POSTER SESSION BOOKLET

http://www.dmi.unict.it/icvss

University of Catania - University of Cambridge

International Computer Vision Summer School 2026

Computer Vision for Spatial and Physical Intelligence

Sicily, 5 - 11 July 2026

International Computer Vision Summer School

Computer Vision is the science and technology of making machines that see. It is concerned with the theory, design and implementation of algorithms that can automatically process visual data to recognize objects, track and recover their shape and spatial layout.

The International Computer Vision Summer School - ICVSS was established in 2007 to provide both an objective and clear overview and an in-depth analysis of the state-of-the-art research in Computer Vision. The courses are delivered by world renowned experts in the field, from both academia and industry, and cover both theoretical and practical aspects of real Computer Vision problems.

The school is organized every year by University of Cambridge (Computer Vision and Robotics Group) and University of Catania (Image Processing Lab). The general entry point for past and future ICVSS editions is:

http://www.dmi.unict.it/icvss

ICVSS Poster Session

The International Computer Vision Summer School is especially aimed to provide a stimulating space for young researchers and Ph.D. Students. Participants have the possibility to present the results of their research, and to interact with their scientific peers, in a friendly and constructive environment.

This booklet contains the abstract of the posters accepted to ICVSS 2026.

Best Presentation Prize A subset of the submitted posters will be selected by the school committee for short oral presentation. A best presentation prize will be given to the best presentations selected by the school committee.

Scholarship A scholarship will be awarded to the best PhD student attending the school. The decision is made by the School Committee at the time of the School, taking into account candidates' CV, poster and oral presentation.

Sicily, June 2026

Roberto Cipolla
Sebastiano Battiato
Giovanni Maria Farinella

List of Posters [1]

1. DEEP PROBABILISTIC SUPERVISION FOR IMAGE CLASSIFICATION Adelöw A., Gamba M., Maki A.

2. SYNTHFORENSICS: BENCHMARKING AND EVALUATING PEOPLE-CENTRIC SYNTHETIC VIDEO DEEPFAKES Leotta R., Sambataro S. A., Ragaglia C. V., Casu M., Petralia Y., Guarnera F., Guarnera L., Battiato S.

3. NLS: NOVEL LATENT SYNTHESIS Anadón X., Batlle V., Mur-Labadia L., Montiel J. M.

4. VISIBLE OBJECT-STATE PROXY GEOMETRY FOR POLICY LEARNING Antypas I. , Averta G. , Garcia N.

5. SCENETOK: A COMPRESSED, DIFFUSABLE TOKEN SPACE FOR 3D SCENES Asim M., Wewer C., Lenssen J.

6. GEOMETRY ESTIMATION USING DENSE CORRESPONDENCES Astermark J., Heyden A., Larsson V.

7. ROOM ENVELOPES: A SYNTHETIC DATASET FOR INDOOR LAYOUT RECONSTRUCTION FROM IMAGES Bahrami Sam., Campbell Dylan.

8. ASSEMBLYHANDS-X: MODELING HAND, BODY, AND VISUAL CONTEXT FOR UNDERSTANDING BIMANUAL HUMAN ACTIVITIES Banno T., Suzuki N., Ohkawa T., Liu R., Kwon T., He K., Shinoda R., Furuta R., Sato Y.

9. PANO3D: UNIFIED 3D RECONSTRUCTION AND PANOPTIC SEGMENTATION Victor Barberteguy, Ahmet Iscen, Mathilde Caron, Alireza Fathi, Gül Varol, Cordelia Schmid

[1] Posters are ordered by surname of the speaker. Each poster is identified by a number.

10. A PAN-EUROPEAN MULTI -SEASONAL LAND COVER MAPPING MODEL BARCO L., GALATOLA M., ARNAUDO E., BRAGAGNOLO A., GARZA P., ROSSI C.

11. COSY: COMPOSITIONAL 3DGS SYNTHESIS FOR DISENTANGLED HUMAN HEAD EDITING Barthel F., De Mello S., Nagano K., Morgenstern W., Hilsmann A., Eisert P.

12. ON THE GENERALIZATION OF OPTICAL FLOW: QUANTIFYING ROBUSTNESS TO DATASET SHIFTS Katrin Bauer, Andrés Bruhn, Jenny Schmalfuss

13. MACES-APO: MULTI-AGENT CO-EVOLUTIONARY SIMULATION FOR AUTOMATIC PROMPT OPTIMIZATION Bayoumi O., Cinque L., Foresti G.F.

14. HIERARCHICAL OBJECT-CENTRIC REPRESENTATION LEARNING Behrad F., Tuytelaars T., Wagemans J.

15. SLAD : SHARED LORA ADAPTERS FOR TASK-SPECIFIC DISTILLATION Bensaid Reda., Bendou Yassir., Gripon Vincent., Leduc-Primeau François.

16. LINEAR MODEL MERGING UNLOCKS SIMPLE AND SCALABLE MULTIMODAL DATA MIXTURE OPTIMIZATION Berasi D., Farina M., Mancini M., Ricci E.

17. ROBUST RECOGNITION OF CARDIAC PATHOLOGIES BASED ON PHONOCARDIOGRAM ANALYSIS AND DEEP LEARNING Beritelli L., Avanzato R., Guarnera L., Battiato S., Beritelli F.

18. ONE TARGET TO ALIGN THEM ALL: LIDAR, RGB AND EVENT CAMERAS EXTRINSIC CALIBRATION FOR AUTONOMOUS DRIVING Bertogalli, Boracchi, Magri

19. ROBOTIC POLICY ADAPTATION VIA WEIGHT-SPACE META-LEARNING. Christian Bianchi, Siamak Yousefi, Alessio Sampieri, Andrea Roberti, Luca Rigazio, Fabio Galasso, Luca Franco

20. DOPO: DENSE ONLINE PREFERENCE OPTIMIZATION FOR CROSS-DATASET MOTION DIFFUSION ADAPTATION Macaluso G., Mandelli L., Bicchierai M., Berretti S., Bagdanov A.D.

21. REWIS3D: RECONSTRUCTION IMPROVES WEAKLY-SUPERVISED SEMANTIC SEGMENTATION Jonas Ernst*, Wolfgang Boettcher*, Lukas Hoyer, Jan Eric Lenssen, Bernt Schiele

22. EDGES OF PHYSICAL INTELLIGENCE Bonazzi P., Magno M.

23. TINY INFERENCE TIME SCALING WITH LATENT VERIFIERS Bucciarelli D, Turri E, Baraldi L, Cornia M, Cucchiara R

24. GAP: GEOMETRIC ANCHOR PRE-TRAINING FOR DATA-EFFICIENT VISUOMOTOR LEARNING OF MANIPULATION TASKS Buoso D., Protopapa A., Di Carlo S., Pistilli F., Averta G

25. AUTOMATED PREDICTION OF PARAVALVULAR REGURGITATION BEFORE TRANSCATHETER AORTIC VALVE IMPLANTATION Michele Cannito, Riccardo Renzulli, Marco Grangetto, Fabrizio D'Ascenzo

26. SHIELDCLIP: SELECTIVE SAFETY ALIGNMENT FOR HARMFUL CONTENT MITIGATION IN MULTIMODAL FOUNDATION MODELS Poppi T., Cappelletti S., Poppi S., Cornia M., Baraldi L., Cucchiara R.

27. PHYSFORMER: LEARNING TO SIMULATE MECHANICS IN WORLD SPACE Chen Y., Lan Y., Vedaldi, A.

28. DIFFUSING DEBIAS (DDB): SYNTHETIC BIAS AMPLIFICATION FOR MODEL DEBIASING Ciranni M., Pastore V.P., Di Via R., Tartaglione E., Odone F., Murino V.

29. WATCH, LEARN, ASSIST: ENABLING BIMANUAL MANIPULATION WITH THE HANNES PROSTHESIS Columbaro M., Vasile F., Boccardo N., Natale L.

30. FLAME: A DEPLOYABLE FRAMEWORK FOR LAYER-WISE AUTOMATIC MIXED-PRECISION QUANTIZATION ON EMBEDDED DEVICES Corti G., Vitali M., Vacis N., Palladino V., Merigo L., Pidò S., Matteucci M.

31. LEARNING GENERALIZABLE DYNAMICS MODELS WITH GRAPH NEURAL NETWORKS FOR NOVEL TOOL DESIGN Cugito N., Allen K.

32. CONVMAMBA: A HYBRID CNN-MAMBA BACKBONE FOR ANIMAL POSE ESTIMATION Daadouch S., Zhao K., Gelautz M., Roth P.M

33. MITIGATING THE MODALITY GAP IN TEXT-DRIVEN SEMANTIC SEGMENTATION D'ASARO F., BOTTINO A., RIZZO G.

34. CAMC2V: CONTEXT-AWARE CONTROLLABLE VIDEO GENERATION Luis Denninger, Sina Mokhtarzadeh Azar, Juergen Gall

35. EXECUTION-AWARE VLA FOR ROBUST MANIPULATION Dey Sombit, Albanese Giuliano, Zaech Jan-Nico, Van Gool Luc , Paudel Danda

36. MAKING FEW-SHOT SEGMENTATION ACTUALLY WORK De Marinis P., Vessio G., Castellano G.

37. GOOD, FAST, AND CHEAP: VISUAL-INERTIAL TRACKING FOR LOW LATENCY SPATIAL AGENTS de Mayo M., Pire T., Cremers D.

38. ORTHOTRACK: CONTINUOUS 6-DOF UAV TRAJECTORY ESTIMATION ANCHORED IN PUBLIC ORTHOPHOTOS Oussema Dhaouadi, Zuria Bauer, Johannes Michael Meier, Olaf Wysocki, Marc Pollefeys, Daniel Cremers

39. HAC: PARAMETER-EFFICIENT HYPERBOLIC ADAPTATION OF CLIP FOR ZERO-SHOT VQA Dibitonto N1., Beyan N2., Murino N3.

40. ENHANCING OUT-OF-DISTRIBUTION DETECTION WITH EXTENDED LOGIT NORMALIZATION Ding Y., Liu X., Unger J., Eilertsen G.

41. GENERATIVE EDITING THROUGH FEW-SHOT DIFFUSION ADAPTATION Elezabi O., Zamfir E., Wu Z., Timfote R.

42. WHEN CAN GENERATED VIDEO BECOME RELIABLE 3D EVIDENCE? Fan C., Favaro P.

43. ONLINE VIDEO DEPTH ANYTHING: TEMPORALLY-CONSISTENT DEPTH PREDICTION WITH LOW MEMORY CONSUMPTION Feiden J., Küchler T., Zavadski D., Savchynskyy B., Rother C.

44. T-FUNS3D: TASK-DRIVEN HIERARCHICAL OPEN-VOCABULARY 3D FUNCTIONALITY SEGMENTATION Feng J., Sabzevari R.

45. TOWARDS 4D ENVIRONMENT RECONSTRUCTION OF GLACIERS FROM SPARSE AND NOISY DATA Foggin A., Smith W.

46. SCENARIOCONTROL: VISION-LANGUAGE CONTROLLABLE VECTORIZED LATENT SCENARIO GENERATION Gao L., Xu Y., Koch W., Ruffino S., Rowe L., Chalaki B., Rivkin D., Ost J., Girgis R., Bijelic M., Heide F.

47. EZ-SP: FAST AND LIGHTWEIGHT SUPERPOINT-BASED 3D SEGMENTATION Geist L., Landrieu L., Robert D.

48. SUPERHUMAN SAFE AND AGILE RACING THROUGH MULTI-AGENT REINFORCEMENT LEARNING Geles I., Bauersfeld L., Wulfmeier M., Scaramuzza D.

49. HUMANMOVEVQA: CAN VIDEO MLLMS REASON ABOUT HUMAN MOVEMENT IN VIDEOS? Pulkit Gera, Faegheh Sardari, Asmar Nadeem, Valentina Bono, Padraig Boulton, Adrian Hilton, Armin Mustafa

50. UNIFIED LIDAR PSEUDO-LABELING FOR ULTRA LONG RANGE TRUCKING PERCEPTION Ghilotti F., Brucker S., Palladin E., Saidy N., Sigal A., Matteucci M., Bijelic M., Heide F.

51. INFINITE-STORY: A TRAINING-FREE CONSISTENT TEXT-TO-IMAGE GENERATION Park J.*, Lee K.*, Gim J.*, Jo H., Oh M., Choi W., Hwang K., Kim J., Choi M., and Im S.

52. CONCEPT DYNAMICS IN DIFFUSION MODELS: FROM TEMPORAL FORMATION TO COMPOSITIONAL BINDING Görgu"n A., Schiele B. , Fischer J.

53. ARTICULATE-3D: DATASET & METHOD FOR INTERACTION UNDERSTANDING HALACHEVA A.-M., MIAO Y., ZAECH J.-N., WANG X., VAN GOOL L., PAUDEL D.

54. PUZZLE SIMILARITY: A CROSS-REFERENCE METRIC FOR ARTIFACT DETECTION IN UNSEEN VIEWS HOUSE FLOW: RECONSTRUCTING MULTIFLOOR HOUSE LAYOUTS Hermann N., Condor J., Didyk P., Engelmann F.

55. SELF-SUPERVISED ONLINE ROBOT-AGNOSTIC TRAVERSABILITY ESTIMATION FOR OPEN-WORLD ENVIRONMENTS Hindel J., Bultmann S., Masnavi H., Cattaneo D., Valada A.

56. TRAINABLE HIGHLY-EXPRESSIVE ACTIVATION FUNCTIONS Chelly I., Finder S. E., Ifergane S., Freifeld O.

57. PHYSICALLY PLAUSIBLE HUMAN–OBJECT INTERACTION GENERATION VIA ATTRIBUTE CLASSIFIER GUIDANCE Ikeuchi K., Ohkawa T., Shinoda R., Sato Y.

58. SEEING WITH SOUND Nazrul Ismail, Owais Ahmed Malik, Ong Wee Hong

59. ZERO-SHOT SIM-TO-REAL DETECTION AND SEGMENTATION FOR ROBOTIC EV BATTERY DISASSEMBLY Isoaho J.1*, Rätz R.1,2, Özen Ö.1, Ochsenbein C.1

60. SOKE-GRPO: REINFORCEMENT LEARNING FOR TEXT-TO-SIGN GENERATION Isotton G., Talon D., Ricci E.

61. OPTIMIZING INCOMPLETE, LARGE-SCALE AND SPARSE MULTI-GRAPH MATCHING Stricker S., Kahl M., Hutschenreiter L., Bernard F., Rother C., Savchynksyy B.

62. INVSPLAT: INVERSE FEED-FORWARD SCENE SPLATTING Karpikova P., Bian W., Xu H., Lensch H., Geiger A.

63. LEARNING GLOBAL CAMERA POSES FROM NOISY VIEW-GRAPHS FOR STRUCTURE FROM MOTION Khatib F., Galun M., Basri R.

64. SELF-AUGMENTED RESIDUAL 3DGS FOR NEXT BEST VIEW SELECTION Jun-Seong K., Oh T.-H., Eduardo P.-P., Jang Y.

65. ALL-DAY DEPTH COMPLETION VIA THERMAL-LIDAR FUSION Kim J., Kweon M., Shin U., Park J.

66. REFLETION-AWARE GENERATIVE NOVEL VIEW SYNTHESIS Kim G., Dong-Yeon S., Oh T. H.

67. NEURO-SYMBOLIC OUT-OF-DISTRIBUTION DETECTION VIA SCENE GRAPH REASONING Kirchheim K., Czarnecki K., Ortmeier F.

68. CAN AI UNDERSTAND GARMENT CONDITION? TOWARDS FULLY AUTOMATED TEXTILE SORTING FOR REUSE Kirillova N., Possegger H.

69. SURGE: IMPROVED SURFACE GEOMETRY IN POINT MAPS Knaebel K., Martin Garcia G., Schmidt C., Fradlin I., Nunes L., de Geus D., Leibe B.

70. SLIDER: SLIDER-GUIDED LATENT IMAGE DISCOVERY FOR EXPLAINABLE RETRIEVAL Kolouju P., Qaiser I., Xing E., Stylianou A., Souvenir R., Pless R., Jacobs N.

71. MMLANDMARKS: A CROSS-VIEW INSTANCE-LEVEL BENCHMARK FOR GEO-SPATIAL UNDERSTANDING Kristoffersen O., Sánchez A., Hannemose M., Dahl A., Papadopoulos D.

72. DO IMAGE EDITING MODELS UNDERSTAND LIGHTING? Küchler T.*, Feiden J.*, Nießner M., Rother C.

73. EARLY PROSTATE CANCER DETECTION USING AI-ASSISTED ABDOMINAL ULTRASOUND Kurucz L.M., Natali T., Mertens L.S., Van Leeuwen P.J., Dashtbozorg B., Ruers T., De Korte C.

74. A TRAINING-FREE STYLE PERSONALIZATION VIA SVD-BASED FEATURE DECOMPOSITION Lee K.*, Park J.*, Gim J.*, Choi W., Hwang K., Kim J., Im S.

75. SUBSPACEAD: TRAINING-FREE FEW-SHOT ANOMALY DETECTION VIA SUBSPACE MODELING Lendering Camile, Akdag Erkut, Bondarev Egor

76. EGOINTERACT: SYNTHETIC EGOCENTRIC VIDEOS GENERATION FOR INTERACTION UNDERSTANDING AND ANTICIPATION Leonardi R., Ragusa F., Materia D., Passanisi A., Fort J., Engel J., Farinella G.M.

77. CONFORMALIZED FLOWMATCHING FOR TRUSTWORTHY SYNTHETIC MEDICAL IMAGE GENERATION Li J., Li H.B.

78. ECHO2ECG: ENHANCING ECG REPRESENTATIONS WITH CARDIAC MORPHOLOGY FROM MULTI-VIEW ECHOS Liman, M., Turgut Ö., Müller A., Martens E., Rueckert D., Müller P.

79. INTENT ANALYSIS IN HISTORICAL VISUAL ARCHIVES Lin T., Aigner W., Sablatnig R.

80. VISUAL LOCALIZATION: FROM CONTINENTS TO CORRESPONDENCES Lindenberger P.

81. TOWARDS MULTIMODAL AI THAT KNOWS WHAT IT DOESN'T KNOW Liu M., Dong H., Fink O., Trapp M.

82. SPATIALLY AWARE WORLD ACTION MODEL VIA GEOMETRIC LATENT DIFFUSION Lopetegui-Gonzalez Javier, Pacaud Paul, Schmid Cordelia

83. VITTR3: VISUAL-INERTIAL TEST-TIME REFINEMENT OF 3D FOUNDATION MODELS Lozano E., Jaenal A., Civera J.

84. GAUSSIAN WORLD: 3DGS FROM SCENE RECONSTRUCTION TO UNDERSTANDING Ma, Mengjiao; Ma, Qi; Li, Yue; Cheng, Jiahuan; Yang, Runyi; Ren, Bin; Popovic, Nikola; Wei, Mingqiang; Sebe, Nicu; Van Gool, Luc; Gevers, Theo; Oswald, Martin R; Paudel, Danda Pani

85. DO 3D LLMS REALLY UNDERSTAND 3D SPATIAL RELATIONSHIPS? Ma X., Sun T., Chen S., Bhalgat Y., Gu J., Chang A.X., Armeni I., Laina I., Peng S., Prisacariu V.

86. LEXUS: LIDAR OUTLIER EXPOSURE FOR 3D UNKNOWN SEGMENTATION IN AUTONOMOUS DRIVING Marinai A., Rai S., Masone C., Tommasi T.

87. LEVERAGING GAZE AND SET-OF-MARK IN VLLMS FOR HUMAN-OBJECT INTERACTION ANTICIPATION FROM EGOCENTRIC VIDEOS Materia D., Ragusa F., Farinella G.M.

88. PROSKILL: SEGMENT-LEVEL SKILL ASSESSMENT IN PROCEDURAL VIDEOS Mazzamuto M., Di Mauro D., Francesca G., Farinella G.M., Furnari A.

89. SIGNIT: A COMPREHENSIVE DATASET AND MULTIMODAL ANALYSIS FOR ITALIAN SIGN LANGUAGE RECOGNITION Micieli A., Farinella G.M., Ragusa F.

90. WASABI: WEAKLY-SUPERVISED ANATOMY PRESERVING LESION SYNTHESIS WITH IMPLICIT ANOMALY LOCALIZATION FOR MEDICAL IMAGE BOOTSTRAPPING AND SEGMENTATION IMPROVEMENT Mitic B. , Prosch H. , Langs G.

91. PRIOR-GUIDED GRASP ADMISSIBILITY FOR SAFE LANDMINE MANIPULATION Miuccio A., Lebecque F., Le Flécher E., Hamesse C., Tsiogkas N., Detry R., Haelterman R.

92. CAN WE TRUST THE NEXT ACTION? GROUNDED FUTURE REASONING FOR RELIABLE EGOCENTRIC ANTICIPATION Mahsa MohammadiEshkaftaki(Mohammadi)

93. LIDO: LEARNING TO IDENTIFY OUT-OF-DISTRIBUTION OBJECTS FOR 3D LIDAR ANOMALY SEGMENTATION Mosco S., Fusaro D., Pretto A.

94. VOID: VIDEO OBJECT AND INTERACTION DELETION Motamed S., Harvey W., Klein B., Van Gool L., Yuan Z., Cheng T

95. WBC-CLIP: A MULTIMODAL VISION-LANGUAGE FRAMEWORK FOR MORPHOLOGY AWARE WHITE BLOOD CELL ANALYSIS Zedda L., Mura D.A., Manzo A., Di Ruberto C., Loddo A.

96. LEARNING FROM SYNTHETIC DATA VIA PROVENANCE-BASED INPUT GRADIENT GUIDANCE Nagano K., Fujii R., Hachiuma R., Sato F., Sekii T., Saito H.

97. VISUAL-INERTIAL EGOCENTRIC METRIC DEPTH ESTIMATION FOR TABLETOP ACTIVITIES OF DAILY LIVING Nalivayko Y., Wochner I.

98. VISUAL ODOMETRY THAT TUNES ITSELF Nascivera S., Bauersfeld L., Delaune J., and Scaramuzza D.

99. DYNAMIC CONTRAST ENHANCEMENT BRIDGE FOR CONTRAST-FREE BREAST MRI Newegy S., Nagarajan B., Radeva P.

100. ITALIANPARKS400K: LARGE SCALE EUROPEAN SPECIES DATASET AND BASELINE FOR AUTOMATED ECOLOGICAL ANALYSIS FROM CAMERA TRAP DATA Niccoli N., Seidenari L., Greco I., Salvatori M., Rovero F.

101. VIDEOMT: YOUR VIT IS SECRETLY ALSO A VIDEO SEGMENTATION MODEL Norouzi N., Zulfikar I. E., Cavagnero N., Kerssies T., Leibe B., Dubbelman G., de Geus D.

102. REAL-TIME UNDERWATER TRASH DETECTION AND SEGMENTATION ON RESOURCE-CONSTRAINED ROVS VIA EDGE AI OPTIMIZATION Jiregna Abdissa Olana, De zan Alberto, Tavaris Denis, Gian Luka Foresti, Carlo Drioli

103. ENABLING LARGE-SCALE ANALYSIS OF HISTORICAL LOGIC DIAGRAMS IN BYZANTINE MANUSCRIPTS Osburg Lilly, Dr. Götzelmann Germaine, Dr. Tonne Danah, Prof. Streit Achim

104. ONE PATCH TO CAPTION THEM ALL A UNIFIED ZERO-SHOT CAPTIONING FRAMEWORK Bianchi L., Pacini G., Carrara F., Messina N., Amato G., Falchi F.

105. PERCEPTION - REASONING DATASET FINGERPRINTS VIA TRAJECTORY ANALYSIS UNDER VLM SCALING Paez-Ubieta I.D.L., Pieters R.

106. EXPLAINING FOR BETTER MODELS, MODELING FOR BETTER EXPLANATIONS Parchami-Araghi Amin

107. 3D-LATTE: LATENT SPACE 3D EDITING FROM TEXTUAL INSTRUCTIONS Parelli M., Oechsle M., Niemeyer M., Tombari F., Geiger A.

108. ENIGMA-360: AN EGO-EXO DATASET FOR HUMAN BEHAVIOR UNDERSTANDING IN INDUSTRIAL SCENARIOS Ragusa F., Leonardi R., Mazzamuto M., Di Mauro D., Quattrocchi C., Passanisi A., D'Ambra I., Furnari A., Farinella G.M.

109. NAME THAT PART: 3D PART SEGMENTATION AND NAMING Paul S., Kaushik P., Vaidya A., Bhattad A., Yuille A.

110. INTERPRETABLE 3D NEURAL OBJECT VOLUMES FOR ROBUST CONCEPTUAL REASONING Pham N., Jesslen A., Schiele B., Kortylewski A., Fischer J.

111. GENERALIZABILITY ANALYSIS OF DEEP LEARNING PREDICTIONS OF HUMAN BRAIN RESPONSES TO AUGMENTED AND SEMANTICALLY NOVEL VISUAL STIMULI Piskovskyi V., Chimisso R., Patania S., Foulsham T., Vizzari G., Ognibene D.

112. R5DGS: SEMANTIC-AWARE 4D GAUSSIAN SPLATTING WITH RIGID BODY CONSTRAINTS FOR EFFICIENT DYNAMIC SCENE RECONSTRUCTION Gridusov D., Popov M., Kolyubin S.

113. UNSPOKEN BIASES IN END-TO-END DRIVING BENCHMARKS Porres D.

114. SEMANITC-AWARE, PHYSICS-INFORMED, GEOMETRY-GROUNDED WEATHER VIDEO SYNTHESIS Chenghao Qian, Nedko Savov, Lingdong Kong, Yeying Jin, Rui Song, Wenjing Li, Zhun Zhong, Jiaqi Ma, Gustav Markkula, Luc Van Gool

115. GROUNDING FOUNDATION MODELS: REPRESENTATIONS FOR SPATIAL AND PHYSICAL INTELLIGENCE Kaixian Qu, Mike Zhang, Zhengyu Fu, Cesar Cadena, Marco Hutter

116. ROBUST AND SECURE MRI: DEEP LEARNING ATROPHY ESTIMATION MEETS K-SPACE FINGERPRINTING Riccardo Raciti

117. PHYSICS-IQ VERIFIED Rädsch T., Asano Y. M., Kuehne H., Bauer S., Jaini P., Geirhos R., Lüth C. T.

118. EGO-EXTRA: VIDEO-LANGUAGE EGOCENTRIC DATASET FOR EXPERT-TRAINEE ASSISTANCE Ragusa F., Mazzamuto M., Forte R., D'Ambra I., Fort J., Engel J., Furnari A., Farinella G. M.

119. INDOOR ROOM RECONSTRUCTION USING SMARTPHONES Xuqian Ren, Juho Kannala, Esa Rahtu

120. EVER WONDERED WHAT YOU ARE DREAMING ABOUT? Riccardi E., Bottini R., Rota P.

121. FROM LAND TO SEA: AI VISION FOR ILLEGAL WASTE DUMPING AND FISH DETECTION Ricciardi Andrea Vincenzo

122. ADDRESSING THE WAYPOINT-ACTION GAP IN END-TO-END AUTONOMOUS DRIVING Rodríguez-Vidal J.D., Villalonga G., Porres D., López A.M.

123. STRUCTXLIP: ENHANCING VISION-LANGUAGE MODELS WITH MULTIMODAL STRUCTURAL CUES Ruan Z., Gao S., Kong Q., Wang Y., Cristani M.

124. ATTENTION-DISCOUNTED ADAPTIVE SAMPLER FOR MASKED DIFFUSION LANGUAGE MODELS Sahin Y., Saikia A. R., Cevher V., Favaro P.

125. WHEN NEGATION IS A GEOMETRY PROBLEM IN VISION-LANGUAGE MODELS Sammani Fawaz, Chamiti Tzoulio, Gavrikov Paul, Deligiannis Nikos

126. ROBUST AI FOR AUTOMATED INFRASTRUCTURE INSPECTION Sánchez A., Sampedro C., Corrochano J.

127. IMPROVING CONTROLLABLE GENERATION: FASTER TRAINING AND BETTER PERFORMANCE VIA X0-SUPERVISION Sangare A., Maglo A., Chaouch M., Luvison B.

128. DRESS-ED: INSTRUCTION-GUIDED EDITING FOR VIRTUAL TRY-ON AND VIRTUAL TRY-OFF Sanguigni F., Lobba D., Ren B., Cornia M., Sebe N., Cucchiara R.

129. VIDEO UNLEARNING VIA LOW-RANK REFUSAL VECTOR Facchiano S., Saravalle S., Migliarini M., De Matteis E., Sampieri A., Pilzer A., Rodola E., Spinelli I., Franco L. , Galasso F.

130. A NOVEL METRIC FOR DETECTING MEMORIZATION IN GENERATIVE MODELS FOR BRAIN MRI SYNTHESIS Scardace A., Puglisi L., Guarnera F., Battiato S., Ravì D.

131. SPLATXTRACT: TRACTABLE GAUSSIAN SPLATTING VIA OPEN-WORLD REGION-OF-INTEREST EXTRACTION AND REFINEMENT Schieber H., Kleinbeck C., Schoellig A. P. , Leutenegger S., Roth D.

132. ADVERSARIAL CORRECTION AND DOMAIN-ADAPTATIVE CURRICULUM (AC-DAC) FINE-TUNING Shen L., Edalati A., Li X., Meyer B., Gross W., Clark J. J.

133. VLM-GUIDED CONTACT RETARGETING AND GENERATIVE PARTNER MODELING FOR DEPLOYABLE HUMANOID-HUMAN INTERACTION Shibata Y., Amaya K., Yamazaki K., Jayanti L., Aoki Y., Isogawa M., Fragkiadaki K.

134. ADAPTIVE VS. STATIC ROBOT-TO-HUMAN HANDOVER: A STUDY ON ORIENTATION AND APPROACH DIRECTION Biagi F., Onfiani D., Silenzi S., Iani C., Biagiotti L.

135. ELASTIC VITS FROM PRETRAINED MODELS WITHOUT RETRAINING Simoncini Walter., Dorkenwald Michael., Blankevoort Tijmen., Snoek Cees GM., Asano Yuki M.

136. ANTHROSPHERE: HUMAN-GUIDED EGOCENTRIC VISION FOR ADAPTIVE AGENTIC XR ASSISTANCE IN INDUSTRIAL HUMAN-IN-THE-LOOP SYSTEMS Sirocchi C., Stacchio L., Migliorelli L., Galdelli A., Mancini A.

137. DEXOAK: OBJECT-AWARE KINEMATIC RETARGETING FOR ROBOT TRAJECTORY GENERATION Spinola F., Katzschmann R., Schmid C.

138. BRIDGING IMPLICIT NEURAL AND EXPLICIT SHAPE REPRESENTATIONS Stippel C., Engel D., Hermosilla P.,

139. MARKUSHGRAPHER-2: END-TO-END MULTIMODAL RECOGNITION OF CHEMICAL STRUCTURES Tim Strohmeyer, Lucas Morin, Gerhard Ingmar Meijer, Valéry Weber, Ahmed Nassar, Peter Staar

140. REGIONREASONER: REGION-GROUNDED MULTI-ROUND VISUAL REASONING Wenfang Sun, Hao Chen, Yingjun Du, Yefeng Zheng, Cees G. M. Snoek

141. SEEING WITH PURPOSE: VISUAL INTELLIGENCE FOR GOAL-DIRECTED ROBOT MOTION Sun B.

142. LEARNING ROBUST GEOMETRIC REPRESENTATIONS USING SYNTHETIC STRUCTURAL DEFECTS Szymanski W., Drwiega G., Wodzinski M.

143. ROGR: RELIGHTABLE 3D OBJECTS USING GENERATIVE RELIGHTING Jiapeng Tang, Matthew Levine, Dor Verbin, Stephan J. Garbin, Matthias Niessner, Ricardo Martin-Brualla, Pratul P. Srinivasan, Philipp Henzler

144. PANOPTIC SEGMENTATION FOR TIRE DEFECT DETECTION Tarassov E., Derville A., Ponchon F., Tilmant C., Chateau T.

145. LOST IN THE TIME DOMAIN: SPECTRAL ADAPTERS FOR VIDEO UNDERSTANDING Thiyakesan Ponbagavathi T., Seibold C., Roitberg A.

146. WHEN NONLOCAL VARIATIONAL MODELS MEET ATTENTION MECHANISMS: SEEING THROUGH DARKNESS AND WATER Torres D., Duran J., Navarro J., Sbert C.

147. CAN AUTOML HELP US FIND EFFICIENT FOREST BIOMASS ESTIMATION MODELS SUSTAINABLY? Traoré, K. R. and Lindauer, M.

148. HALO-FREE ALL-IN-FOCUS AND 3D IMAGING FROM FOCAL STACKS Ueda S., Saito H., Schmalstieg D., Mori S.

149. INHIBITED SELF-ATTENTION: SHARPENING FOCUS IN VISION TRANSFORMERS van der Wal, P.R.D., Strisciuglio, N., Azzopardi, G.

150. CONCEPTPOSE: TRAINING-FREE ZERO-SHOT OBJECT POSE ESTIMATION USING CONCEPT VECTORS Kuang L., Velikova Y., Saleh M., Zaech JN., Paudel D., Busam B.

151. VENI: VARIATIONAL ENCODER FOR NATURAL ILLUMINATION Walker P., Gardner J. A. D., Ardelean A., Smith W. A. P., Egger B.

152. HOW TO TRAIN A SOTA FOUNDATIONAL VLM Han Wang

153. SELF-SUPERVISED LEARNING BASED ON TRANSFORMED IMAGE RECONSTRUCTION FOR EQUIVARIANCE-COHERENT FEATURE REPRESENTATION Qin Wang, Alessio Quercia, Benjamin Bruns, Abigail Morrison, Hanno Scharr, Kai Krajsek

154. PAWS: PERCEPTION OF ARTICULATION IN THE WILD AT SCALE FROM EGOCENTRIC VIDEOS Wang Y., Miao Y., Zhao W., Yang W., Wang Z., Pajarinen J., Van Gool L., Paudel D., Kannala J., Wang X., Solin A.

155. BENCHMARKING AND BOOSTING PHYSICAL REASONING FROM CONDITIONAL VIDEO OBSERVATIONS Fanyue Wei, Kai Xu, Yizhuo Zhang, Pengzhan Sun, Junbin Xiao, Angela Yao

156. RADIANCE FIELDS FOR ROBOTICS Wilder-Smith M, Patil V, Morkva S, Bhardwaj A, Mittal M, Tateno K, Niemeyer M, Oechsle M, Tombari F, Hutter M

157. BREWING STRONGER FEATURES: DUAL-TEACHER DISTILLATION FOR MULTISPECTRAL EARTH OBSERVATION Wolf Filip, Rolih Blaž, Čehovin Zajc Luka

158. FINER: MLLMS HALLUCINATE UNDER FINE-GRAINED NEGATIVE QUERIES Xiao, Rui; Kim, Sanghwan; Xian, Yongqin; Akata, Zeynep; Alaniz, Stephan

159. VIBE SPACE: CREATIVELY CONNECTING AND EXPRESSING VISUAL CONCEPTS Xu K., Yang H., Lu A., Grossberg M.D., Bai Y., Shi J.

160. MODEL-AGNOSTIC POST-HOC PRUNING AND OPTIMIZATION FOR SINGLE-VIEW FEED-FORWARD 3D GAUSSIAN SPLATTING Yagawa R., Cheng H., Schmalstieg D., Saito H., Mori S.

161. DEXNINJA: LEARNING ROBUST DEXTEROUS CUTTING POLICY WITH A REAL-TO-SIM-TO-REAL DATA ENGINE Lou H., Yang R., Zhong W., Liu C., Liu Y., Liu W., Ma W., Xia J., Wu K., Paudel D.P., Van Gool L., Zhao H., Li Y.

162. STAR: SEAMLESS SPATIAL-TEMPORAL AWARE MOTION RETARGETING WITH PENETRATION AND CONSISTENCY CONSTRAINTS Yang X., Wang Q., Yang J., Slabaugh G., Yuan S.

163. BOOTSTRAPPING ARTICULATED 3D RECONSTRUCTION FROM IMAGES Zadrozny J., Mac Aodha O., Bilen H.

164. EGONIGHT: TOWARDS EGOCENTRIC VISION UNDERSTANDING AT NIGHT WITH A CHALLENGING BENCHMARK Deheng Zhang*, Yuqian Fu*, Runyi Yang, Yang Miao, Tianwen Qian, Xu Zheng, Guolei Sun, Ajad Chhatkuli, Xuanjing Huang, Yu-Gang Jiang, Luc Van Gool, Danda Pani Paudel

165. TOWARDS IN-THE-WILD EGOCENTRIC 3D HAND-OBJECT POSE ESTIMATION Bansal S. , Zhu Z. , Tripathi S. , Zhao J. , Black M. , Damen D.

166. HIERARCHICAL STEP DECOMPOSITION FOR TRAINING-FREE ONLINE VIDEO STEP GROUNDING Zhou L., Zanella L., Rota P., Mancini M., Ricci E.

167. EVENT-AIDED SHARP RADIANCE FIELD RECONSTRUCTION FOR FAST-FLYING DRONES Zou R., Cannici M., Scaramuzza D.

168. VGGSOUNDER: AUDIO-VISUAL EVALUATIONS FOR FOUNDATION MODELS Daniil Zverev, Thaddäus Wiedemer, Ameya Prabhu, Matthias Bethge, Wieland Brendel, A. Sophia Koepke

DEEP PROBABILISTIC SUPERVISION FOR IMAGE CLASSIFICATION

Adelöw A., Gamba M., Maki A.

Abstract: We propose Deep Probabilistic Supervision (DPS), a principled learning framework that constructs sample-specific target distributions via approximate Bayesian inference on the model's own predictions and remains independent of hard targets after initialization. Extensive evaluations demonstrate that DPS significantly improves generalization, calibration, and robustness against label noise across standard benchmarks.

Contact: aadelow@kth.se

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 1

SYNTHFORENSICS: BENCHMARKING AND EVALUATING PEOPLE-CENTRIC SYNTHETIC VIDEO DEEPFAKES

Leotta R., Sambataro S. A., Ragaglia C. V., Casu M., Petralia Y., Guarnera F., Guarnera L., Battiato S.

Abstract: SynthForensics is a people-centric deepfake benchmark of 20,445 synthetic videos from 8 text-to-video and 7 image-to-video open-source generators, paired-source from FF++/DFD reals, two-stage human-validated, in four compression versions. In a human study, raters prefer it in 71–77% of comparisons against nine existing benchmarks. Across 15 detectors, face-based methods drop on average 27 AUC points from FF++ to SynthForensics, exposing a critical gap in current detection.

Contact: salvatore.sambataro@phd.unict.it

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 2

NLS: NOVEL LATENT SYNTHESIS

Anadón X., Batlle V., Mur-Labadia L., Montiel J. M.

Abstract: Novel View Synthesis (NVS) predicts photorealistic images, in contrast we propose Novel Latent Synthesis, in which we directly predict a latent representation used in the downstream task (depth, segmentation). Because of bypassing photorealistic prediction it is more efficient and easier to learn. Our 27M-param predictor (100× smaller than the RGB baseline[1]) trains in 3 days on one V100, runs 4× faster, and performs on par with the NVS baseline on depth estimation and semantic segmentation.

Contact: xanadon@unizar.es

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 3

VISIBLE OBJECT-STATE PROXY GEOMETRY FOR POLICY LEARNING

Antypas I. , Averta G. , Garcia N.

Abstract: Visible Object-State Proxy Geometry (VOSP-G) is a compact representation of object-centric effects in unconstrained interaction videos. Rather than relying on privileged 6D pose or semantic labels, VOSP-G uses tracked object masks and visible 3D geometry. Centroid trajectories capture translational dynamics, while principal-component axes and spatial extents provide orientation-sensitive cues, describing where an object is, how it is oriented, and how it changes during interaction.

Contact: jordanantypas@gmail.com

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 4

SCENETOK: A COMPRESSED, DIFFUSABLE TOKEN SPACE FOR 3D SCENES

Asim M., Wewer C., Lenssen J.

Abstract: We propose a latent scene representation from which novel views can be rendered, allowing efficient generation in the latent token space and decoupling view rendering from generation.

Contact: masim@mpi-inf.mpg.de

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 5

GEOMETRY ESTIMATION USING DENSE CORRESPONDENCES

Astermark J., Heyden A., Larsson V.

Abstract: Dense matching has recently become the gold-standard for image matching, with downstream estimation often outperforming both sparse and feed-forward methods [1]. However, using dense matches in existing geometry pipelines leads to redundancies and multi-view inconsistencies. Recent methods address this by refining quantized tracks [2, 3] or sampling in a star-topology [4]. We instead address the first issue by summarizing geometric constraints, and the second by multi-warp aware track sampling.

Contact: jonathan.astermark@math.lth.se

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 6

ROOM ENVELOPES: A SYNTHETIC DATASET FOR INDOOR LAYOUT RECONSTRUCTION FROM IMAGES

Bahrami Sam., Campbell Dylan.

Abstract: Reconstruction recovers visible surfaces but misses occluded structure—walls, floors, ceilings. As these are planar and simple, they should be cheap to predict. We present Room Envelopes, a synthetic dataset pairing each RGB image with two pointmaps: the visible surface and the structural layout (the first surface with fittings and fixtures removed). This gives direct supervision for feed-forward monocular estimators predicting both surfaces, capturing scene extent and object placement.

Contact: sam.bahrami@anu.edu.au

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 7

ASSEMBLYHANDS-X: MODELING HAND, BODY, AND VISUAL CONTEXT FOR UNDERSTANDING BIMANUAL HUMAN ACTIVITIES

Banno T., Suzuki N., Ohkawa T., Liu R., Kwon T., He K., Shinoda R., Furuta R., Sato Y.

Abstract: We introduce AssemblyHands-X, the first markerless 3D hand-body benchmark for understanding bimanual human activities, built with an annotation pipeline that robustly reconstructs temporally consistent SMPL-X hand-body motion from multi-view RGB videos. Our action recognition benchmark shows that jointly modeling hand, body, and visual cues consistently outperforms using each modality in isolation, highlighting the importance of multimodal synergy for understanding bimanual activities.

Contact: banno@iis.u-tokyo.ac.jp

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 8

PANO3D: UNIFIED 3D RECONSTRUCTION AND PANOPTIC SEGMENTATION

Victor Barberteguy, Ahmet Iscen, Mathilde Caron, Alireza Fathi, Gül Varol, Cordelia Schmid

Abstract: Recent advances in 3D feedforward reconstruction neural networks have achieved remarkable success in dense reconstruction from images without any camera parameters. Yet, equipping these models with robust semantic understanding remains an open problem. Here we introduce an approach that performs 3D reconstruction and 3D panoptic segmentation in a unified framework. We build on existing 3D reconstruction models and augment them with a set-based mask decoder. The approach is jointly trained with a geometric and semantic loss, which are shown to be mutually beneficial. More precisely, the features are initialized from the geometric information and then finetuned to capture jointly geometry and semantics. We demonstrate the generality of our approach by successfully applying our framework both to online and all-to-all attention reconstruction backbones. Our method achieves state-of-the-art performance in 3D panoptic segmentation across ScanNet, ScanNet200, and ScanNet++ datasets. Ablation studies show that such joint training of a unified model equips 3D feedforward reconstruction neural networks with panoptic segmentation and yields mutually beneficial improvements.

Contact: victorbbt@google.com

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 9

A PAN-EUROPEAN MULTI -SEASONAL LAND COVER MAPPING MODEL

BARCO L., GALATOLA M., ARNAUDO E., BRAGAGNOLO A., GARZA P., ROSSI C.

Abstract: Seasonal distribution shifts are a critical obstacle to land cover mapping. GEOID-Land is a large-scale multi-seasonal Sentinel-2 dataset covering 749 European cities across diverse climatic zones, annotated with 10 land cover classes derived from Urban Atlas 2018 and refined through Copernicus HRL products. Convolutional and transformer-based semantic segmentation networks (i.e., SegFormer (MiT-B5) and UPerNet (ConvNeXt-L)) are benchmarked across single-season, multi-season, and season-aware training regimes. Results indicate that seasonal robustness stems more from data diversity than architectural complexity: random seasonal exposure during training matches or surpasses dedicated season-aware mechanisms, and test-time aggregation of predictions across all four seasonal views yields additional gains.

Contact: luca.barco@polito.it

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 10

COSY: COMPOSITIONAL 3DGS SYNTHESIS FOR DISENTANGLED HUMAN HEAD EDITING

Barthel F., De Mello S., Nagano K., Morgenstern W., Hilsmann A., Eisert P.

Abstract: Generative models are inherently entangled, as they aim to copy the training data distribution. This makes it almost impossible to synthesize out of distribution samples like “a masculine face with long hair”. To solve this we split the generator into multiple sub-generators for hair, face, torso, and glasses. This allows us to edit one component without changing the others.

Contact: florian.tim.barthel@hhi.fraunhofer.de

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 11

ON THE GENERALIZATION OF OPTICAL FLOW: QUANTIFYING ROBUSTNESS TO DATASET SHIFTS

Katrin Bauer, Andrés Bruhn, Jenny Schmalfuss

Abstract: Background. The generalization of learning-based methods that estimate the optical flow is generally evaluated by their accuracy on unseen datasets.

Contribution. We demonstrate a linear correlation between the prediction accuracy on in-distribution (ID) and out-of-distribution (OOD) data. This allows us to define effective robustness for optical flow, a new metric measuring OOD robustness separately from accuracy enabling new insights into optical flow architectures.

Contact: katrin.bauer@vis.uni-stuttgart.de

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 12

MACES-APO: MULTI-AGENT CO-EVOLUTIONARY SIMULATION FOR AUTOMATIC PROMPT OPTIMIZATION

Bayoumi O., Cinque L., Foresti G.F.

Abstract: MACES-APO views automatic prompt optimization [1,2,3,4] as a multi-agent co-evolutionary simulation. From a one-sentence goal, a society of LLM agents [5] interacts over 200 cycles, building memories and heuristics, while a macro-evolutionary layer spawns, merges, and prunes them. Skill is scored by an executable engine: objectively from the game state where possible, or from agents' graded interactions. Finally, the top-skill agent's heuristics are distilled into a deployable system prompt.

Contact: bayoumi@di.uniroma1.it

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 13

HIERARCHICAL OBJECT-CENTRIC REPRESENTATION LEARNING

Behrad F., Tuytelaars T., Wagemans J.

Abstract: Visual perception in embodied agents can benefit from structured, hierarchical object-centric representations. Yet current models typically represent objects at a single level, without modelling hierarchical part-whole relationships common in real-world scenes. We propose a hierarchical slot attention model that explicitly encodes part-whole structure, enabling richer unsupervised scene decomposition and laying the groundwork for embodied systems that perceive the physical world closer to how humans do.

Contact: fatemeh.behrad@kuleuven.be

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 14

SLAD : SHARED LORA ADAPTERS FOR TASK-SPECIFIC DISTILLATION

Bensaid Reda., Bendou Yassir., Gripon Vincent., Leduc-Primeau François.

Abstract: SLAD introduces a task-specific distillation framework that improves knowledge transfer from large foundation models to smaller ones. The method uses shared LoRA adapters between teacher and student models to preserve feature alignment during adaptation, reducing representation mismatch caused by fine-tuning. Through joint training and parameter sharing, SLAD achieves state-of-the-art accuracy on classification and segmentation tasks while training up to 2× faster than conventional fine-tuning approaches.

Contact: reda.bensaid@imt-atlantique.fr

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 15

LINEAR MODEL MERGING UNLOCKS SIMPLE AND SCALABLE MULTIMODAL DATA MIXTURE OPTIMIZATION

Berasi D., Farina M., Mancini M., Ricci E.

Abstract: Selecting an optimal data mixture is critical for fine-tuning Multimodal Large Language Models, but doing so is computationally expensive due to the large search space and high training costs. We demonstrate that model merging can efficiently score and rank data mixtures, as merged models strongly correlate with mixture-trained models. This enables scalable and efficient mixture optimization without the need for repeated expensive training runs.

Contact: davide.berasi@unitn.it

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 16

ROBUST RECOGNITION OF CARDIAC PATHOLOGIES BASED ON PHONOCARDIOGRAM ANALYSIS AND DEEP LEARNING

Beritelli L., Avanzato R., Guarnera L., Battiato S., Beritelli F.

Abstract: Cardiovascular diseases (CVDs) remain a leading cause of global mortality, underscoring the critical need for reliable and accessible diagnostic tools like Phonocardiogram (PCG). This paper proposes a deep learning framework for the classification of seven distinct cardiac conditions from PCG signals. We present a systematic benchmark comparing a One-Dimensional Convolutional Neural Network (1D-CNN), on raw waveforms, and several Two-Dimensional (2D) architectures, on spectrograms, evaluated under both random and, more critically, clinically realistic subject-independent splitting protocols. Our results highlight a key trade-off for practical deployment: while 2D spectrogram-based models yield the best absolute performance on clean data, the simpler 1D-CNN shows remarkable stability against respiratory noise across varying Signal-to-Noise Ratio (SNR) levels. This is confirmed by its ability to maintain a robust F1-score of 83.8% even at 5 dB SNR. This superior stability makes the 1D-CNN an attractive and reliable choice for resource-constrained or monitoring scenarios where noise is prevalent. The 7-class dataset is available at [link will be published upon acceptance].

Contact: ludovica.beritelli@phd.unict.it

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 17

ONE TARGET TO ALIGN THEM ALL: LIDAR, RGB AND EVENT CAMERAS EXTRINSIC CALIBRATION FOR AUTONOMOUS DRIVING

Bertogalli, Boracchi, Magri

Abstract: We propose a one-shot extrinsic calibration framework for event cameras, LiDARs, and RGB cameras, designed to jointly estimate their relative poses. The method relies on a novel 3D calibration target that combines planar geometric structures, ChArUco markers, and active LED patterns, enabling reliable feature extraction across heterogeneous sensors. The framework is validated on a custom real-world autonomous driving dataset, demonstrating accurate and robust multi-modal alignment.

Contact: andrea.bertogalli@unipr.it

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 18

ROBOTIC POLICY ADAPTATION VIA WEIGHT-SPACE META-LEARNING.

Christian Bianchi, Siamak Yousefi, Alessio Sampieri, Andrea Roberti, Luca Rigazio, Fabio Galasso, Luca Franco

Abstract: WIZARD eliminates costly fine-tuning for Vision-Language-Action models by instantly generating task-specific LoRA weights from just a text instruction and a video. This single-pass approach achieves up to a 14x improvement on unseen tasks, enabling true zero-shot, real-world deployment.

Contact: ch.bianchi02@gmail.com

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 19

DOPO: DENSE ONLINE PREFERENCE OPTIMIZATION FOR CROSS-DATASET MOTION DIFFUSION ADAPTATION

Macaluso G., Mandelli L., Bicchierai M., Berretti S., Bagdanov A.D.

Abstract: Adapting text-to-motion models to new domains typically requires expensive motion capture data. We propose DOPO (Dense Online Preference Optimization), a framework that fine-tunes pretrained motion diffusion models using only target textual prompts, eliminating ground-truth data needs, cutting training time by 10 times, and outperforming RL baselines across BABEL, HumanML3D, and MotionX.

Contact: mirko.bicchierai@unifi.it

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 20

REWIS3D: RECONSTRUCTION IMPROVES WEAKLY-SUPERVISED SEMANTIC SEGMENTATION

Jonas Ernst*, Wolfgang Boettcher*, Lukas Hoyer, Jan Eric Lenssen, Bernt Schiele

Abstract: Rewis3d is a framework that uses feed-forward 3D reconstruction to improve weakly supervised semantic segmentation on 2D images. Since dense pixel-level annotations are costly, sparse annotations offer an efficient alternative but leave a performance gap. We close it by using 3D scene reconstruction as auxiliary supervision: geometric structure recovered from 2D videos propagates sparse labels across scenes via a dual student-teacher architecture enforcing 2D–3D semantic consistency. Rewis3d achieves state-of-the-art sparse-supervision performance, outperforming prior work by 2–7% with no extra inference cost.

Contact: wolfgang.boettcher@mpi-inf.mpg.de

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 21

EDGES OF PHYSICAL INTELLIGENCE

Bonazzi P., Magno M.

Abstract: Real-world physical intelligence demands ultra-low latency, energy efficiency, and robustness. We illustrate key principles through separate contributions based on (1) event cameras, (2) in-sensor computing, (3) low-latency FPGA interfaces, and (4) a fully ternary Vision Transformer on a microcontroller.

Contact: pbonazzi@ethz.ch

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 22

TINY INFERENCE TIME SCALING WITH LATENT VERIFIERS

Bucciarelli D, Turri E, Baraldi L, Cornia M, Cucchiara R

Abstract: Inference-time scaling improves image generation by allocating more computation at test time: generating multiple candidate images and using a verifier to select the sample that best matches the prompt. VHS makes this more efficient by making verification latent.

Contact: davide.bucciarelli@unimore.it

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 23

GAP: GEOMETRIC ANCHOR PRE-TRAINING FOR DATA-EFFICIENT VISUOMOTOR LEARNING OF MANIPULATION TASKS

Buoso D., Protopapa A., Di Carlo S., Pistilli F., Averta G

Abstract: Geometric Anchor Pre-training (GAP) is a simple, action-free warm-up stage that regularizes a spatial adapter to extract robust, object-centric keypoints from dense visual tokens. By translating high-dimensional vision features into precise geometric coordinates, GAP bridges generalist Vision Foundation Models and policy learning, drastically simplifying imitation learning.

Contact: davide.buoso@polito.it

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 24

AUTOMATED PREDICTION OF PARAVALVULAR REGURGITATION BEFORE TRANSCATHETER AORTIC VALVE IMPLANTATION

Michele Cannito, Riccardo Renzulli, Marco Grangetto, Fabrizio D'Ascenzo

Abstract: Post-TAVI paravalvular regurgitation remains challenging to predict from preoperative CT, as conventional workflows rely on manual measurements and may miss 3D textural patterns. This work proposes an automated 3D CNN pipeline using cardiac CT volumes centered on the aortic valve. By combining full volumetric context with domain-specific pretraining, the model aims to learn clinically relevant features for preprocedural PVR risk stratification.

Contact: michele.cannito@unito.it

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 25

SHIELDCLIP: SELECTIVE SAFETY ALIGNMENT FOR HARMFUL CONTENT MITIGATION IN MULTIMODAL FOUNDATION MODELS

Poppi T., Cappelletti S., Poppi S., Cornia M., Baraldi L., Cucchiara R.

Abstract: Multimodal foundation models can generate harmful content due to unsafe associations learned from large-scale web data. Current safety alignment methods have several limitations: global alignment strategies distort benign regions of the embedding space; safety is modality-dependent, but most methods assume symmetric text–image harmfulness; generated samples are often treated as unsafe by default, leading to over-sanitization. We therefore need selective safety alignment that preserves safe content while redirecting only harmful inputs.

Contact: silvia.cappelletti@unimore.it

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 26

PHYSFORMER: LEARNING TO SIMULATE MECHANICS IN WORLD SPACE

Chen Y., Lan Y., Vedaldi, A.

Abstract: A physics-grounded diffusion transformer for 4D multi-object, multi-material mesh dynamics generation in world coordinates.

Contact: yiming@robots.ox.ac.uk

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 27

DIFFUSING DEBIAS (DDB): SYNTHETIC BIAS AMPLIFICATION FOR MODEL DEBIASING

Ciranni M., Pastore V.P., Di Via R., Tartaglione E., Odone F., Murino V.

Abstract: Spurious correlations between attributes and target labels cause DL classifiers to learn biased representations, leading to weak generalization. We introduce Diffusing DeBias (DDB), a plug-in for unsupervised debiasing methods that uses conditional diffusion models to generate synthetic bias-aligned data for training a stronger bias amplifier, which drives the debiasing process. By avoiding conflicting sample memorization on the original dataset, DDB achieves SotA across multiple benchmarks.

Contact: massimiliano.ciranni@edu.unige.it

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 28

WATCH, LEARN, ASSIST: ENABLING BIMANUAL MANIPULATION WITH THE HANNES PROSTHESIS

Columbaro M., Vasile F., Boccardo N., Natale L.

Abstract: Bimanual manipulation is challenging for individuals with upper-limb loss, especially when coordinating a prosthesis with the healthy hand. We present a vision-based imitation-learning framework for the Hannes prosthetic hand that maps egocentric and eye-in-hand observations to motor commands for collaborative two-handed tasks. Using multiview demonstrations of hand coordination, our policy enables successful real-world bimanual manipulation, providing preliminary evidence of feasibility.

Contact: martina.columbaro@iit.it

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 29

FLAME: A DEPLOYABLE FRAMEWORK FOR LAYER-WISE AUTOMATIC MIXED-PRECISION QUANTIZATION ON EMBEDDED DEVICES

Corti G., Vitali M., Vacis N., Palladino V., Merigo L., Pidò S., Matteucci M.

Abstract: FLAME is a training-free ONNX-based framework for automatic mixed-precision quantization on edge devices. It ranks layer sensitivity from FP32/INT8 activation shifts and uses a genetic search to trade accuracy, size, and CPU inference time. Validated on YOLOv8n-Pose, MobileNetV2, and DeepLabV3 with deployment on STM32N6 and GAP9, FLAME recovers much of the accuracy lost by uniform INT8 while preserving model compression.

Contact: greta.corti@polimi.it

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 30

LEARNING GENERALIZABLE DYNAMICS MODELS WITH GRAPH NEURAL NETWORKS FOR NOVEL TOOL DESIGN

Cugito N., Allen K.

Abstract: Model-based planning needs accurate dynamics models. For deformable object manipulation, analytic methods lack accuracy while learned models hallucinate outside training data. We learn dynamics from real-world RGB-D data using graph neural networks across varied tool-object interactions. Training on a small tool set generalizes better than single-tool training and predicts unseen tool interactions. We leverage model differentiability to optimize tool geometry, designing novel tools for downstream tasks.

Contact: ncugito@ethz.ch

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 31

CONVMAMBA: A HYBRID CNN-MAMBA BACKBONE FOR ANIMAL POSE ESTIMATION

Daadouch S., Zhao K., Gelautz M., Roth P.M

Abstract: While Convolutional Neural Networks provide strong inductive bias for local features, they still struggle with modeling long-range dependencies. More recently, there has been increasing interest in using Mamba due to its capacity to model global context. To unite both strengths, we propose ConvMamba, a hierarchical CNN-Mamba hybrid, and evaluate it on ImageNet-1K classification and animal pose estimation (AP-10K).

Contact: Salma.Daadouch@vetmeduni.ac.at

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 32

MITIGATING THE MODALITY GAP IN TEXT-DRIVEN SEMANTIC SEGMENTATION

D'ASARO F., BOTTINO A., RIZZO G.

Abstract: VLP-based segmentation models suffer from misalignment between pixel and text features, exhibiting the Modality Gap, where image and text features form separate clusters in the shared embedding space. We propose a Mask-Text Contrastive (MTC) module that aligns image regions with their concepts via an InfoNCE loss. Lightweight and plug-and-play, MTC consistently improves segmentation on ADE20K and COCO-Stuff 10k.

Contact: federico.dasaro@polito.it

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 33

CAMC2V: CONTEXT-AWARE CONTROLLABLE VIDEO GENERATION

Luis Denninger, Sina Mokhtarzadeh Azar, Juergen Gall

Abstract: Recently, image-to-video (I2V) diffusion models have demonstrated impressive scene understanding and generative quality, incorporating image conditions to guide generation. However, these models primarily animate static images without extending beyond their provided context. Introducing additional constraints, such as camera trajectories, can enhance diversity but often degrade visual quality, limiting their applicability for tasks requiring faithful scene representation. We propose CamC2V, a context-to-video (C2V) model that integrates multiple image conditions as context with 3D constraints alongside camera control to enrich both global semantics and fine-grained visual details. This enables more coherent and context-aware video generation. Moreover, we motivate the necessity of temporal awareness for an effective context representation. Our comprehensive study on the RealEstate10K dataset demonstrates a (FVD) improvement in visual quality and camera controllability.

Contact: luis.denninger@tum.de

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 34

EXECUTION-AWARE VLA FOR ROBUST MANIPULATION

Dey Sombit, Albanese Giuliano, Zaech Jan-Nico, Van Gool Luc , Paudel Danda

Abstract: Vision-Language-Action (VLA) models often assume that commanded actions are executed faithfully, overlooking real-world execution gaps caused by latency, tracking error, actuator limits, and robot-specific dynamics. We introduce EA-VLA, a framework that incorporates execution-state feedback during policy inference, enabling adaptation to varying execution dynamics. Trained with simulation-based domain randomization, EA-VLA learns to compensate for execution discrepancies and significantly improves robustness in both simulation and real-world evaluations.

Contact: sombit.dey@insait.ai

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 35

MAKING FEW-SHOT SEGMENTATION ACTUALLY WORK

De Marinis P., Vessio G., Castellano G.

Abstract: Few-shot segmentation (FSS) segments novel classes from only a handful of labeled examples. We address four open challenges: LabelAnything unifies diverse visual prompts for flexible N-way K-shot inference; TaP boosts any FSS model via support-conditioned LoRA adaptation; DistillFSS enables efficient cross-domain inference with no support images at test time; AffEx provides the first interpretability method for FSS, extracting structured attribution maps from support images.

Contact: pasquale.demarinis@uniba.it

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 36

GOOD, FAST, AND CHEAP: VISUAL-INERTIAL TRACKING FOR LOW LATENCY SPATIAL AGENTS

de Mayo M., Pire T., Cremers D.

Abstract: Visual inertial (SLAM) tracking is a fundamental componentfor any computing device needing to understand its surroundings. Originally motivated by the application of VR, we showcase how what should be, in theory, a straightforward and solved problem according to current literature, presents major gaps in deployability, performance, and accuracy, even on simple scenarios. We propose a project that encompasses an in-development state-of-the-art system together with additional extensions.

Contact: mateo.demayo@tum.de

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 37

ORTHOTRACK: CONTINUOUS 6-DOF UAV TRAJECTORY ESTIMATION ANCHORED IN PUBLIC ORTHOPHOTOS

Oussema Dhaouadi, Zuria Bauer, Johannes Michael Meier, Olaf Wysocki, Marc Pollefeys, Daniel Cremers

Abstract: Continuous 6-DoF pose estimation is essential for autonomous UAV operations. Yet, existing visual odometry and SLAM methods accumulate drift and yield only relative, up-to-scale trajectories. Single-frame geo-localization, in turn, discards temporal continuity and remains too slow for real-time use. We present OrthoTrack, a training-free system that estimates continuous 6-DoF UAV trajectories using only publicly available orthophotos and surface models as a map prior. OrthoTrack matches keyframes against the orthophoto and lifts correspondences to metric 3D via the surface model. It then propagates these map-anchored correspondences to intermediate frames with optical flow, producing absolute, metrically scaled poses at every frame without GPS or post-hoc alignment. We also introduce the MovingDrone Dataset, a large-scale benchmark pairing photorealistic UAV sequences with dense 6-DoF ground truth and co-registered multi-modal geodata including multi-temporal orthophotos. On MovingDrone and real-world benchmarks, OrthoTrack runs in real time on a single GPU. It outperforms all baselines by a large margin, even those receiving oracle scale and alignment. By relying on publicly available geodata, OrthoTrack enables deployment to new regions without site-specific adaptation. Code and data will be released upon acceptance.

Contact: oussema.dhaouadi@tum.de

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 38

HAC: PARAMETER-EFFICIENT HYPERBOLIC ADAPTATION OF CLIP FOR ZERO-SHOT VQA

Dibitonto N1., Beyan N2., Murino N3.

Abstract: Hyperbolic representations are more expressive than Euclidean ones for VLMs, better capturing hierarchical and relational structures. However, Hyperbolic CLIP models are expensive, as they require training from scratch. What if pretrained CLIP could transition into a better geometry without full retraining? We present HAC, a parameter-efficient framework that lifts pretrained CLIP into hyperbolic space via lightweight fine-tuning, improving zero-shot VQA by up to +1.9 points on reasoning tasks.

Contact: francesco.dibitonto@univr.it

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 39

ENHANCING OUT-OF-DISTRIBUTION DETECTION WITH EXTENDED LOGIT NORMALIZATION

Ding Y., Liu X., Unger J., Eilertsen G.

Abstract: Out-of-distribution (OOD) detection is crucial for reliable machine learning systems. In this work, we identify feature collapse in Logit Normalization (LogitNorm) and propose ELogitNorm, a hyperparameter-free extension with feature distance awareness. Our method improves OOD detection and in-distribution confidence calibration while preserving classification accuracy, outperforming existing training-time approaches on standard benchmarks.

Contact: yd402@cam.ac.uk

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 40

GENERATIVE EDITING THROUGH FEW-SHOT DIFFUSION ADAPTATION

Elezabi O., Zamfir E., Wu Z., Timfote R.

Abstract: Text-guided diffusion models have advanced image editing by enabling intuitive control through language. However, despite their strong capabilities, we surprisingly find that SOTA methods struggle with simple, everyday transformations such as rain or blur. We attribute this limitation to weak and inconsistent textual supervision during training, which leads to poor alignment between language and vision. Existing solutions often rely on extra finetuning or stronger text conditioning, but suffer from high data and computational requirements. We argue that diffusion-based editing capabilities aren't lost but merely hidden from text. The door to cost-efficient visual editing remains open, and the key lies in a vision-centric paradigm that perceives and reasons about visual change as humans do, beyond words. Inspired by this, we introduce Visual Diffusion Conditioning (VDC), a training-free framework that learns conditioning signals directly from visual examples for precise, language-free image editing. Given a paired example -one image with and one without the target effect- VDC derives a visual condition that captures the transformation and steers generation through a novel condition-steering mechanism. An accompanying inversion-correction step mitigates reconstruction errors during DDIM inversion, preserving fine detail and realism. Additionally, we propose Optimal Path Transition (OPT), a framework that adapts the diffusion trajectory by constructing a linear transition between a point on the input trajectory and the desired output distribution. The adaptation is achieved through LoRA fine-tuning of the attention heads to apply the translation during sampling. This framework provides more general editing capabilities while maintaining the adaptation efficiency.

Contact: omar.elezabi@uni-wuerzburg.de

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 41

WHEN CAN GENERATED VIDEO BECOME RELIABLE 3D EVIDENCE?

Fan C., Favaro P.

Abstract: Starting from OrbitForge, this poster traces a research path from reconstructing text-to-video clips into closed-orbit 3D Gaussian Splatting scenes, to benchmarking when generated videos behave like usable 3D acquisitions, to diagnosing a field-of-view range tail in feed-forward geometry cameras against matched real-video controls. The resulting story asks when generated video can become reliable 3D evidence.

Contact: chenrui.fan@unibe.ch

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 42

ONLINE VIDEO DEPTH ANYTHING: TEMPORALLY-CONSISTENT DEPTH PREDICTION WITH LOW MEMORY CONSUMPTION

Feiden J., Küchler T., Zavadski D., Savchynskyy B., Rother C.

Abstract: Recently, Video Depth Anything (VDA) has demonstrated strong performance on long video sequences. However, it relies on batch-processing which prohibits its use in an online setting. In this work, we overcome this limitation and introduce online VDA (oVDA). The key innovation is to employ techniques from Large Language Models (LLMs), namely, caching latent features during inference and masking frames at training. We demonstrate that oVDA runs at 42 FPS on an NVIDIA A100 and at 20 FPS on an NVIDIA Jetson edge device.

Contact: johann-friedrich.feiden@iwr.uni-heidelberg.de

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 43

T-FUNS3D: TASK-DRIVEN HIERARCHICAL OPEN-VOCABULARY 3D FUNCTIONALITY SEGMENTATION

Feng J., Sabzevari R.

Abstract: T-FunS3D is a task-driven hierarchical framework for open-vocabulary 3D functionality segmentation, enabling robots to localize functional object parts in indoor scenes. By building an open-vocabulary scene graph from 3D point clouds and RGB-D images, it identifies task-relevant objects and their functional components using pre-trained vision-language models. On SceneFun3D [1], T-FunS3D achieves competitive accuracy while reducing runtime and memory usage.

Contact: jingkun.feng@tudelft.nl

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 44

TOWARDS 4D ENVIRONMENT RECONSTRUCTION OF GLACIERS FROM SPARSE AND NOISY DATA

Foggin A., Smith W.

Abstract: Monitoring of glaciers requires modelling of surface geometry, its changes both gradual and sudden, and understanding of constantly varying lighting conditions. This work is building towards a 4D Gaussian Splatting scene representation with a neural flow field, explicit modelling of surface changes over time, and a vision transformer to handle lighting changes. This unified model enables linking of independent observations from drone flights and time-lapse cameras.

Contact: alistair.foggin@york.ac.uk

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 45

SCENARIOCONTROL: VISION-LANGUAGE CONTROLLABLE VECTORIZED LATENT SCENARIO GENERATION

Gao L., Xu Y., Koch W., Ruffino S., Rowe L., Chalaki B., Rivkin D., Ost J., Girgis R., Bijelic M., Heide F.

Abstract: We introduce ScenarioControl, the first vision-language control mechanism for learned driving scenario generation. It transforms text prompts or images into realistic 3D scenarios with road structure, agent placement, and traffic conditions, and. A cross-global control mechanism enables fine-grained scene control while preserving realism and temporal consistency.

Contact: gaolilli@outlook.com

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 46

EZ-SP: FAST AND LIGHTWEIGHT SUPERPOINT-BASED 3D SEGMENTATION

Geist L., Landrieu L., Robert D.

Abstract: Superpoint methods provide an efficient alternative to point- or voxel-based 3D semantic segmentation, but are often bottlenecked by their CPU-bound partition step. We propose a learnable, fully GPU partitioning algorithm, achieving 13× faster partitioning than prior methods. Combined with a lightweight superpoint classifier, the full pipeline scales to multi-million-point scenes while matching point-based SOTA accuracy with 72× faster inference and 120× fewer parameters.

Contact: louis.geist@enpc.fr

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 47

SUPERHUMAN SAFE AND AGILE RACING THROUGH MULTI-AGENT REINFORCEMENT LEARNING

Geles I., Bauersfeld L., Wulfmeier M., Scaramuzza D.

Abstract: Motivation Autonomous systems achieve superhuman performance in isolation, but suffer catastrophic collisions when multiple agents share dynamic environments.

Contributions We present superhuman safe multi-player quadrotor racing through league-based self-play. Our agents develop anticipatory collision avoidance and overtaking, outperforming a champion-level human pilot at speeds up to 22 m/s and accelerations of 7 g while reducing collisions by over 50%.

Contact: geles@ifi.uzh.ch

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 48

HUMANMOVEVQA: CAN VIDEO MLLMS REASON ABOUT HUMAN MOVEMENT IN VIDEOS?

Pulkit Gera, Faegheh Sardari, Asmar Nadeem, Valentina Bono, Padraig Boulton, Adrian Hilton, Armin Mustafa

Abstract: A caption like " a p e r s o n p la y i n g t e n n i s " collapses complex motion into a coarse label — a rapid lateral sprint to recover a ball is fundamentally different from a slow approach to the net, yet both share the same tag. The foundational components of physical action — where a person moves, how their trajectory evolves, and how their orientation changes over time — are largely absent from video- language data, bottlenecking MLLMs for sports analytics, autonomous navigation & safety. We introduce HumanMoveVQA, the first benchmark for global human trajectory & orientation reasoning from an exocentric view, posing questions in a first-frame-anchored 3D world coordinate system. Benchmark: 10,203 QA pairs over 7 reasoning categories in a first-frame-anchored 3D world coordinate system. Pipeline: lifts 2D video to world-consistent 3D motion tracks & deterministically generates distractor-controlled QA. Finding: SOTA MLLMs sit near chance; SFT on our data triples the average score (11.7 $\rightarrow$ 43.0).

Contact: pg00807@surrey.ac.uk

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 49

UNIFIED LIDAR PSEUDO-LABELING FOR ULTRA LONG RANGE TRUCKING PERCEPTION

Ghilotti F., Brucker S., Palladin E., Saidy N., Sigal A., Matteucci M., Bijelic M., Heide F.

Abstract: Safe highway autonomy for heavy trucks requires scene under-standing of hundreds of meters for anticipatory planning due to higher operating velocities and extended braking distances. Existing datasets focus on urban scenes with perception ranges limited only up to 100 meters. Manual 3D labeling becomes increasingly expensive with longer ranges and more actors in the scene and does not scale with required data amounts for generative AI. We need scalable, geometry grounded pseudo labels that work at ultra long ranges.

Contact: filippoghilotti0@gmail.com

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 50

INFINITE-STORY: A TRAINING-FREE CONSISTENT TEXT-TO-IMAGE GENERATION

Park J.*, Lee K.*, Gim J.*, Jo H., Oh M., Choi W., Hwang K., Kim J., Choi M., and Im S.

Abstract: We present Infinite-Story, a training-free framework for consistent T2I generation in multi-prompt storytelling. To address identity/style inconsistency, we propose Identity Prompt Replacement and attention guidance combining Adaptive Style Injection and Synchronized Guidance Adaptation on a scale-wise autoregressive model. Experiments show state-of-the-art results and over 6x faster inference than the previous method, achieving 1.72 seconds per image, proving practical for visual storytelling.

Contact: jongmin4422@dgist.ac.kr

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 51

CONCEPT DYNAMICS IN DIFFUSION MODELS: FROM TEMPORAL FORMATION TO COMPOSITIONAL BINDING

Görgu"n A., Schiele B. , Fischer J.

Abstract: Modern diffusion models produce striking images yet remain unreliable on compositional prompts with several objects and attributes. We study concept dynamics in these models through targeted interventions. We first ask when a concept forms and stabilizes: our method PCI inserts a concept at a chosen step and tests whether it persists. We then ask why concepts fail to combine, tracing the mechanisms in joint attention that break attribute binding.

Contact: agoerguen@mpi-inf.mpg.de

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 52

ARTICULATE-3D: DATASET & METHOD FOR INTERACTION UNDERSTANDING

HALACHEVA A.-M., MIAO Y., ZAECH J.-N., WANG X., VAN GOOL L., PAUDEL D.

Abstract: We address interaction understanding in 3D indoor environments through two key contributions: (1) Articulate3D, an expert-curated dataset with object- and part-level semantic segmentations, connectivity graphs, and detailed articulation specifications, including explicitly defined graspable regions; and (2) a unified framework for joint part segmentation and articulation prediction. We demonstrate applications in robotic interaction simulation and LLM-based scene editing.

Contact: anna-maria.halacheva@insait.ai

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 53

PUZZLE SIMILARITY: A CROSS-REFERENCE METRIC FOR ARTIFACT DETECTION IN UNSEEN VIEWS HOUSE FLOW: RECONSTRUCTING MULTIFLOOR HOUSE LAYOUTS

Hermann N., Condor J., Didyk P., Engelmann F.

Abstract: In our recent work, we introduce a cross-reference metric to spatially assess 3D reconstruction quality in unobserved views via patch-wise perceptual similarity to the training views, outperforming direct-reference metrics in human-alignment experiments and enabling recursive in-painting for artifact removal. Our current work, HouseFlow, scales editable scene generation to 100k multifloor buildings, learning flow-matched corner fields from point clouds as a scaffold for layout graphs.

Contact: nicolai.hermann@usi.ch

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 54

SELF-SUPERVISED ONLINE ROBOT-AGNOSTIC TRAVERSABILITY ESTIMATION FOR OPEN-WORLD ENVIRONMENTS

Hindel J., Bultmann S., Masnavi H., Cattaneo D., Valada A.

Abstract: COTRATE is an online learning framework for continuous traversability estimation from unlabeled robot experiences. A robot-agnostic terrain assessment module infers traversability scores from proprioceptive signals, supervising a visual network through our alignment loss. A diversity-aware feature selection strategy mitigates forgetting with minimal overhead. We evaluate COTRATE on 50,000 images across 11 outdoor terrains and navigation tasks in three environments on two robotic platforms.

Contact: hindel@cs.uni-freiburg.de

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 55

TRAINABLE HIGHLY-EXPRESSIVE ACTIVATION FUNCTIONS

Chelly I., Finder S. E., Ifergane S., Freifeld O.

Abstract: Nonlinear activation functions (AFs) are pivotal to deep neural nets, enabling them to approximate complex functions. Existing fixed/trainable AFs either rely on fixed-shape functions or require parameters that scale with model size. We propose DiTAC, a trainable AF based on highly expressive and efficient diffeomorphic transformations. DiTAC achieves significant improvements on various datasets and tasks, including semantic segmentation, image generation, image classification, and regression.

Contact: shiraif@post.bgu.ac.il

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 56

PHYSICALLY PLAUSIBLE HUMAN–OBJECT INTERACTION GENERATION VIA ATTRIBUTE CLASSIFIER GUIDANCE

Ikeuchi K., Ohkawa T., Shinoda R., Sato Y.

Abstract: Human-Object Interactions (HOIs) are shaped by physical attributes like object mass. Yet diffusion models rely only on surface geometry, ignoring mass-driven dynamics. We propose Attribute Classifier Guidance, a plug-and-play framework that steers pre-trained diffusion sampling via gradients of a lightweight attribute classifier. Validated on object mass, it yields more physically plausible motion—e.g., upright postures for lighter objects—while maintaining competitive generation quality.

Contact: kikeuchi@iis.u-tokyo.ac.jp

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 57

SEEING WITH SOUND

Nazrul Ismail, Owais Ahmed Malik, Ong Wee Hong

Abstract: Visual systems often struggle in low-light while echoes inherently encode geometric structure in the absence of light. With the advent of vision foundation models, we transfer the spatial intelligence for these models to audio by Cross-modal distillation. Existing audio-only depth estimation methods remain sparse, and current cross-modal Knowledge Distillation (KD) approaches often suffer from severe modality misalignment. In this work, we investigate whether binaural echoes alone can recover dense scene depth when guided by a vision teacher through latent-space distillation.

Contact: nazftri.is@gmail.com

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 58

ZERO-SHOT SIM-TO-REAL DETECTION AND SEGMENTATION FOR ROBOTIC EV BATTERY DISASSEMBLY

Isoaho J.1*, Rätz R.1,2, Özen Ö.1, Ochsenbein C.1

Abstract: To overcome the lack of vision training data for automated EV battery recycling, we propose a pure sim-to-real pipeline. Digital twins in NVIDIA Isaac Sim are used to generate synthetic data leveraging domain randomization. Requiring no real-world annotations, models trained only in simulation achieve 86.1% F1 for detecting bolts on real Hyundai Kona modules, and 87.9% mAP@50 for segmenting bolts across entire Kona packs. These models drive autonomous robotic disassembly.

Contact: jesse.isoaho@sipbb.ch

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 59

SOKE-GRPO: REINFORCEMENT LEARNING FOR TEXT-TO-SIGN GENERATION

Isotton G., Talon D., Ricci E.

Abstract: This work addresses Text-to-Sign Language Video Generation, which aims to produce temporally coherent signing sequences from text while preserving sentence meaning. We apply Group Relative Policy Optimization (GRPO) to improve our SOKE baseline, a 3D pose-based sign language video generation model. Optimization is guided by a semantic retrieval reward that measures alignment between input text and generated sign sequences.

Contact: gloria.isotton@unitn.it

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 60

OPTIMIZING INCOMPLETE, LARGE-SCALE AND SPARSE MULTI-GRAPH MATCHING

Stricker S., Kahl M., Hutschenreiter L., Bernard F., Rother C., Savchynksyy B.

Abstract: Consistently identifying the same physical structures across many views, scans, or instances is fundamental to spatial intelligence and known formally as the multi-graph matching problem. We argue, partly by proof, that large-scale methods must address sparsity and incompleteness, and thus introduce GREEDA, a general solver instantiable as a direct and a synchronization method. In our experiments, both variants tie their respective competitors at small scale and lead at large scale.

Contact: mkahl@mpi-inf.mpg.de

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 61

INVSPLAT: INVERSE FEED-FORWARD SCENE SPLATTING

Karpikova P., Bian W., Xu H., Lensch H., Geiger A.

Abstract: We present a feed-forward multi-view framework for inverse rendering that directly predicts 3D Gaussian primitives with intrinsic material attributes — albedo, metallic, and roughness. Unlike optimization-based methods requiring costly per-scene fitting, or 2D approaches lacking an explicit 3D representation and struggling with multi-view consistency, our model jointly recovers consistent geometry and reflectance in a single forward pass, enabling novel view synthesis with relighting.

Contact: poliikwork@gmail.com

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 62

LEARNING GLOBAL CAMERA POSES FROM NOISY VIEW-GRAPHS FOR STRUCTURE FROM MOTION

Khatib F., Galun M., Basri R.

Abstract: Camera pose estimation is central to 3D reconstruction and view synthesis. We present a deep global SfM method that aggregates noisy pairwise relative poses into consistent camera poses. The model is trained without ground-truth supervision and is followed by triangulation and bundle adjustment. It is efficient, scalable, and achieves accurate results on standard benchmarks.

Contact: fadi.khateeb@weizmann.ac.il

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 63

SELF-AUGMENTED RESIDUAL 3DGS FOR NEXT BEST VIEW SELECTION

Jun-Seong K., Oh T.-H., Eduardo P.-P., Jang Y.

Abstract: SA-ResGS stabilizes uncertainty-aware Next-Best-View selection by generating geometry-consistent SA-Points from observed and extrapolated views, and improves sparse wide-baseline training through residual supervision that strengthens gradients for high-uncertainty Gaussians. Experiments show consistent gains in reconstruction quality and view selection robustness across multiple benchmarks.

Contact: junseong.kim@postech.ac.kr

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 64

ALL-DAY DEPTH COMPLETION VIA THERMAL-LIDAR FUSION

Kim J., Kweon M., Shin U., Park J.

Abstract: RGB-based depth completion often degrades under low-light and rainy conditions, while ground-truth depth maps suffer from missing measurements that limit supervision. To address these challenges, we investigate thermal-LiDAR depth completion and propose COPS, a framework that exploits a depth foundation model through depth-aware contrastive learning and pseudo-supervision. COPS enhances depth boundary clarity and completion accuracy, and extensive benchmarks demonstrate its robustness across diverse environments.

Contact: jangjoa41@pusan.ac.kr

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 65

REFLETION-AWARE GENERATIVE NOVEL VIEW SYNTHESIS

Kim G., Dong-Yeon S., Oh T. H.

Abstract: We propose a reflection-aware method for generative novel view synthesis (NVS) in mirror scenes. Existing generative NVS methods lack mirror-awareness and generate reflection-inconsistent novel view images. To overcome this limitation, we treat a mirror image as two complementary views, enabling reflection-consistent NVS. Our method inherits the strong generalizability of the multi-view diffusion backbone while requiring no finetuning.

Contact: geonukim@kaist.ac.kr

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 66

NEURO-SYMBOLIC OUT-OF-DISTRIBUTION DETECTION VIA SCENE GRAPH REASONING

Kirchheim K., Czarnecki K., Ortmeier F.

Abstract: A vision system that acts on what it sees is only trustworthy if it can flag inputs unlike anything it was trained on—the out-of-distribution (OOD) problem. Standard detectors judge this from a network's opaque internal features. We instead read the image as objects and their relations (a scene graph) and check plausibility against simple, human-readable rules—for example, "a face sits at the center." A scene that breaks high-weight rules is flagged, and SGR shows exactly which rule it broke.

Contact: konstantin.kirchheim@ovgu.de

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 67

CAN AI UNDERSTAND GARMENT CONDITION? TOWARDS FULLY AUTOMATED TEXTILE SORTING FOR REUSE

Kirillova N., Possegger H.

Abstract: Automated inspection and decision-making for large-scale second-hand clothing sorting support a circular economy. In collaboration with industrial sorting facilities, we investigate computer vision approaches, including segmentation, category classification, defect detection, monocular depth estimation, and anomaly detection, for fully automated garment condition assessment to enable reuse, resale, and recycling in the fashion industry. We aim to understand how effectively state-of-the-art visual models can support reliable reuse-oriented sorting across diverse garment categories and conditions handled by a robotic system.

Contact: nadezda.kirillova@tugraz.at

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 68

SURGE: IMPROVED SURFACE GEOMETRY IN POINT MAPS

Knaebel K., Martin Garcia G., Schmidt C., Fradlin I., Nunes L., de Geus D., Leibe B.

Abstract: We focus on a qualitative issue in current feedforward 3D reconstruction models: they estimate global geometry well, but local structures in their point maps can still look noisy or warped, often with oscillatory artifacts, especially around thin structures. This is clearly visible qualitatively but only weakly reflected in common metrics. We suggest a metric to make these errors more explicit, and two components to reduce these errors: a loss formulation and a Neighborhood Attention Decoder.

Contact: knaebel@vision.rwth-aachen.de

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 69

SLIDER: SLIDER-GUIDED LATENT IMAGE DISCOVERY FOR EXPLAINABLE RETRIEVAL

Kolouju P., Qaiser I., Xing E., Stylianou A., Souvenir R., Pless R., Jacobs N.

Abstract: Text-to-image search is challenging when there are many attributes to describe that are not well captured in the gallery results. SLIDER is an image search interface where users are shown query-relevant concepts in the form of sliders. Sliders allow users to emphasize or de-emphasize the presence of concepts in the gallery. We explore the SLIDER interface, the underlying retrieval and reranking methods, and the planned user study and evaluation to assess the utility of sliders as search tools.

Contact: pranavi.kolouju@slu.edu

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 70

MMLANDMARKS: A CROSS-VIEW INSTANCE-LEVEL BENCHMARK FOR GEO-SPATIAL UNDERSTANDING

Kristoffersen O., Sánchez A., Hannemose M., Dahl A., Papadopoulos D.

Abstract: Geospatial models today train and evaluate in isolation across modalities. We introduce MMLandmarks: 18,557 U.S. landmarks with one-to-one Ground / Satellite / GPS / Text correspondence. A simple contrastive baseline shows strong performance on crossview retrieval and geolocalization.

Contact: ofhkr@dtu.dk

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 71

DO IMAGE EDITING MODELS UNDERSTAND LIGHTING?

Küchler T.*, Feiden J.*, Nießner M., Rother C.

Abstract: Do AI image editing models understand real-world lighting? We introduce 3DLP (3D-anchored Light Probe), a benchmark and a new light-effect-labeled HDR dataset of 1K indoor image pairs capturing physical light changes (on/off). Using two new evaluation metrics that account for exposure and white balance shifts, we evaluate eight cutting-edge editing models. We find that while leading models understand light transport remarkably well, others completely fail to reason about realistic physics.

Contact: tim.kuechler@iwr.uni-heidelberg.de

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 72

EARLY PROSTATE CANCER DETECTION USING AI-ASSISTED ABDOMINAL ULTRASOUND

Kurucz L.M., Natali T., Mertens L.S., Van Leeuwen P.J., Dashtbozorg B., Ruers T., De Korte C.

Abstract: Introduction. Prostate cancer is a major health concern requiring accurate and accessible methods for early detection and risk stratification. Prostate volume (PV) is a key parameter in multivariate risk assessment, traditionally measured using transrectal ultrasound (TRUS). While TRUS provides precise measurements, its invasive nature affects patient comfort. Transabdominal ultrasound (TAUS) offers a non-invasive alternative but is limited by lower image quality and operator dependence. This study presents a deep-learning-based framework for automatic PV estimation using TAUS, aiming to improve non-invasive prostate cancer risk stratification.

Methods. A dataset of TAUS videos from 228 patients (median age 67, 95-percentile range 55–81.2) was curated, with expert-delineated prostate boundaries and diameter calculations as ground truth. The framework integrates deep-learning models for prostate segmentation in both axial and sagittal planes, automatic diameter estimation, and PV calculation. Segmentation performance was evaluated using Dice correlation coefficient (%) and while volume estimation accuracy was assessed through volumetric error (mL).

Results. The axial model outperformed the sagittal model, achieving a Dice score of 0.76 $\pm$ 0.16 versus 0.68 $\pm$ 0.21, a Dice-MidPlane of 0.91 $\pm$ 0.06 versus 0.83 $\pm$ 0.10. The framework estimated PV with a mean volumetric error of -5.4 mL (95% limits of agreement: -25.6 to 36.4 mL), resulting in a relative error of 7%.

Conclusion. These findings highlight the potential of deep learning for accurate, non-invasive PV estimation, supporting improved prostate cancer risk assessment. Particularly, the use of axial-only TAUS assessment enhances the accessibility of abdominal PSAD estimation, without compromising accuracy. Our framework may enable non-invasive PSAD assessment to support risk stratification and guide MRI referral in early prostate cancer detection.

Contact: l.kurucz@nki.nl

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 73

A TRAINING-FREE STYLE PERSONALIZATION VIA SVD-BASED FEATURE DECOMPOSITION

Lee K.*, Park J.*, Gim J.*, Choi W., Hwang K., Kim J., Im S.

Abstract: We present a training-free framework for style-personalized image generation using a scale-wise autoregressive model. We identify a key step where the dominant singular values of the feature encode style. We introduce two lightweight modules: Principal Feature Blending for precise style modulation, and Structural Attention Correction for structural consistency. Our method achieves competitive style and prompt fidelity without extra training, offering faster inference and deployment flexibility.

Contact: kyoungmin@dgist.ac.kr

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 74

SUBSPACEAD: TRAINING-FREE FEW-SHOT ANOMALY DETECTION VIA SUBSPACE MODELING

Lendering Camile, Akdag Erkut, Bondarev Egor

Abstract: Detecting visual anomalies in industrial inspection often requires training with only a few normal images per category. Recent few-shot methods achieve strong results employing foundation-model features, but typically rely on memory banks, auxiliary datasets, or multi-modal tuning of vision-language models. We therefore question whether such complexity is necessary given the feature representations of vision foundation models. To answer this question, we introduce SubspaceAD, a training-free method, that operates in two simple stages. First, patch-level features are extracted from a small set of normal images by a frozen DINOv2 backbone. Second, a Principal Component Analysis (PCA) model is fit to these features to estimate the low-dimensional subspace of normal variations. At inference, anomalies are detected via the reconstruction residual with respect to this subspace, producing interpretable and statistically grounded anomaly scores. Despite its simplicity, SubspaceAD achieves state-of-the-art performance across one-shot and few-shot settings without training, prompt tuning, or memory banks. In the one-shot anomaly detection setting, SubspaceAD achieves image-level and pixel-level AUROC of 97.1% and 97.5% on the MVTec-AD dataset, and 93.2% and 98.2% on the VisA dataset, respectively, surpassing prior state-of-the-art results.

Contact: c.r.lendering@tue.nl

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 75

EGOINTERACT: SYNTHETIC EGOCENTRIC VIDEOS GENERATION FOR INTERACTION UNDERSTANDING AND ANTICIPATION

Leonardi R., Ragusa F., Materia D., Passanisi A., Fort J., Engel J., Farinella G.M.

Abstract: Collecting large-scale egocentric video datasets with dense spatial and temporal annotations is costly, slow, and often constrained by environmental biases, privacy constraints, and limited coverage of interaction patterns. While synthetic data has shown strong potential in several vision domains, its use for egocentric perception remains relatively underexplored, especially for tasks requiring temporally coherent human-object interactions. In this work, we introduce EgoInteract, a controllable simulator for egocentric video generation designed to model fine-grained egocentric interactions and their temporal dynamics. The simulator enables precise control over camera, human body and hand motion, object manipulation, and scene composition across diverse environments. Building on this framework, we generate a synthetic egocentric video dataset with dense spatial and temporal annotations for temporal action segmentation, next-active object detection, interaction anticipation, and hand-object interaction detection. We evaluate models trained with simulated data on multiple real-world egocentric benchmarks spanning diverse environments, object categories, and interaction patterns. Results show consistent improvements over strong baselines across tasks and datasets, demonstrating the effectiveness and transferability of our simulation-based approach.

Contact: rosario.leonardi@unict.it

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 76

CONFORMALIZED FLOWMATCHING FOR TRUSTWORTHY SYNTHETIC MEDICAL IMAGE GENERATION

Li J., Li H.B.

Abstract: • Research question: Which regions can be trusted in CT-to-PET synthesis? • Method: We combine Flow Matching with conformal prediction to transform sampling uncertainty into calibrated trust maps. • Result: The resulting trust maps highlight reliable regions and flag uncertain areas for cautious interpretation.

Contact: jialin-li@nus.edu.sg

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 77

ECHO2ECG: ENHANCING ECG REPRESENTATIONS WITH CARDIAC MORPHOLOGY FROM MULTI-VIEW ECHOS

Liman, M., Turgut Ö., Müller A., Martens E., Rueckert D., Müller P.

Abstract: ECG is low-cost but cannot measure cardiac morphological phenotypes like LVEF, which require Echo. Predicting them from ECG would enable accessible screening. Existing SSL methods align ECGs to single-view Echos, limiting supervision to local anatomy. We propose Echo2ECG, a multimodal SSL framework enriching ECG representations with multi-view Echos. Our ECG encoder outperforms SOTA baselines on structural phenotype classification across 3 datasets, while 18x smaller than the largest baseline.

Contact: michelle.liman@tum.de

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 78

INTENT ANALYSIS IN HISTORICAL VISUAL ARCHIVES

Lin T., Aigner W., Sablatnig R.

Abstract: Historical film is not a neutral record. Each shot carries intent, its makers' motivations and messages, and recovering it helps us understand the past. My PhD reads intent along three dimensions, content (what is shown), cinematographic settings (how it is filmed), and expert context (who, where, when), using computer vision and visual analytics. The works here show that, even on degraded archives where standard models fail, each dimension can be analyzed accurately and interpretably.

Contact: tylin@cvl.tuwien.ac.at

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 79

VISUAL LOCALIZATION: FROM CONTINENTS TO CORRESPONDENCES

Lindenberger P.

Abstract: Determining the precise geographic location of an image at a global scale remains an unsolved problem. Such a system has to handle vast areas, missing data, and a lack of distinctive visual features. We cover three works that integrate towards this end-goal: A continent-level localization system that combines classification and cross-view retrieval at an unprecedented scale, a practical keypoint detector with metric covariances, and a sparse-to-dense matcher that disambiguates unmatchable pairs.

Contact: philipp.lindenberger@inf.ethz.ch

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 80

TOWARDS MULTIMODAL AI THAT KNOWS WHAT IT DOESN'T KNOW

Liu M., Dong H., Fink O., Trapp M.

Abstract: Multimodal models often fail silently, producing overconfident errors on OOD and hard ID inputs. To build trustworthy AI, we introduce two complementary frameworks: Feature Mixing (FM), an extremely fast multimodal outlier synthesis method for OOD detection and segmentation (NeurIPS 2025), and Adaptive Confidence Regularization (ACR) for robust multimodal failure detection (CVPR 2026). These approaches enable safe rejection and improve robust generalization.

Contact: moru.liu@tum.de

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 81

SPATIALLY AWARE WORLD ACTION MODEL VIA GEOMETRIC LATENT DIFFUSION

Lopetegui-Gonzalez Javier, Pacaud Paul, Schmid Cordelia

Abstract: SA-WAM grounds World-Action Models in explicit 3D structure, using either depth or PointMaps. A nonlinear metric encoding makes 3D signals compatible with the frozen VAE tokenizer while preserving near-field resolution for manipulation. The resulting spatially aware WAM achieves state-of-the-art RoboCasa performance and improves visual consistency in future-state prediction. These gains also transfer to a real-world UR5 robotic-arm setting.

Contact: javier-alejandro.lopetegui-gonzalez@inria.fr

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 82

VITTR3: VISUAL-INERTIAL TEST-TIME REFINEMENT OF 3D FOUNDATION MODELS

Lozano E., Jaenal A., Civera J.

Abstract: 3D Foundation Models predict geometry from unposed, uncalibrated images in one pass. To generalize, they rely on relative priors, not metric information. Vision alone is thus constrained: a static forward pass cannot recover absolute scale or scene-specific dynamics. Inertial sensing resolves both, anchoring metric dimensions and physical motion. We refine the model's weights at test-time to enforce cross-modal consistency with the IMU stream, yielding metric, physically consistent geometry.

Contact: e.lozano@unizar.es

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 83

GAUSSIAN WORLD: 3DGS FROM SCENE RECONSTRUCTION TO UNDERSTANDING

Ma, Mengjiao; Ma, Qi; Li, Yue; Cheng, Jiahuan; Yang, Runyi; Ren, Bin; Popovic, Nikola; Wei, Mingqiang; Sebe, Nicu; Van Gool, Luc; Gevers, Theo; Oswald, Martin R; Paudel, Danda Pani

Abstract: We are building a system for robust 3D scene reconstruction, understanding, and natural language interaction within complex environments. Our goal is to develop a foundation model that can tokenize complex 3D scenes, perform instance detection, and enable spatial reasoning to answer complex language queries—ultimately working to make 3D scene understanding as accessible and powerful as its 2D counterpart.

Contact: mengjiao.ma@insait.ai

Presentation Type: Poster

Date: Monday 6 July 2026

Time: 21:30

Poster Session: 1

Poster Number: 84

DO 3D LLMS REALLY UNDERSTAND 3D SPATIAL RELATIONSHIPS?

Ma X., Sun T., Chen S., Bhalgat Y., Gu J., Chang A.X., Armeni I., Laina I., Peng S., Prisacariu V.

Abstract: We test whether 3D-LLMs truly understand 3D space. We find that existing benchmarks contain linguistic shortcuts. We introduce Real-3DQA, a benchmark for 3D-dependent reasoning across viewpoints. Real-3DQA reveals poor 3D understanding and motivates 3D-reweighted fine-tuning.

Contact: some5596@ox.ac.uk

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 85

LEXUS: LIDAR OUTLIER EXPOSURE FOR 3D UNKNOWN SEGMENTATION IN AUTONOMOUS DRIVING

Marinai A., Rai S., Masone C., Tommasi T.

Abstract: LExUS is a simple method to enhance Out-of-Distrubution (OOD) segmentation of a LiDAR segmentation model. Inspired by Outlier Exposure in 2D, 3D scenes are augmented by inserting OOD objects from an auxiliary dataset into In-Distribution (ID) scenes and then are used to train a model to jointly segment OOD and inliers objects through a novel fine-tuning loss.

Contact: alessandro.marinai@polito.it

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 86

LEVERAGING GAZE AND SET-OF-MARK IN VLLMS FOR HUMAN-OBJECT INTERACTION ANTICIPATION FROM EGOCENTRIC VIDEOS

Materia D., Ragusa F., Farinella G.M.

Abstract: Human-Object Interaction anticipation – predicting which object will be interacted with before contact occurs – is key for proactive assistance. We leverage Set-of-Mark prompting and the user's gaze to tackle visual grounding and user intent understanding limitations in existing VLLM-based approaches. To better capture the temporal dynamics preceding interactions, we introduce an inverse exponential sampling strategy for video frames. Our proposed pipeline is training-free and model-agnostic.

Contact: daniele.materia@studium.unict.it

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 87

PROSKILL: SEGMENT-LEVEL SKILL ASSESSMENT IN PROCEDURAL VIDEOS

Mazzamuto M., Di Mauro D., Francesca G., Farinella G.M., Furnari A.

Abstract: Skill assessment in procedural videos is crucial for the objective evaluation of human performance in settings such as manufacturing and procedural daily tasks. Current research on skill assessment has predominantly focused on sports and lacks large-scale datasets for complex procedural activities. Existing studies typically involve only a limited number of actions, focus on either pairwise assessments (e.g., A is better than B) or on binary labels (e.g., good execution vs needs improvement). In response to these shortcomings, we introduce PROSKILL, the first benchmark dataset for action-level skill assessment in procedural tasks. PROSKILL provides absolute skill assessment annotations, along with pairwise ones. This is enabled by a novel and scalable annotation protocol that allows for the creation of an absolute skill assessment ranking starting from pairwise assessments. This protocol leverages a Swiss Tournament scheme for efficient pairwise comparisons, which are then aggregated into consistent, continuous global scores using an ELO-based rating system. We use our dataset to benchmark the main state-of-the-art skill assessment algorithms, including both ranking-based and pairwise paradigms. The suboptimal results achieved by the current state-of-the-art highlight the challenges and thus the value of PROSKILL in the context of skill assessment for procedural videos. All data and code are available at https://fpv-iplab.github.io/ProSkill/.

Contact: michele.mazzamuto@unict.it

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 88

SIGNIT: A COMPREHENSIVE DATASET AND MULTIMODAL ANALYSIS FOR ITALIAN SIGN LANGUAGE RECOGNITION

Micieli A., Farinella G.M., Ragusa F.

Abstract: We present SignIT, a new dataset for Italian Sign Language (LIS) recognition. It contains 644 videos (3.33 hours) annotated with 94 sign classes grouped into five semantic categories. The dataset also provides 2D hand, face, and body keypoints. We introduce a benchmark based on state-of-the-art recognition models using RGB frames and pose information. Experimental results highlight the importance of temporal cues and reveal the challenges posed by this diverse and realistic LIS dataset.

Contact: micieli.alessia@studium.unict.it

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 89

WASABI: WEAKLY-SUPERVISED ANATOMY PRESERVING LESION SYNTHESIS WITH IMPLICIT ANOMALY LOCALIZATION FOR MEDICAL IMAGE BOOTSTRAPPING AND SEGMENTATION IMPROVEMENT

Mitic B. , Prosch H. , Langs G.

Abstract: We propose a class-conditioned diffusion framework for pathological lesion synthesis in medical images with an integrated lesion localization mechanism that doesn't require pixel-level annotations during training. Additionally, the framework supports pseudo-healthy reconstruction, enabling controlled transformation between disease classes while preserving underlying anatomy. Experiments show realistic, diverse samples that improve classification and segmentation in low-data settings.

Contact: branko.mitic@meduniwien.ac.at

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 90

PRIOR-GUIDED GRASP ADMISSIBILITY FOR SAFE LANDMINE MANIPULATION

Miuccio A., Lebecque F., Le Flécher E., Hamesse C., Tsiogkas N., Detry R., Haelterman R.

Abstract: We present a prior-guided solution for safely parameterizing a parallel-jaw grasp on a PMN1 anti-personnel landmine from a single stereo observation. Instead of generating arbitrary 6-DoF grasps, the proposed pipeline verifies whether the current view supports the contact and orientation constraints required for hazardous-object manipulation. The method segments the mine, estimates its 3D position, infers the protrusion axis, and accepts a view only when the gripper can be centered so that both fingers contact the cylindrical body while avoiding the pressure plate and detonator protrusion. For accepted views, the system outputs a camera-frame grasp parameterization with an admissible grasp center and gripper orientation perpendicular to the protrusion axis. Evaluated across four outdoor scenarios, the method rejects views in which a safe grasp cannot be established.

Contact: alessandra.miuccio@mil.be

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 91

CAN WE TRUST THE NEXT ACTION? GROUNDED FUTURE REASONING FOR RELIABLE EGOCENTRIC ANTICIPATION

Mahsa MohammadiEshkaftaki(Mohammadi)

Abstract: Egocentric action anticipation aims to predict what a camera wearer will do before an action begins. However, accuracy alone is not enough for real world use. This poster focuses on three common problems: ambiguous motion, implausible action sequences, and unstable predictions over time. It presents a reliability focused view in which predictions should be grounded in objects, consistent across frames, physically plausible, and fast enough for real time interaction.

Contact: mm1510@exeter.ac.uk

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 92

LIDO: LEARNING TO IDENTIFY OUT-OF-DISTRIBUT OBJECTS FOR 3D LIDAR ANOMALY SEGMENTATION

Mosco S., Fusaro D., Pretto A.

Abstract: We present LIDO, a novel approach for 3D LiDAR Anomaly Segmentation that directly works on the feature space with feature prototypes and contrastive learning to distinguish known and unknown classes and identify anomalous objects.

Contact: simone.mosco@phd.unipd.it

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 93

VOID: VIDEO OBJECT AND INTERACTION DELETION

Motamed S., Harvey W., Klein B., Van Gool L., Yuan Z., Cheng T

Abstract: Current video inpainting ignores physical causality, leading to impossible scenes when interacting objects are removed. We propose VOID, an interaction-aware framework utilizing Vision-Language Models (VLMs) to predict downstream consequences. By mapping these effects into a dynamic "quadmask," VOID guides a video diffusion model to re-simulate the scene, producing physically accurate, counterfactual videos where the environment reacts realistically to the missing object.

Contact: sam.motamed@insait.ai

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 94

WBC-CLIP: A MULTIMODAL VISION-LANGUAGE FRAMEWORK FOR MORPHOLOGY AWARE WHITE BLOOD CELL ANALYSIS

Zedda L., Mura D.A., Manzo A., Di Ruberto C., Loddo A.

Abstract: WBC-CLIP is a dual-encoder vision-language framework for white blood cell analysis. It pairs WBC images with LLM-generated descriptions of quantitative morphological features and learns aligned image-text embeddings through contrastive training. On WBCAtt, WBC-CLIP reaches up to 69.92% average macro F1, outperforming BioMedCLIP and MedGemma. On the out-of-distribution LeukemiaAttri dataset, it achieves up to 52.72% average macro F1, supporting robust morphology-aware analysis.

Contact: davideantonio.mura@unica.it

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 95

LEARNING FROM SYNTHETIC DATA VIA PROVENANCE-BASED INPUT GRADIENT GUIDANCE

Nagano K., Fujii R., Hachiuma R., Sato F., Sekii T., Saito H.

Abstract: Research Question: Can auxiliary information obtained when training on synthetic data—such as which regions were synthesized, where the target objects were placed, etc.—be used as supervisory signals? We propose a training method to automatically correct the model's focus area using these supervisory signals and enhance the robustness of a weakly supervised model.

Contact: koshiro.nagano@gmail.com

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 96

VISUAL-INERTIAL EGOCENTRIC METRIC DEPTH ESTIMATION FOR TABLETOP ACTIVITIES OF DAILY LIVING

Nalivayko Y., Wochner I.

Abstract: Neurodegenerative diseases impact the ability to perform the tabletop activities of daily living such as eating. Exoskeleton assistive devices can suppress the pathological movement components but require extensive scene information, e.g. depth. By exploiting the biomechanical constraints of a person sitting at a table and utilizing monocular visual data and inertial measurements from the Neon eye-tracking eyeglasses, we aim to estimate the depth of the objects on the table in real-time.

Contact: yaroslava.nalivayko@uni-tuebingen.de

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 97

VISUAL ODOMETRY THAT TUNES ITSELF

Nascivera S., Bauersfeld L., Delaune J., and Scaramuzza D.

Abstract: We present the first system that maps an input image to the best parameter configuration for your favorite SLAM/VO pipeline.

Contact: snascivera@ifi.uzh.ch

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 98

DYNAMIC CONTRAST ENHANCEMENT BRIDGE FOR CONTRAST-FREE BREAST MRI

Newegy S., Nagarajan B., Radeva P.

Abstract: Contrast-enhanced breast MRI requires intravenous gadolinium, adding cost and patient risk. DCEB (Dynamic Contrast Enhancement Bridge) predicts the post-contrast scan from the pre-contrast image alone, with no tumor mask at test time. It integrates a Schrödinger-bridge diffusion model that learns only the contrast uptake, a tumor localizer that guides enhancement, and an attention network preserving realistic tissue, achieving high image fidelity, tumor fidelity, and localization.

Contact: salma.newegy@bsc.es

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 99

ITALIANPARKS400K: LARGE SCALE EUROPEAN SPECIES DATASET AND BASELINE FOR AUTOMATED ECOLOGICAL ANALYSIS FROM CAMERA TRAP DATA

Niccoli N., Seidenari L., Greco I., Salvatori M., Rovero F.

Abstract: ItalianParks400k is a large European camera-trap dataset with 400k+ wildlife images, 35k+ sequences, 600k+ boxes, 61 species, and metric distance annotations. We provide an open pipeline for detection, species recognition, tracking, and 3D localization, enabling automated ecological analysis beyond species classification.

Contact: niccolo.niccoli@unifi.it

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 100

VIDEOMT: YOUR VIT IS SECRETLY ALSO A VIDEO SEGMENTATION MODEL

Norouzi N., Zulfikar I. E., Cavagnero N., Kerssies T., Leibe B., Dubbelman G., de Geus D.

Abstract: VidEoMT is a simple encoder-only video segmentation model that removes complex tracking modules. It enables temporal modeling by propagating queries from previous frames and fusing them with learnable queries for new content. This gives tracking-like behavior without added complexity, achieving competitive accuracy while being 5×–10× faster and reaching up to 160 FPS with a ViT-L backbone.

Contact: n.norouzi@tue.nl

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 101

REAL-TIME UNDERWATER TRASH DETECTION AND SEGMENTATION ON RESOURCE-CONSTRAINED ROVS VIA EDGE AI OPTIMIZATION

Jiregna Abdissa Olana, De zan Alberto, Tavaris Denis, Gian Luka Foresti, Carlo Drioli

Abstract: Marine debris poses a growing threat to ocean ecosystems, creating an urgent need for autonomous and efficient underwater monitoring. We present a deployment-aware Edge AI framework for real-time trash detection and segmentation on resource-constrained ROVs. The approach unifies different marine debris datasets and combines lightweight YOLO26n multi-task learning with quantization-based optimization for embedded deployment. Achieving 0.916 mAP@50 and 0.934 mask mAP@50, the system was validated through real-field underwater experiments under varying visibility conditions, occlusions, and ROV motion, supporting autonomous monitoring and cleanup missions.

Contact: olana.jiregnaabdissa@spes.uniud.it

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 102

ENABLING LARGE-SCALE ANALYSIS OF HISTORICAL LOGIC DIAGRAMS IN BYZANTINE MANUSCRIPTS

Osburg Lilly, Dr. Götzelmann Germaine, Dr. Tonne Danah, Prof. Streit Achim

Abstract: This project develops a pipeline for handwritten diagram recognition in Byzantine copies of Aristotle's Organon, enabling large-scale analysis of diagrams. The manuscripts pose significant challenges due to material degradation, handwriting variability, and heterogeneous digitization methods. Using a newly created dataset, segmentation, classification, and detection-based approaches are compared. Results indicate that detection-based approaches achieve the strongest overall performance.

Contact: lilly.osburg@kit.edu

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 103

ONE PATCH TO CAPTION THEM ALL A UNIFIED ZERO-SHOT CAPTIONING FRAMEWORK

Bianchi L., Pacini G., Carrara F., Messina N., Amato G., Falchi F.

Abstract: Traditional captioning requires massive image-text datasets to train. Localized tasks also require annotations for each granularity. Without a unified framework, each level of granularity requires a separated model and training. Zero-shot captioners bypass the data bottleneck by (A) training a text decoder only on text embeddings and (B) applying it to visual features at inference. Such models are still image-centric. Patch-ioner is the first patch-centric zero-shot captioning framework.

Contact: giacomo.pacini@isti.cnr.it

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 104

PERCEPTION - REASONING DATASET FINGERPRINTS VIA TRAJECTORY ANALYSIS UNDER VLM SCALING

Paez-Ubieta I.D.L., Pieters R.

Abstract: Multimodal datasets currently mix perception and reasoning demands [1], which are challenging to detect given current evaluation approaches [2]. However, observing how output trajectories change across model capacities may show a dataset fingerprint. In this work, model scaling analysis of the same data reveals family-dependent trajectories across different dataset mixes. These results can fingerprint datasets and help guide model / image-size selection.

Contact: ignacio.paezubieta@tuni.fi

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 105

EXPLAINING FOR BETTER MODELS, MODELING FOR BETTER EXPLANATIONS

Parchami-Araghi Amin

Abstract: Deep networks achieve remarkable performance, yet their decision-making is opaque, highlighting the importance of explanation methods. In our first two works, we show that explanations are more than a tool for interpretation: they can guide training and even transfer knowledge between models. Our latest work explores the reverse direction: improving model design for better explanations, which leads to a competitive model that provides faithful traces of interpretable concepts for its decisions.

Contact: amin.parchami@mpi-inf.mpg.de

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 106

3D-LATTE: LATENT SPACE 3D EDITING FROM TEXTUAL INSTRUCTIONS

Parelli M., Oechsle M., Niemeyer M., Tombari F., Geiger A.

Abstract: Given a 3D asset and an edit instruction, 3D-LATTE produces high-quality, 3D-consistent and precise edits. Existing methods primarily leverage 2D or multi-view diffusion priors leading to multi-view inconsistencies and difficulty with morphological edits. We directly operate in the latent space of a native 3D diffusion model by inverting the input 3D representation and leveraging 3D attention control. Our motivation is that 3D attention maps capture information about the layout and composition of the 3D scene.

Contact: maryparelli@gmail.com

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 107

ENIGMA-360: AN EGO-EXO DATASET FOR HUMAN BEHAVIOR UNDERSTANDING IN INDUSTRIAL SCENARIOS

Ragusa F., Leonardi R., Mazzamuto M., Di Mauro D., Quattrocchi C., Passanisi A., D'Ambra I., Furnari A., Farinella G.M.

Abstract: Understanding human behavior from complementary egocentric (ego) and exocentric (exo) points of view enables the development of systems that can support workers in industrial environments and enhance their safety. However, progress in this area is hindered by the lack of datasets capturing both views in realistic industrial scenarios. To address this gap, we propose ENIGMA-360, a new ego-exo dataset acquired in a real industrial scenario. The dataset is composed of 180 egocentric and 180 exocentric procedural videos temporally synchronized offering complementary information of the same scene. The 360 videos have been labeled with temporal and spatial annotations, enabling the study of different aspects of human behavior in industrial domain. We provide baseline experiments for 3 foundational tasks for human behavior understanding: 1) Temporal Action Segmentation, 2) Keystep Recognition and 3) Egocentric Human-Object Interaction Detection, showing the limits of state-of-the-art approaches on this challenging scenario. These results highlight the need for new models capable of robust ego-exo understanding in real-world environments. We publicly release the dataset and its annotations at https://fpv-iplab.github.io/ENIGMA-360/.

Contact: ale.passa001@gmail.com

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 108

NAME THAT PART: 3D PART SEGMENTATION AND NAMING

Paul S., Kaushik P., Vaidya A., Bhattad A., Yuille A.

Abstract: We address semantic 3D part segmentation: decomposing objects into parts with meaningful names. Prior methods either provide unlabeled decompositions or retrieve individual parts. ALIGN-Parts casts naming as set alignment — shapes decompose into partlets (implicit 3D parts) matched to LLM affordance descriptions via bipartite assignment, fusing 3D geometry, multi-view appearance, and text. A text-alignment loss enables open-vocabulary matching: fast, one-shot, and a scalable 3D annotation engine.

Contact: soumava2016@gmail.com

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 109

INTERPRETABLE 3D NEURAL OBJECT VOLUMES FOR ROBUST CONCEPTUAL REASONING

Pham N., Jesslen A., Schiele B., Kortylewski A., Fischer J.

Abstract: CAVE unifies robustness and interpretability by learning sparse, faithful concepts from 3D neural object volumes. It produces consistent explanations across OOD settings and introduces 3D Consistency, a mesh-based metric for evaluating spatial consistency of concepts without human-annotated part annotations.

Contact: nhipham@mpi-inf.mpg.de

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 110

GENERALIZABILITY ANALYSIS OF DEEP LEARNING PREDICTIONS OF HUMAN BRAIN RESPONSES TO AUGMENTED AND SEMANTICALLY NOVEL VISUAL STIMULI

Piskovskyi V., Chimisso R., Patania S., Foulsham T., Vizzari G., Ognibene D.

Abstract: We investigate DL-based brain encoding models as a framework for predicting how image enhancement influences visual cortex activation. We use best-performing models to predict responses to augmentations involving faces, words, and semantically novel objects. Results provide evidence of the generalizability of brain encoding vision models, supporting their use in optimizing visual processing, perception-aware design, and AR/VR systems, as well as in advancing human-inspired visual intelligence.

Contact: v.piskovskyi@campus.unimib.it

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 111

R5DGS: SEMANTIC-AWARE 4D GAUSSIAN SPLATTING WITH RIGID BODY CONSTRAINTS FOR EFFICIENT DYNAMIC SCENE RECONSTRUCTION

Gridusov D., Popov M., Kolyubin S.

Abstract: Reconstructing and predicting dynamic 3D scenes from multi-view videos is crucial for robotics and AR/VR. Physics-informed Gaussian Splatting predicts future frames well but is computationally heavy and lacks semantics. We introduce R5DGS, a 4D Gaussian model with compact identity encodings for accurate object association and CLIP-based text retrieval. Motion is predicted via object centroids and propagated to Gaussians, yielding an 11 FPS speedup without quality loss.

Contact: mfpopov@itmo.ru

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 112

UNSPOKEN BIASES IN END-TO-END DRIVING BENCHMARKS

Porres D.

Abstract: CARLA has democratized end-to-end autonomous driving research, enabling reproducible closed-loop evaluation at scale. However, benchmarks built on it share implicit design assumptions that systematically distort model rankings regardless of real-world deployability. We identify five categories of unspoken bias in standard CARLA benchmarks, and using CIL++ as a representative policy, show that each leads to misleading conclusions about which architectures and training strategies are superior.

Contact: dporres@cvc.uab.es

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 113

SEMANITC-AWARE, PHYSICS-INFORMED, GEOMET GROUNDED WEATHER VIDEO SYNTHESIS

Chenghao Qian, Nedko Savov, Lingdong Kong, Yeying Jin, Rui Song, Wenjing Li, Zhun Zhong, Jiaqi Ma, Gustav Markkula, Luc Van Gool

Abstract: Weather synthesis aims to add weather effects to input videos while preserving scene identity, structure, and motion. The key limitation of existing methods is the lack of diversity in weather appearance and effective control over weather dynamics (e.g., temporal evolution and particle motion). Most approaches rely on text prompts, which are inherently underspecified and often fail to produce detailed weather characteristics. Additionally, general-purpose video editors optimized for clean and aesthetic outputs tend to suppress heavy weather phenomena, making dense particle effects difficult to generate. To address these, we propose a Semantic-Aware, Physics-Informed, and Geometry-Grounded framework that steers an off-the-shelf video editor to synthesize diverse global appearances and detailed particle dynamics. We factorize the synthesis into three conditional signals, so that each provides a distinct and stable source of guidance: semantics specifies what the weather should look like, dynamics governs how it evolves over time, and geometry determines where it should appear in the scene. Specifically, we introduce (1) semantic-aware appearance anchoring to establish the target appearance from scene semantics and user input; (2) physics-informed dynamic simulation to generate particle effects by simulating a Gaussian-represented particle field under gravity, wind, and turbulence; and (3) geometry-grounded video synthesis to align the simulated particles with target scene geometry and synthesize the final video. Experiments demonstrate that our method produces diverse weather effects with high physical and visual realism. Furthermore, the synthesized data significantly improves the robustness of semantic segmentation for autonomous driving under adverse weather conditions, yielding mIoU gains of up to 14.5%.

Contact: yohji.qian@gmail.com

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 114

GROUNDING FOUNDATION MODELS: REPRESENTATIONS FOR SPATIAL AND PHYSICAL INTELLIGENCE

Kaixian Qu, Mike Zhang, Zhengyu Fu, Cesar Cadena, Marco Hutter

Abstract: Foundation models (LLMs and VLMs) hold broad knowledge but cannot act in the physical world. The bridge is representation: text maps, object-probability maps, probabilistic functional scene graphs, and experiential memory each ground a foundation model, turning abstract knowledge into spatial and physical intelligence that real robots can use to perceive, navigate, reason, and act.

Contact: kaixqu@ethz.ch

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 115

ROBUST AND SECURE MRI: DEEP LEARNING ATROPHY ESTIMATION MEETS K-SPACE FINGERPRINTING

Riccardo Raciti

Abstract: Integrating deep learning into SIENA optimizes atrophy (PBVC) estimation, reducing runtimes and increasing clinical consistency. Concurrently, to protect data integrity against healthcare fraud, we propose extracting scanner noise fingerprints from K-Space. This validates both the superior performance of the upgraded clinical pipeline and its security against the manipulation or synthetic generation of MRI scans.

Contact: riccardo.raciti@phd.unict.it

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 116

PHYSICS-IQ VERIFIED

Rädsch T., Asano Y. M., Kuehne H., Bauer S., Jaini P., Geirhos R., Lüth C. T.

Abstract: Video generative models have become a new frontier not only for video generation, but also for a wide range of downstream tasks, including world modeling. To support progress on these tasks, video models must capture the physical structure and dynamics of the real world. Evaluating this form of understanding remains an emerging challenge and has motivated the Physics-IQ benchmark, which explicitly quantifies physical understanding by comparing model-generated videos against real-world recordings of physical experiments. In this work, we present a systematic audit of Physics-IQ, identify key shortcomings, and propose three improvements that sharpen the measurement of physical understanding in video generative models. Specifically, we improve prompt and ground-truth quality to reduce confounding factors, and introduce a sample-level scoring scheme that weights each sample and metric equally. The resulting benchmark, Physics-IQ Verified, refines 57.6% of all samples and improves 34.8% of prompts. In a comparison study across six image-to-video generative models, we observe moderate but meaningful ranking changes compared with the original benchmark, with Kendall's tau = 0.46. Physics-IQ Verified provides a more reliable signal for evaluating physically accurate video generative models and supports more robust progress toward physical understanding in generative video modeling. The benchmark code is available on the Physics-IQ Verified GitHub repository.

Contact: tim.raedsch@tum.de

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 117

EGO-EXTRA: VIDEO-LANGUAGE EGOCENTRIC DATASET FOR EXPERT-TRAINEE ASSISTANCE

Ragusa F., Mazzamuto M., Forte R., D'Ambra I., Fort J., Engel J., Furnari A., Farinella G. M.

Abstract: We present Ego-EXTRA, a video-language Egocentric Dataset for EXpert-TRAinee assistance. Ego-EXTRA features 50 hours of unscripted egocentric videos of subjects performing procedural activities (the trainees) while guided by real-world experts who provide guidance and answer specific questions using natural language. Following a "Wizard of OZ" data collection paradigm, the expert enacts a wearable intelligent assistant, looking at the activities performed by the trainee exclusively from their egocentric point of view, answering questions when asked by the trainee, or proactively interacting with suggestions during the procedures. This unique data collection protocol enables Ego-EXTRA to capture a high-quality dialogue in which expert-level feedback is provided to the trainee. Two-way dialogues between experts and trainees are recorded, transcribed, and used to create a novel benchmark comprising more than 45k high-quality Visual Question Answer sets, which we use to evaluate Multimodal Large Language Models. The results show that Ego-EXTRA is challenging and highlight the limitations of current models when used to provide expert-level assistance to the user. The Ego-EXTRA dataset is publicly available to support the benchmark of egocentric video-language assistants: https://fpv-iplab.github.io/Ego-EXTRA/.

Contact: francesco.ragusa@unict.it

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 118

INDOOR ROOM RECONSTRUCTION USING SMARTPHONES

Xuqian Ren, Juho Kannala, Esa Rahtu

Abstract: Geometric priors are often used to enhance 3D reconstruction. However, the accuracy of depth estimates from mobile devices is typically poor for highly detailed geometry, and monocular estimators often suffer from poor multi-view consistency and precision. We proposed Dn-splatter and AGS-Mesh for joint surface depth and normal refinement of Gaussian Splatting methods for accurate 3D reconstruction and novel view synthesis of indoor scenes.

Contact: xuqian.ren@tuni.fi

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 119

EVER WONDERED WHAT YOU ARE DREAMING ABOUT?

Riccardi E., Bottini R., Rota P.

Abstract: When we look at an image, our brain encodes it as a pattern of neural activity. Can we reverse this process and reconstruct the picture directly from the neural signals? Yes! Current approaches (1) align brain and image embeddings in a shared latent feature space, (2) leverage pretrained generative models to synthesize images. Can these frameworks also recover internally generated pictures, such as imagined scenes and dreams?

Contact: ester.riccardi@unitn.it

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 120

FROM LAND TO SEA: AI VISION FOR ILLEGAL WASTE DUMPING AND FISH DETECTION

Ricciardi Andrea Vincenzo

Abstract: Environmental monitoring is a challenging task for AI-based surveillance systems across diverse ecosystems. We address two problems:

Illegal Waste Dumping Detection (IWDD): We introduce MIVIA-IWDD-500, the first public video benchmark for IWDD, with a baseline achieving F_1=0.84, alongside WACV 2026 contest teams.

Underwater Fish Detection: We present YOLO-JUICE, a jointly optimized framework combining underwater image enhancement (UIE) and multi-scale fish detection, achieving an AP_50 of 0.89.

Contact: anricciardi@unisa.it

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 121

ADDRESSING THE WAYPOINT-ACTION GAP IN END-TO-END AUTONOMOUS DRIVING

Rodríguez-Vidal J.D., Villalonga G., Porres D., López A.M.

Abstract: End-to-End Autonomous Driving systems are grouped by the nature of their outputs: (i) waypoint-based models that predict a future trajectory [1], and (ii) action-based models that output control [2].

Recent popular benchmarks are only waypoint-based [3], which makes action-based policies harder to train and compare, slowing their progress.

To bridge this waypoint–action gap, we propose a novel, differentiable framework that rolls out predicted action sequences to their corresponding waypoint trajectories.

Contact: jdrodriguez@cvc.uab.cat

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 122

STRUCTXLIP: ENHANCING VISION-LANGUAGE MODELS WITH MULTIMODAL STRUCTURAL CUES

Ruan Z., Gao S., Kong Q., Wang Y., Cristani M.

Abstract: StructXLIP enhances vision-language retrieval by injecting multimodal structural cues into CLIP fine-tuning. It extracts edge maps as visual structure proxies and aligns them with structure-centric captions through edge-text alignment, local region-text matching, and color-edge consistency. This improves robust cross-modal retrieval.

Contact: zanxi.ruan@univr.it

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 123

ATTENTION-DISCOUNTED ADAPTIVE SAMPLER FOR MASKED DIFFUSION LANGUAGE MODELS

Sahin Y., Saikia A. R., Cevher V., Favaro P.

Abstract: Masked diffusion language models accelerate decoding by unmasking multiple tokens at once, but token-wise confidence can miss dependencies between jointly committed positions. We introduce ADAS, a training-free sampler that uses self-attention to discount the confidences of remaining candidates while preserving existing stopping rules. Across reasoning and code benchmarks, ADAS improves low-step performance with minimal runtime overhead.

Contact: yusuf.sahin@unibe.ch

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 124

WHEN NEGATION IS A GEOMETRY PROBLEM IN VISION-LANGUAGE MODELS

Sammani Fawaz, Chamiti Tzoulio, Gavrikov Paul, Deligiannis Nikos

Abstract: Joint Vision-Language Embedding models such as CLIP typically fail at understanding negation in text queries—for example, failing to distinguish "no" in the query: "a plain blue shirt with no logos". Prior work has largely addressed this limitation through data-centric approaches, fine-tuning CLIP on large-scale synthetic negation datasets. However, these efforts are commonly evaluated using retrieval-based metrics that cannot reliably reflect whether negation is actually understood. In this paper, we identify two key limitations of such evaluation metrics and investigate an alternative evaluation framework based on Multimodal LLMsas-a-judge, which typically excel at understanding simple yes/no questions about image content, providing a fair evaluation of negation understanding in CLIP models. We then ask whether there already exists a direction in the CLIP embedding space associated with negation. We find evidence that such a direction exists, and show that it can be manipulated through test-time intervention via representation engineering to steer CLIP toward negation-aware behavior without any fine-tuning. Finally, we test negation understanding on non-common image-text samples to evaluate generalization under distribution shifts

Contact: fawaz.sammani@vub.be

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 125

ROBUST AI FOR AUTOMATED INFRASTRUCTURE INSPECTION

Sánchez A., Sampedro C., Corrochano J.

Abstract: Automated inspection of critical infrastructure from aerial imagery needs perception that stays reliable in production. We present a robust computer-vision pipeline that answers each failure mode with a targeted defence: Image-Quality Assessment (IQA) filters degraded inputs, synthetic data covers rare classes, Out-Of-Distribution (OOD) detection flags unknowns, and ensembles give calibrated confidence. Currently being adopted across European TSOs and deployed in production with Spain's TSO.

Contact: asanchez@unusuals.ai

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 126

IMPROVING CONTROLLABLE GENERATION: FASTER TRAINING AND BETTER PERFORMANCE VIA X0-SUPERVISION

Sangare A., Maglo A., Chaouch M., Luvison B.

Abstract: Training controllable diffusion models can be resource intense and slow to convergence. Besides, diffusion models can be trained by supervising with the clean image (x0), the noise, or the velocity. We show, formally and experimentally, that supervising with x0 accelerates the training convergence speed and scales better with the batch size. Additionally, we propose a metric, the mean area under the convergence curve (mAUCC), to measure the training convergence speed.

Contact: amadou-siaka.sangare@universite-paris-saclay.fr

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 127

DRESS-ED: INSTRUCTION-GUIDED EDITING FOR VIRTUAL TRY-ON AND VIRTUAL TRY-OFF

Sanguigni F., Lobba D., Ren B., Cornia M., Sebe N., Cucchiara R.

Abstract: Recent advances in Virtual Try-On (VTON) and Virtual Try-Off (VTOFF) have enabled photo-realistic garment synthesis and reconstruction, yet no existing model jointly handles garment transfer and instruction-based editing in a single step. We introduce Dress-ED — the first large-scale benchmark unifying VTON, VTOFF, and text-guided garment editing, where each sample pairs garment and person images with their edited counterparts and a natural-language instruction.

Contact: sanguignifulvio@gmail.com

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 128

VIDEO UNLEARNING VIA LOW-RANK REFUSAL VECTOR

Facchiano S., Saravalle S., Migliarini M., De Matteis E., Sampieri A., Pilzer A., Rodola E., Spinelli I., Franco L. , Galasso F.

Abstract: Video generative models risk producing harmful content due to web-scale training. Existing unlearning methods are either bypassable or require costly fine-tuning. We propose a training-free weight-update framework for concept removal in video diffusion models. From five prompt pairs, a refusal vector is integrated via closed-form update; contrastive low-rank factorization ensures selective suppression without quality loss. Unsafe generations drop significantly on various benchmarks.

Contact: stefano.saravalle@uniroma1.it

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 129

A NOVEL METRIC FOR DETECTING MEMORIZATION IN GENERATIVE MODELS FOR BRAIN MRI SYNTHESIS

Scardace A., Puglisi L., Guarnera F., Battiato S., Ravì D.

Abstract: Generative models are increasingly used in medical imaging. However, recent empirical studies highlight a critical vulnerability: these models can memorize sensitive training data, posing risks of patient data disclosure. In this work, we propose DeepSSIM, a self-supervised metric for quantifying memorization in generative models. On synthetic brain MRI generated by a memorization-prone LDM, DeepSSIM outperformed state-of-the-art metrics, improving F1 score by 52% over the best existing method.

Contact: antonio.scardace@phd.unict.it

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 130

SPLATXTRACT: TRACTABLE GAUSSIAN SPLATTING VIA OPEN-WORLD REGION-OF-INTEREST EXTRACTION AND REFINEMENT

Schieber H., Kleinbeck C., Schoellig A. P. , Leutenegger S., Roth D.

Abstract: We present a task-conditioned refinement for Gaussian Splatting (GS) that enables robots or human operators to selectively extract region-of-interest (ROI). Given a GS map, SplatXtract supports ROI refinement, preserving map consistency while meeting close to real-time constraints required for interactive perception. SplatXtract decouples semantic ROIs from the GS map, allowing flexibility. integration with external and novel perception models.

Contact: hannah.schieber@tum.de

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 131

ADVERSARIAL CORRECTION AND DOMAIN-ADAPTATIVE CURRICULUM (AC-DAC) FINE-TUNING

Shen L., Edalati A., Li X., Meyer B., Gross W., Clark J. J.

Abstract: We propose AC-DAC (Adversarial Correction and Domain-Adaptive Curriculum) fine-tuning, a lightweight two-stage framework for improving trained neural networks. Misclassified samples are first adversarially corrected to form a refined source domain, followed by domain adaptation to the original domain. AC-DAC improves accuracy by over 5% on CIFAR, 1% on CINIC-10, and 1-2% on ImageNet subsets, while remaining effective for quantized models and improving adversarial robustness.

Contact: lulan.shen@mail.mcgill.ca

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 132

VLM-GUIDED CONTACT RETARGETING AND GENERATIVE PARTNER MODELING FOR DEPLOYABLE HUMANOID-HUMAN INTERACTION

Shibata Y., Amaya K., Yamazaki K., Jayanti L., Aoki Y., Isogawa M., Fragkiadaki K.

Abstract: Humanoid-human interaction suffers from severe partial observability during close contact. We present REACT, a framework for learning deployable interaction policies from human demonstrations using semantic contact structure. It features VLM-guided contact retargeting and distills policies into egocentric depth models via dynamic human rendering. Generative partner-state modeling ensures robustness under heavy occlusion. REACT successfully executes diverse interactions on a Unitree G1.

Contact: yuto19990715@gmail.com

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 133

ADAPTIVE VS. STATIC ROBOT-TO-HUMAN HANDOVER: A STUDY ON ORIENTATION AND APPROACH DIRECTION

Biagi F., Onfiani D., Silenzi S., Iani C., Biagiotti L.

Abstract: Robot-to-human handovers are object-centric: the robot freezes the object at a preset pose and the receiver must adapt. We propose an adaptive, task-aware handover that orients the object to the hand and the next task while keeping motion predictable, via a Bézier path with imposed approach directions and on-the-fly orientation alignment. With 14 participants and two objects, the adaptive policy lowered NASA-TLX workload (-18%) and blink rate and raised trust over a static baseline.

Contact: simone.silenzi@unimore.it

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 134

ELASTIC VITS FROM PRETRAINED MODELS WITHOUT RETRAINING

Simoncini Walter., Dorkenwald Michael., Blankevoort Tijmen., Snoek Cees GM., Asano Yuki M.

Abstract: Foundation Models are only released in fixed, pre-defined sizes. We propose a method to make pretrained ViTs elastic in under 5 minutes on a A100 GPU. We use self-supervised gradients, thus no labels or classification head required. We can produce strong sparse models without re-training.

Contact: w.simoncini@uva.nl

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 135

ANTHROSPHERE: HUMAN-GUIDED EGOCENTRIC VISION FOR ADAPTIVE AGENTIC XR ASSISTANCE IN INDUSTRIAL HUMAN-IN-THE-LOOP SYSTEMS

Sirocchi C., Stacchio L., Migliorelli L., Galdelli A., Mancini A.

Abstract: AnthroSphere presents a human-centered framework for Industry 5.0, bringing together egocentric perception, XR-based interaction and human-in-the-loop feedback to support operators in complex industrial tasks. Through agentic orchestration, it connects anomaly detection, action understanding, posture/safety analysis and adaptive model management toward contextual assistance and continuous improvement.

Contact: c.sirocchi@pm.univpm.it

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 136

DEXOAK: OBJECT-AWARE KINEMATIC RETARGETING FOR ROBOT TRAJECTORY GENERATION

Spinola F., Katzschmann R., Schmid C.

Abstract: Converting abundant human hand–object demonstrations into physically executable robot trajectories is key to scaling dexterous robot learning. DexOAK retargets them into contact-consistent robot references: a graph-Laplacian objective preserves object-relative contact, distilled into a feed-forward network refined with contact losses. Better kinematic trajectories lead to better downstream RL, improving success-rate over ManipTrans from 20.6% to 48.2% on OakInk-v2.

Contact: federica.spinola@inria.fr

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 137

BRIDGING IMPLICIT NEURAL AND EXPLICIT SHAPE REPRESENTATIONS

Stippel C., Engel D., Hermosilla P.,

Abstract: Every 3D representation has its own strengths. Instead of focusing on inventing or improving a representation, we focus on the translation between them: moving geometry exactly from the most optimizable but slowest form, a neural SDF, to the fastest but most rigid one, a mesh, and back.

Contact: christian.stippel@tuwien.ac.at

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 138

MARKUSHGRAPHER-2: END-TO-END MULTIMODAL RECOGNITION OF CHEMICAL STRUCTURES

Tim Strohmeyer, Lucas Morin, Gerhard Ingmar Meijer, Valéry Weber, Ahmed Nassar, Peter Staar

Abstract: Automatically extracting chemical structures from documents is essential for the large-scale analysis of the literature in chemistry. Automatic pipelines have been developed to recognize molecules represented either in figures or in text independently. However, methods for recognizing chemical structures from multimodal descriptions (Markush structures) lag behind in precision and cannot be used for automatic large-scale processing. In this work, we present MarkushGrapher-2, an end-to-end approach for the multimodal recognition of chemical structures in documents. First, our method employs a dedicated OCR model to extract text from chemical images. Second, the text, image, and layout information are jointly encoded through a Vision-Text-Layout encoder and an Optical Chemical Structure Recognition vision encoder. Finally, the resulting encodings are effectively fused through a two-stage training strategy and used to auto-regressively generate a representation of the Markush structure. To address the lack of training data, we introduce an automatic pipeline for constructing a large-scale dataset of real-world Markush structures. In addition, we present IP5-M, a large manually-annotated benchmark of real-world Markush structures, designed to advance research on this challenging task. Extensive experiments show that our approach substantially outperforms state-of-the-art models in multimodal Markush structure recognition, while maintaining strong performance in molecule structure recognition. Code, models, and datasets are released publicly.

Contact: tstrohmeyer@ethz.ch

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 139

REGIONREASONER: REGION-GROUNDED MULTI-ROUND VISUAL REASONING

Wenfang Sun, Hao Chen, Yingjun Du, Yefeng Zheng, Cees G. M. Snoek

Abstract: Large vision-language models have achieved remarkable progress in visual reasoning, yet most existing systems rely on single-step or text-only reasoning, limiting their ability to iteratively refine understanding across multiple visual contexts. To address this limitation, we introduce a new multi-round visual reasoning benchmark with training and test sets spanning both detection and segmentation tasks, enabling systematic evaluation under iterative reasoning scenarios. We further propose RegionReasoner, a reinforcement learning framework that enforces grounded reasoning by requiring each reasoning trace to explicitly cite the corresponding reference bounding boxes, while maintaining semantic coherence via a global-local consistency reward. This reward extracts key objects and nouns from both global scene captions and region-level captions, aligning them with the reasoning trace to ensure consistency across reasoning steps. RegionReasoner is optimized with structured rewards combining grounding fidelity and global-local semantic alignment. Experiments on detection and segmentation tasks show that RegionReasoner-7B, together with our newly introduced benchmark RegionDial-Bench, considerably improves multi-round reasoning accuracy, spatial grounding precision, and global-local consistency, establishing a strong baseline for this emerging research direction.

Contact: w.sun2@uva.nl

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 140

SEEING WITH PURPOSE: VISUAL INTELLIGENCE FOR GOAL-DIRECTED ROBOT MOTION

Sun B.

Abstract: How can robots use vision not only to understand the world, but to move through it? This requires identifying goals from visual input, grounding language intent into physical motion, and adapting across scenes. We present three works on learning intent-aware robot motion from vision: OpenFrontier grounds language goals into navigable frontiers, LIME predicts fine-grained camera motion from ego-video, and VidBot transfers reconstructed human motion to robot manipulation.

Contact: boysun@ethz.ch

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 141

LEARNING ROBUST GEOMETRIC REPRESENTATIONS USING SYNTHETIC STRUCTURAL DEFECTS

Szymanski W., Drwiega G., Wodzinski M.

Abstract: Introduction Deep learning models in medical imaging often rely on complete and high-quality anatomical data. In practice, clinical scans frequently contain missing structures, occlusions, or artefacts due to acquisition limitations, pathology, or preprocessing. Such imperfections can affect robustness of models trained for geometric tasks such as segmentation, reconstruction, or representation learning. Recent advances in self-supervised pretraining and foundation models highlight the importance of learning from incomplete data.

Methods In this work, we propose a configurable framework for simulating structural defects in medical imaging data. The approach follows geometric masked self-supervised learning with reconstruction-based autoencoding. The framework is implemented in PyTorch and integrated into the training loop as structured augmentation. Synthetic defects are generated from configurable geometric primitives and applied probabilistically to training volumes. The configuration enables control over defect size through ratio sampling distributions, spatial placement, probability of occurrence, and primitive selection using probability distributions. Multiple shapes can be combined to generate diverse and reproducible masked perturbations of anatomical structures. Augmentations are applied only to defect regions rather than entire volumes, making the process computationally efficient. The framework currently supports voxel grids and point clouds, with ongoing extensions to other geometric representations such as surface meshes.

Results Preliminary results are available for voxel grid and point cloud representations. The defect generation framework has been integrated into nnU-Net and Point Transformer v3 training pipelines, enabling structured defect simulation during model optimisation. Ongoing experiments focus on quantitative

evaluation of robustness across geometric learning tasks. Although regular geometric defects yield higher Dice, heterogeneous shapes reflect clinically realistic variability.

Summary Controlled simulation of structural defects represents a promising strategy for improving reliability of geometric deep learning methods in medical imaging. Potential clinical impact includes improved robustness of AI systems when handling incomplete or corrupted imaging data, supporting more reliable deployment in real-world clinical settings.

Contact: szymanski@agh.edu.pl

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 142

ROGR: RELIGHTABLE 3D OBJECTS USING GENERATIVE RELIGHTING

Jiapeng Tang, Matthew Levine, Dor Verbin, Stephan J. Garbin, Matthias Niessner, Ricardo Martin-Brualla, Pratul P. Srinivasan, Philipp Henzler

Abstract: We introduce ROGR, a novel approach that reconstructs a relightable 3D model of an object captured from multiple views, driven by a generative relighting model that simulates the effects of placing the object under novel environment illuminations. Our method samples the appearance of the object under multiple lighting environments, creating a dataset that is used to train a lighting-conditioned Neural Radiance Field (NeRF) that outputs the object's appearance under any input enviromental lighting. The lighting-conditioned NeRF uses a novel dual-branch architecture to encode the general lighting effects and specularities separately.

The optimized lighting-conditioned NeRF enables efficient feed-forward relighting under arbitrary environment maps without requiring per-illumination optimization or light transport simulation. We evaluate our approach on the established TensoIR and Stanford-ORB datasets, where it improves upon the state-of-the-art on most metrics, showcase our approach on real-world object captures.

Contact: tangjiapengtjp@gmail.com

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 143

PANOPTIC SEGMENTATION FOR TIRE DEFECT DETECTION

Tarassov E., Derville A., Ponchon F., Tilmant C., Chateau T.

Abstract: Computer vision is essential for automated industrial inspection, primarily driven by supervised deep learning algorithms. These models require large amounts of annotated data but annotating industrial defects is expensive, requires domain experts, and is often prone to ambiguities depending on the imaging techniques and defect types.

This PhD aims to develop novel methods to address annotation ambiguities and imaging limitations which can't be resolved by data cleaning or augmentation alone. This work will primarily use tire defect detection as a case study, but the methods developed will be applicable to a wide range of tasks.

The data which serves as the basis for this work is collected by an automated tire imaging system. It uses line-scan cameras and 3D profilometers to capture images and point clouds of the tire's entire surface. The resulting industrial annotated image dataset is made of 14 highly unbalanced classes.

This poster explores the use of panoptic segmentation for defect detection to specifically address annotation ambiguities of defects with poorly defined boundaries. Finally, the challenges of adapting this paradigm to industrial contexts are detailed.

Contact: elie.tarassov@doctorant.uca.fr

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 144

LOST IN THE TIME DOMAIN: SPECTRAL ADAPTERS FOR VIDEO UNDERSTANDING

Thiyakesan Ponbagavathi T., Seibold C., Roitberg A.

Abstract: Adapting image-pretrained vision foundation models to video via parameter-efficient fine-tuning is a promising alternative to full fine-tuning. We identify two failures in existing adapters: spectral blindness, focusing on frequency extremes, and order blindness, which ignores temporal direction. Frame2Freq inserts FFT adapters between frozen blocks: mid-frequency recovery fixes spectral blindness, phase preservation fixes order blindness, achieving strong performance on 5 fine-grained datasets.

Contact: thiyakesan@uni-hildesheim.de

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 145

WHEN NONLOCAL VARIATIONAL MODELS MEET ATTENTION MECHANISMS: SEEING THROUGH DARKNESS AND WATER

Torres D., Duran J., Navarro J., Sbert C.

Abstract: In this work, we present a unified perspective that connects classical nonlocal variational models with modern deep learning architectures for low-light enhancement and underwater image restoration. We first introduce a general variational formulation based on image decomposition, where iterative optimization is driven by proximal updates. This framework is extended through deep unfolding, in which proximal operators are replaced by learnable neural networks. To better preserve structural details and sharp edges, we incorporate a nonlocal gradient fidelity term that enforces alignment between the reconstructed image gradients and a reference vector field through nonlocal interactions. Building on this formulation, we address two challenging applications. For low-light enhancement, we adopt a Retinex-based decomposition and integrate cross-attention mechanisms to model nonlocal dependencies. For underwater image restoration, we employ a dehazing model incorporating a Mamba-based architecture, which efficiently captures long-range dependencies through selective state space modeling.

Contact: daniel.torres@uib.cat

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 146

CAN AUTOML HELP US FIND EFFICIENT FOREST BIOMASS ESTIMATION MODELS SUSTAINABLY?

Traoré, K. R. and Lindauer, M.

Abstract: AutoML promises to help AI practitioners design competitive models fast and consistently. Accurate Above-ground Biomass (AGB) estimation is key in monitoring important indicators of environmental sustainability, such as tracking ecological changes. It can be estimated globally using multimodal and multitemporal satellite imagery, but requires expert-defined DL models. We explore various mechanisms for a cost-effective AutoML procedure that searches for alternative, efficient AGB estimators.

Contact: krb.traore@protonmail.com

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 147

HALO-FREE ALL-IN-FOCUS AND 3D IMAGING FROM FOCAL STACKS

Ueda S., Saito H., Schmalstieg D., Mori S.

Abstract: All-in-focus (AiF) imaging from focal stacks can recover sharp color and depth beyond optical limits, but learning-based methods remain unstable due to scarce real-world data and often suffer from halo artifacts around object boundaries. We propose a volumetric formulation with multi-layer defocus reasoning, enabling robust reconstruction across different scene domains. Trained only on synthetic data, our method improves AiF image and depth recovery on real scenes and benefits novel view synthesis of close-up scenes.

Contact: shiori.ueda@keio.jp

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 148

INHIBITED SELF-ATTENTION: SHARPENING FOCUS IN VISION TRANSFORMERS

van der Wal, P.R.D., Strisciuglio, N., Azzopardi, G.

Abstract: Vision Transformers (ViTs) have demonstrated remarkable performance in computer vision tasks. However, their self-attention mechanism often diffuses focus across background regions, relying on spurious correlations rather than object-relevant cues. Inspired by inhibitory mechanisms observed in biological vision systems, we propose the Inhibited Self-Attention (ISA), a novel self-attention that integrates inhibitory signals to enhance feature selectivity and suppress spurious responses.

In contrast to conventional self-attention, which relies solely on positive attention values due to softmax normalization, our approach retains and utilizes negative attention scores to suppress irrelevant features and sharpen focus on objects of interest. Experiments across multiple datasets, including ImageNet-1k and COCO, and several robustness benchmarks demonstrate that ISA enhances object-centric selectivity, reduces shortcut reliance, and improves out-of-distribution generalization.

Our analysis of relevance maps confirms that ViTs with ISA exhibit sharper, more localized focus on object-relevant regions while reducing distractions from non-relevant (background) features, enabling more reliable models. Code is available at github.com/prdvanderwal/inhibited-self-attention.

Contact: p.r.d.van.der.wal@rug.nl

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 149

CONCEPTPOSE: TRAINING-FREE ZERO-SHOT OBJECT POSE ESTIMATION USING CONCEPT VECTORS

Kuang L., Velikova Y., Saleh M., Zaech JN., Paudel D., Busam B.

Abstract: Object pose estimation is a fundamental task in computer vision and robotics, yet most methods require extensive, dataset-specific training. Concurrently, large-scale vision language models show remarkable zero-shot capabilities. In this work, we bridge these two worlds by introducing ConceptPose, a framework for object pose estimation that is both training-free and model-free. ConceptPose leverages a vision-language-model (VLM) to create open-vocabulary 3D concept maps, where each point is tagged with a concept vector derived from saliency maps. By establishing robust 3D-3D correspondences across concept maps, our approach allows precise estimation of 6DoF relative pose. Without any object or dataset-specific training, our approach achieves state-of-the-art results on common zero shot relative pose estimation benchmarks, outperforming the strongest baseline by a relative 62

Contact: dani.velikova@tum.de

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 150

VENI: VARIATIONAL ENCODER FOR NATURAL ILLUMINATION

Walker P., Gardner J. A. D., Ardelean A., Smith W. A. P., Egger B.

Abstract: Inverse rendering is ill-posed, illumination priors can help simplify it. We propose a rotation-equivariant variational autoencoder that models natural illumination on the sphere. To preserve the SO(2)-equivariance of environment maps, we use a novel Vector Neuron Vision Transformer as encoder and a rotation-equivariant neural field as decoder. Compared to previous methods, our variational autoencoder enables smoother interpolation in latent space and offers a more well-behaved latent space.

Contact: paul@wwwalker.de

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 151

HOW TO TRAIN A SOTA FOUNDATIONAL VLM

Han Wang

Abstract: Training a state-of-the-art foundational Vision-Language Model is increasingly a data engineering problem rather than a single-model-design problem. From the hands-on engineering perspective of us, the key shift is from human-in-the-loop annotation to model-in-the-loop data construction, where strong models are used to generate, score, filter, rewrite, and expand multimodal training data at scale. This poster summarizes a benchmark-driven data flywheel: candidate data recipes are first validated through controlled training runs and fixed evaluation suites; only transformations that improve benchmark performance are retained. We further categorize data into OCR and grounding data produced by specialist expert models, general VQA data labeled by frontier VLMs, instruction/video/table-chart data refined through automatic filtering and rewriting, and context-augmented annotation for injecting wiki-like world knowledge. In this strategy, auxiliary context is provided to the annotator model to obtain correct answers, but removed from the final training sample to prevent shortcut learning. The case studies demonstrate that scalable, validated, and benchmark-aligned data curation directly translates into stronger VLM performance.

Contact: han.wang@unitn.it

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 152

SELF-SUPERVISED LEARNING BASED ON TRANSFORMED IMAGE RECONSTRUCTION FOR EQUIVARIANCE-COHERENT FEATURE REPRESENTATION

Qin Wang, Alessio Quercia, Benjamin Bruns, Abigail Morrison, Hanno Scharr, Kai Krajsek

Abstract: Self-supervised learning (SSL) methods have achieved remarkable success in learning image representations allowing invariances in them — but therefore discarding transformation information that some computer vision tasks actually require. While recent approaches attempt to address this limitation by learning equivariant features using linear operators in feature space, they impose restrictive assumptions that constrain flexibility and generalization. We introduce a weaker definition for the transformation relation between image and feature space denoted as equivariance-coherence. We propose a novel SSL auxillary task that learns equivariance-coherent representations through intermediate transformation reconstruction, which can be integrated with existing joint embedding SSL methods. Our key idea is to reconstruct images at intermediate points along transformation paths, e.g. when training on 30° rotations, we reconstruct the 10° and 20° rotation states. Reconstructing intermediate states requires the transformation information used in augmentations, rather than suppressing it, and therefore fosters features containing the augmented transformation information. Our method decomposes feature vectors into invariant and equivariant parts, training them with standard SSL losses and reconstruction losses, respectively. We demonstrate substantial improvements on synthetic equivariance benchmarks while maintaining competitive performance on downstream tasks requiring invariant representations. The approach seamlessly integrates with existing SSL methods (iBOT, DINOv2) and consistently enhances performance across diverse tasks, including segmentation, detection, depth estimation, and video dense prediction. Our framework provides a practical way for augmenting SSL methods with equivariant capabilities while preserving invariant performance.

Contact: qi.wang@fz-juelich.de

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 153

PAWS: PERCEPTION OF ARTICULATION IN THE WILD AT SCALE FROM EGOCENTRIC VIDEOS

Wang Y., Miao Y., Zhao W., Yang W., Wang Z., Pajarinen J., Van Gool L., Paudel D., Kannala J., Wang X., Solin A.

Abstract: Articulation perception recovers the motion and structure of articulated objects (e.g., drawers), key to 3D scene understanding. Existing methods need supervised training with 3D annotations, limiting scalability. We propose PAWS, a training-free method that extracts object articulations directly from hand–object interactions in large-scale in-the-wild egocentric videos. On HD-EPIC and Arti4D it performs competitively and benefits downstream articulation prediction and robot manipulation.

Contact: yihao.wang@aalto.fi

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 154

BENCHMARKING AND BOOSTING PHYSICAL REASONING FROM CONDITIONAL VIDEO OBSERVATIONS

Fanyue Wei, Kai Xu, Yizhuo Zhang, Pengzhan Sun, Junbin Xiao, Angela Yao

Abstract: We present {CoPhyBench}, a {Co}nditional {Phy}sics-based reasoning {Bench}mark for evaluating Video-LLMs. CoPhyBench probes physics reasoning from three perspectives: 1) Prediction, to predict future events from observable cues to assess real-world causality; 2) Physical Calculation, to estimate times and positions by translating visual conditions into quantitative estimates of event evolutions; and 3) Counterfactual Reasoning, to infer outcomes under hypothetical changes, to test generalizable physical understanding beyond superficial correlations. We curate a high-quality dataset of 1,300 carefully verified QA pairs grounded in 232 diverse real-world physics videos, spanning kinematics and dynamics. We further propose an uncertainty-guided finetuning strategy based on QA pairs of conditional observations. Experiments on leading Video-LLMs show strong causal predictions but substantial gaps in precise calculations and counterfactual reasoning. Overall, results underscore the difficulty of moving from semantic alignment to physics-grounded reasoning. This motivates a call for new training paradigms to incorporate physics reasoning. Our dataset and resources will be released.

Contact: wfanyue@gmail.com

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 155

RADIANCE FIELDS FOR ROBOTICS

Wilder-Smith M, Patil V, Morkva S, Bhardwaj A, Mittal M, Tateno K, Niemeyer M, Oechsle M, Tombari F, Hutter M

Abstract: Autonomous robots need real-time spatial awareness. To overcome traditional radiance fields' scaling, dynamic, and semantic limits, we present the Radiance Fields for Robotics ecosystem: DiskChunGS for kilometer-scale embedded mapping; MOSIAC-GS for rapid 4D dynamic reconstruction; Gaussians All The Way Down for fast, dense semantic embedding; ViserDex showing 3DGS for RL-based manipulation; and an XR Teleoperation toolkit for native headset data streaming across any robot or deployment.

Contact: mwilder@ethz.ch

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 156

BREWING STRONGER FEATURES: DUAL-TEACHER DISTILLATION FOR MULTISPECTRAL EARTH OBSERVATION

Wolf Filip, Rolih Blaž, Čehovin Zajc Luka

Abstract: Foundation models are transforming Earth Observation (EO), yet the diversity of EO sensors and modalities makes a single universal model unrealistic. Multiple specialized EO foundation models (EOFMs) will likely coexist, making efficient knowledge transfer across modalities essential. Most existing EO pretraining relies on masked image modeling, which emphasizes local reconstruction but provides limited control over global semantic structure. To address this, we propose a dual-teacher contrastive distillation framework for multispectral imagery that aligns the student's pretraining objective with the contrastive self-distillation paradigm of modern optical vision foundation models (VFMs). Our approach combines a multispectral teacher with an optical VFM teacher, enabling coherent cross-modal representation learning. Experiments across diverse optical and multispectral benchmarks show that our model adapts to multispectral data without compromising performance on optical-only inputs, achieving state-of-the-art results in both settings, with average improvements of 3.64 percentage points in semantic segmentation, 1.2 in change detection, and 1.31 in classification. This demonstrates that contrastive distillation provides a principled and efficient approach to scalable representation learning across heterogeneous EO data sources.

Contact: filip.wolf@fri.uni-lj.si

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 157

FINER: MLLMS HALLUCINATE UNDER FINE-GRAINED NEGATIVE QUERIES

Xiao, Rui; Kim, Sanghwan; Xian, Yongqin; Akata, Zeynep; Alaniz, Stephan

Abstract: Multimodal large language models (MLLMs) struggle with hallucinations, particularly with fine-grained queries, a challenge underrepresented by existing benchmarks that focus on coarse image-related questions. We introduce FInegrained NEgative queRies (FINER), alongside two benchmarks: FINER-CompreCap and FINER-DOCCI. Using FINER, we analyze hallucinations across four settings: multi-object, multi-attribute, multi-relation, and "what" questions. Our benchmarks reveal that MLLMs hallucinate when fine-grained mismatches co-occur with genuinely present elements in the image. To address this, we propose FINER-Tuning, leveraging Direct Preference Optimization (DPO) on FINER-inspired data. Finetuning four frontier MLLMs with FINER-Tuning yields up to 24.2% gains (InternVL3.5-14B) on hallucinations from our benchmarks, while simultaneously improving performance on eight existing hallucination suites, and enhancing general multimodal capabilities across six benchmarks. Code, benchmark, and models are available at https://explainableml.github.io/finer-project/.

Contact: rui.xiao@tum.de

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 158

VIBE SPACE: CREATIVELY CONNECTING AND EXPRESSING VISUAL CONCEPTS

Xu K., Yang H., Lu A., Grossberg M.D., Bai Y., Shi J.

Abstract: Creating new visual concepts often requires connecting distinct ideas through the most relevant shared attributes — the vibe. We introduce Vibe Blending, a novel task for generating coherent and meaningful hybrids that reveals these shared attributes between images. Achieving such blends is challenging for current methods, which struggle to identify and traverse nonlinear paths linking distant concepts in latent space. We propose Vibe Space, a hierarchical graph manifold that learns low-dimensional geodesics in feature spaces like CLIP, enabling smooth and semantically consistent transitions between concepts. To evaluate creative quality, we design a cognitively inspired framework combining human judgments, LLM reasoning, and a geometric path-based difficulty score. Vibe Space produces blends that humans consistently rate as more creative and coherent than current methods.

Contact: katexu@seas.upenn.edu

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 159

MODEL-AGNOSTIC POST-HOC PRUNING AND OPTIMIZATION FOR SINGLE-VIEW FEED-FORWARD 3D GAUSSIAN SPLATTING

Yagawa R., Cheng H., Schmalstieg D., Saito H., Mori S.

Abstract: Single-view feed-forward 3DGS predicts a fixed number of Gaussians per ray, causing severe spatial redundancy. Existing compaction methods focus exclusively on multi-view settings. Because they rely on cross-view consistency for geometric fusion or learned allocation, they fail in single-view setups. We propose a model-agnostic post-hoc sparsification pipeline. Decoupled from the base architecture, it seamlessly integrates with any off-the-shelf single-view estimator to resolve redundancy.

Contact: rintoyagawa@keio.jp

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 160

DEXNINJA: LEARNING ROBUST DEXTEROUS CUTTING POLICY WITH A REAL-TO-SIM-TO-REAL DATA ENGINE

Lou H., Yang R., Zhong W., Liu C., Liu Y., Liu W., Ma W., Xia J., Wu K., Paudel D.P., Van Gool L., Zhao H., Li Y.

Abstract: DexNinja is a real-to-sim-to-real framework for learning robust dexterous food cutting from a few real demonstrations. It reconstructs real objects, randomizes category-level physical parameters, and augments training with a differentiable simulator that couples robot dynamics, tactile contact, and deformable cutting. On food slicing, tens of demos plus simulated episodes achieve over 60% success, improve real-only training by over 20%, and generalize to unseen shapes and sizes.

Contact: runyi.yang@insait.ai

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 161

STAR: SEAMLESS SPATIAL-TEMPORAL AWARE MOTION RETARGETING WITH PENETRATION AND CONSISTENCY CONSTRAINTS

Yang X., Wang Q., Yang J., Slabaugh G., Yuan S.

Abstract: STaR is a spatio-temporal skinned motion retargeting framework that preserves motion semantics while improving geometric plausibility and temporal coherence. It combines dense shape representations, a limb penetration constraint, and a trajectory-level temporal consistency loss to reduce interpenetration and jitter across characters with diverse body shapes.

Contact: xiaohang.yang@qmul.ac.uk

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 162

BOOTSTRAPPING ARTICULATED 3D RECONSTRUCTION FROM IMAGES

Zadrozny J., Mac Aodha O., Bilen H.

Abstract: Articulated 3D reconstruction typically requires massive datasets. Our iterative framework uses only unannotated 2D images and a template mesh. We align it to weakly predicted dual point maps yielding synthetic data to self-refine our predictor. We approach fully-supervised baselines and outperform generic models on complex articulations. Our framework guarantees topological correctness and scales effortlessly to novel categories by drastically reducing 3D data requirements.

Contact: Jakub.Zadrozny@ed.ac.uk

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 163

EGONIGHT: TOWARDS EGOCENTRIC VISION UNDERSTANDING AT NIGHT WITH A CHALLENGING BENCHMARK

Deheng Zhang*, Yuqian Fu*, Runyi Yang, Yang Miao, Tianwen Qian, Xu Zheng, Guolei Sun, Ajad Chhatkuli, Xuanjing Huang, Yu-Gang Jiang, Luc Van Gool, Danda Pani Paudel

Abstract: Most egocentric vision benchmarks focus on daytime scenarios, overlooking the low-light conditions common in real-world applications. We introduce EgoNight, the first compr ehensive benchmark for nighttime egocentric vision, centered on visual question answering (VQA). EgoNight features day–night aligned videos, including both Blender-rendered synthetic data and real-world recordings, enabling higher-quality night annotations and direct comparison across illumination conditions. Using these paired videos, we construct EgoNight-VQA with a day-augmented night auto-labeling engine, extensive human verification, and double-checked QA pairs. The dataset contains 3658 QA pairs across 90 videos, covering 12 QA types and requiring over 300 hours of annotation. Evaluations of state-of-the-art multimodal large language models reveal substantial performance drops from day to night, highlighting the difficulty of reasoning under low-light conditions. EgoNight also includes two auxiliary tasks: day–night correspondence retrieval and nighttime egocentric depth estimation. We hope EgoNight provides a strong foundation for illumination-robust egocentric vision research. Code and data are available at https://dehezhang2.github.io/EgoNight/.

Contact: deheng.zhang@insait.ai

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 164

TOWARDS IN-THE-WILD EGOCENTRIC 3D HAND-OBJECT POSE ESTIMATION

Bansal S. , Zhu Z. , Tripathi S. , Zhao J. , Black M. , Damen D.

Abstract: Estimating 3D hand-object pose is bottlenecked by simplistic in-lab datasets and methods ignoring hand pose priors. We address this via: (1) EPIC-Contact, a dataset of 2.3K in-the-wild videos with 3D hand-object poses and bijective contact annotations; (2) HOPFormer, a feed-forward model using strong hand priors to explicitly capture hand-object interactions, achieving state-of-the-art 3D hand-object pose estimation across multiple datasets.

Contact: jiahe.zhao@bristol.ac.uk

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 165

HIERARCHICAL STEP DECOMPOSITION FOR TRAINING-FREE ONLINE VIDEO STEP GROUNDING

Zhou L., Zanella L., Rota P., Mancini M., Ricci E.

Abstract: Video Step Grounding (VSG) identifies which procedural steps occur in a video and localizes when they happen. We propose Hierarchical Step Decomposition (HSD), a training-free online framework that perform online VSG by constructing semantically coherent events from streaming video and refines uncertain predictions by decomposing candidate steps into fine-grained sub-steps, improving recognition, localization, and efficiency.

Contact: long.zhou@unitn.it

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 166

EVENT-AIDED SHARP RADIANCE FIELD RECONSTRUCTION FOR FAST-FLYING DRONES

Zou R., Cannici M., Scaramuzza D.

Abstract: Fast-flying drones are valuable for inspection and exploration tasks because they can cover large areas efficiently within limited battery life. However, fast flight causes motion-blurred images and noisy pose estimates, which break standard 3D reconstruction. We propose a unified framework that jointly models motion blur and refines pose estimation using frames and events. We achieve sharp scene reconstruction and recover accurate camera trajectories without requiring ground-truth supervision.

Contact: zou@ifi.uzh.ch

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 167

VGGSOUNDER: AUDIO-VISUAL EVALUATIONS FOR FOUNDATION MODELS

Daniil Zverev, Thaddäus Wiedemer, Ameya Prabhu, Matthias Bethge, Wieland Brendel, A. Sophia Koepke

Abstract: The emergence of audio-visual foundation models underscores the importance of reliably assessing their multi-modal understanding. The VGGSound dataset is commonly used as a benchmark for evaluation audio-visual classification. However, our analysis identifies several limitations of VGGSound, including incomplete labelling, partially overlapping classes, and misaligned modalities. These lead to distorted evaluations of auditory and visual capabilities. To address these limitations, we introduce VGGSounder, a comprehensively re-annotated, multi-label test set that extends VGGSound and is specifically designed to evaluate audio-visual foundation models. VGGSounder features detailed modality annotations, enabling precise analyses of modality-specific performance. Furthermore, we reveal model limitations by analysing performance degradation when adding another input modality with our new modality confusion metric.

Contact: zverev@in.tum.de

Presentation Type: Poster

Date: Tuesday 7 July 2026

Time: 21:30

Poster Session: 2

Poster Number: 168