in the Course Report.

Therefore assessment item will consist of two components:

- **Written Course Report (20%):** Students will prepare a final report describing and discussing the theoretical and practical activities completed during the course.

- **Oral presentation and defence (20%):** Students will present and defend the theoretical and experimental work included in the report. The oral defence will assess their understanding of the methods used, the decisions made, the results obtained and the conclusions presented.

The Course Report may be prepared individually or in groups, according to the instructions provided during the course. When the work is completed in groups, each student must demonstrate their individual understanding and contribution during the oral defence. A minimum grade of **5 out of 10** is required for the TG assessment item.

PG/PI – Progressive Presentations on AI in Multimedia (10%)

Students will give progressive in-class presentations related to the development of the Course Report. These presentations will be used to provide guidance and feedback and to assess students' communication skills, progress and understanding of the activities being developed. An additional presentation outside regular class hours may also be required. A minimum grade of **5 out of 10** is required.

Assessment criteria for the regular exam period. Progressive assessment

Assessment criteria

Progressive evaluation

Progressive evaluation will be the preferred assessment method, as it is better suited to supporting the students' learning process throughout the course. It will consist of the following activities:

PG – Progressive presentations on AI in Multimedia (10%)

Students will give short presentations in class throughout the course to present the progress of the technical and practical activities developed for their Course Report. These presentations will allow lecturers to review the work, provide guidance and feedback, and help students improve their activities before the final submission. They will also assess students' understanding of their work and their ability to explain the methods, decisions and preliminary results obtained. A minimum grade of **5 out of 10** is required.

TG – Final Course Report and Oral Defence on AI in Multimedia (40% = 20% + 20%)

The course expects students to make responsible and transparent use of Artificial Intelligence tools in their academic work. Since these tools may support the preparation of reports and practical activities, the assessment will include an oral presentation and defence to verify each student's understanding of the theoretical concepts, experimental developments, methodological decisions and results included in the Course Report.

This assessment item will consist of two components:

- **Written Course Report (20%):** Students will prepare a final report integrating and discussing the theoretical concepts and the technical and practical activities completed during the course. Draft versions of the report may

be submitted and reviewed at different stages, allowing students to receive feedback and progressively improve their work before the final submission.

- **Report Oral presentation and defense (20%):** Students will present and defend the theoretical and experimental work included in the report. Under the progressive evaluation system, this final oral presentation may take place during the last weeks of the course, as a continuation of the progressive class presentations, or on the date established for the final examination.

The Course Report may be prepared individually or in groups, according to the instructions provided during the course. When the work is completed in groups, each student must demonstrate their individual understanding and contribution during the oral defence. A minimum grade of **5 out of 10** is required for the TG assessment item.

EX – Individual assessment of the course contents (50%)

The individual assessment will evaluate students' understanding of all the contents of the course. It may consist of an individual written examination or, under the progressive evaluation system, the written examination may be fully replaced by a set of short individual assessments carried out throughout the course. These continuous assessments will evaluate students' progressive understanding of the main concepts, technologies and applications of Multimedia AI. A minimum grade of **3.5 out of 10** is required.

Activities

- Class presentations for monitoring and feedback on theory and practical activities (10%)
- Final report on AI in Multimedia (20%)
- Individual written assessment based on all course contents (50%)
- Course Report oral presentation and defense (20%)

Assessment criteria for the regular exam period. Global assessment

Assessment criteria

Assessment Criteria for the Ordinary Examination Period: Global Assessment

The global assessment modality will evaluate the same course competences as the progressive evaluation, but it will not require continuous attendance or participation in the progressive assessment activities. It will consist of the following:

EX – Individual Written Examination (50%)

Students will complete an individual written examination covering all the contents of the course. The examination will assess their understanding of the main concepts, technologies and applications studied throughout the course. In the global assessment modality, the EX grade will be based exclusively on this final examination and will not include continuous assessments completed during the course. A minimum grade of **3.5 out of 10** is required.

TG – Written Course Report on AI in Multimedia (20%)

Students will submit a final report integrating and discussing the theoretical concepts and the technical and practical activities developed for the course. Under this assessment modality, only the final version of the report will be evaluated. The report may be prepared individually or in groups, according to the course instructions.

PG/PI – Course Presentations and Final Report Defense (30%)

Students will be evaluated about both their global understanding of main concepts along the course, and the specific theoretical and experimental work included in their Course Report. This assessment item incorporates both the **10% assigned to the course presentation** and the **20% assigned to the oral defence of the report** under the progressive evaluation system.

The course expects students to make responsible and transparent use of Artificial Intelligence tools in their academic work. Since these tools may support the preparation of reports and practical activities, the assessment of the Course Report will explicitly include an oral presentation and defence to verify each student's understanding of the theoretical concepts, experimental developments, methodological decisions and results included in the Course Report.

The oral presentation and defence may take place during two different regular course classes (one for the course and another for the final course report), when students are able to attend them, or during the dates established for the ordinary examination period. A minimum grade of **5 out of 10** is required for the combined TG and PG/PI assessment components.

The presentation may be completed in groups (**PG**) or individually (**PI**), depending on how the report has been prepared. However, each student must demonstrate their individual understanding of the theoretical concepts, experimental developments, methods used, decisions made, results obtained and conclusions presented.

Activities

- Final report on AI in Multimedia (20%)
- Individual written assessment based on all course contents (50%)
- Course Presentations and Final Report Defense (global) (30%)

Assessment criteria for the retake (extraordinary) exam period

Assessment criteria

Assessment during the **extraordinary examination** period will follow the same procedure, assessment activities, weightings and minimum grade requirements established for the global assessment modality in the ordinary examination period.

Academic Fraud

Any assessment activity or submitted work may be followed by an additional oral assessment to verify that it has been completed by the student without unauthorised assistance. In accordance with the UPM assessment regulations, where academic fraud is confirmed in an assessment activity, the student or students involved will receive a final grade of zero for the corresponding examination period, whether ordinary or extraordinary.

Activities

- Final report on AI in Multimedia (20%)
- Individual written assessment based on all course contents (50%)
- Course Presentations and Final Report Defense (extraordinary) (30%)

RESOURCES

Bibliography

An Introduction to Statistical Learning: with Applications in Python

<https://www.statlearning.com/>

Deep Learning: Foundations and Concepts by Christopher M. Bishop and Hugh Bishop

<https://www.bishopbook.com/>

Web

scikit-learn Machine Learning in Python

<https://scikit-learn.org/stable/>

PyTorch Tutorials

<https://pytorch.org/tutorials/>

Andrej Karpathy blog About Hacker's guide to Neural Networks

<https://karpathy.github.io/>

Hugging Face AI Community

<https://huggingface.co/>

IMAN at Moodle

<https://moodle.upm.es/titulaciones/oficiales/> Reading list of recent papers on Multimodal LLMs, multimodal agents and generative multimedia. Updated every academic year; published in Moodle at the start of the course

Generative Deep Learning

<https://github.com/3p5ilon/ML-books/blob/main/generative-deep-learning-teaching-machines-to-paint-write-compose-and-play.pdf>

Dive into Deep Learning

<https://d2l.ai/>

Bibliography

Li, C., et al. (2024). Multimodal Foundation Models: From Specialists to General-Purpose Assistants. Foundations and Trends in Computer Graphics and Vision, 16(1–2), 1–214.

Yao, H., et al. (2025). A Survey on Agentic Multimodal Large Language Models. arXiv:2510.10991.

Web

librosa + torchaudio — feature extraction and audio I/O for the classical-ML and DSP activities. <https://librosa.org> · <https://pytorch.org/audio>
<https://librosa.org> <https://pytorch.org/audio>

Meta AudioCraft (MusicGen / AudioGen) Text-to-music and text-to-audio generation

<https://github.com/facebookresearch/audiocraft>

Bibliography

Chih-Kai Yang, Neo S. Ho, and Hung-yi Lee. "Towards Holistic Evaluation of Large Audio-Language Models: A Comprehensive Survey." In Proceedings of EMNLP 2025 (Main Conference). arXiv:2505.15957.

SUSTAINABLE DEVELOPMENT GOALS



Quality Education



Decent Work and Economic Growth



Industry, Innovation and
Infrastructure



Reduced Inequalities



Responsible Consumption and
Production



Peace, Justice and Strong Institutions

OTHER INFORMATION

IMAN contributes primarily to **SDG 4 — Quality Education**, particularly Target 4.4, by providing students with up-to-date technical and professional skills in a rapidly developing field, relevant to employment, innovation and entrepreneurship.

The course also supports **SDG 8 — Decent Work and Economic Growth**, particularly Target 8.2, by preparing students to apply advanced digital technologies that can contribute to technological upgrading, innovation and productivity. Its focus on current AI models, experimentation and practical projects is also directly related to **SDG 9 — Industry, Innovation and Infrastructure**, particularly Target 9.5, which promotes scientific research and the technological capabilities of industrial sectors.

Through the study of bias, fairness, accessibility and inclusive design in multimedia AI systems, the course contributes to **SDG 10 — Reduced Inequalities**, particularly Targets 10.2 and 10.3, concerning social inclusion, equal opportunities and the reduction of discriminatory practices.

Finally, the ethics component examines the responsible development and use of multimedia AI. This includes its possible social and environmental impacts, connecting the course with **SDG 12 — Responsible Consumption and Production**, particularly Target 12.8. The discussion of synthetic media, misinformation, transparency and access to reliable information also relates to **SDG 16 — Peace, Justice and Strong Institutions**, particularly Target 16.10.



UNIVERSIDAD
POLITÉCNICA
DE MADRID

