Part B-1

1. Excellence #@REL-EVA-RE@#

1.1 Quality and pertinence of the project's research and innovation objectives (and the extent to which they are ambitious, and go beyond the state of the art)

Introduction: Accurate visual *scene understanding* is vital in today's technology-driven world, particularly through *semantic segmentation*, which groups pixels based on their semantic categories. Applications span across *robotics*, *autonomous vehicles*, *assistive technologies*, *satellite image analysis*, and *medical imaging*. For instance, in robotics, precise scene understanding allows robots to navigate and perform tasks autonomously in various environments. Autonomous vehicles rely on accurate segmentation for detecting and categorizing road elements and pedestrians, enhancing navigation and safety. Assistive technologies for visually impaired individuals and elderly people, such as visual and mobility aids (e.g., smart canes), use segmentation to provide real-time descriptions and guidance, enhancing users' functionality, independence, and quality of life. In satellite image analysis, segmentation is vital for environmental monitoring and disaster management, while in healthcare, it aids in diagnosing diseases by identifying pathological regions. The significance of these applications is reflected in *European initiatives* such as Horizon Europe and the Digital Europe Programme, which focus on advancing artificial intelligence (AI) to tackle societal challenges, improve quality of life, and drive economic growth.

Project Overview: The proposed research project, **RAVIOLI** (**R**etrieval-**A**ugmented **VI**sion-Language Models for **O**pen-vocabulary **L**ocalizat**I**on), aims to significantly advance the field of segmentation by innovatively integrating *retrieval-based predictions* from a memory with the *original predictions* of a vision-language model (VLM) through a learnable **fusion model**. Addressing a critical gap in existing methods, which often struggle to adapt to new or complex classes and domains, RAVIOLI seeks to enhance the *accuracy*, *adaptability*, and *granularity* of models. The project will be hosted by the Visual Recognition Group (VRG) at the Czech Technical University in Prague (CTU) under the supervision of **Prof. Giorgos Tolias**. The fellow, **Bill Psomas**, with a strong background in computer vision (CV) and deep learning (DL), is well-equipped to lead this research, which will further supported by a secondment at AImageLab, University of Modena and Reggio Emilia (UNIMORE).

State-of-the-Art (SOTA): While semantic segmentation methods have progressed[1], they remain limited by the need for labeled datasets with predefined limited classes, making expansion to new categories costly and labor-intensive. This *closed-ended* approach contrasts with human scene understanding, which can adapt to dynamic and varied environments. VLMs like CLIP[2], trained on easily-acquired image-text pairs, have demonstrated remarkable capabilities in image-level recognition tasks, enabling *zero-shot* and *open-vocabulary recognition*[20]. However, these models struggle with pixel-level tasks like semantic segmentation.

Recent works attempt to leverage VLMs for *open-vocabulary segmentation*. Some methods[3,4,5] first group pixels with a region-proposal network to reduce the complexity and then make label predictions for regions based on the open-vocabulary. These methods carry the limitations of region-proposal networks which are trained with a closed-

NOTE: Publications in orange are by the ER & Supervisors. [1]Minaee et al. Image segmentation using deep learning: A survey. TPAMI, 2021. /[2]Radford et al. Learning transferable visual models from natural language supervision. ICML, 2021. /[3]Ghiasi et al. Scaling Open-Vocabulary Image Segmentation with Image-Level Labels. ECCV, 2022. /[4]Liang et al. Open-Vocabulary Semantic Segmentation with Mask-adapted CLIP. CVPR, 2023. /[5]Zhang et al. The role of virtual try-on technology in online purchase decision from consumers' aspect. Internet Research, 2019. /[6]Xu et al. GroupViT: Semantic Segmentation Emerges from Text Supervision. CVPR, 2022. /[7]Dong et al. MaskCLIP: Masked Self-Distillation Advances Contrastive Language-Image Pretraining. CVPR, 2023. /[8]Cha et al. Learning to Generate Text-grounded Mask for Open-world Semantic Segmentation from Only Image-Text Pairs. CVPR, 2023. [10]Rombach et al. High-Resolution Image Synthesis with Latent Diffusion Models. CVPR, 2022. /[11]Jia et al. Visual prompt tuning. ECCV, 2022. /[12]Luo et al. SegCLIP: Patch Aggregation with Learnable Centers for Open-Vocabulary Semantic Segmentation. ICML, 2023. /[13]Mukhoti et al. Open Vocabulary Semantic Segmentation with Patch Aligned Contrastive Learning. CVPR, 2023. /[14]Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR, 2021. /[15]Gal et al. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. ICLR, 2023. /[16]Bousselham et al. Grounding Everything: Emerging Localization Properties in Vision-Language Transformers. CVPR, 2024. /[17]Graves et al. Neural turing machines. arXiv. /[18]Khandelwal et al. Generalization through memorization : Nearest neighbor language models. ICLR, 2020. /[19]Wu et al. Memorizing transformers. ICLR, 2022. /[20]Stojnic et al. Label Propagation for Zero-shot Classification with Vision-Language Models. CVPR, 2024. /[21]Long et al. Retrieval augmented classification for long-tail visual recognition. CVPR 2022. /[22]Conti et al. Vocabulary-free image classification. NeurIPS, 2023. /[23]Iscen et al. Retrieval-enhanced contrastive vision-text models. ICLR, 2024. /[24]Bertrand et al. Test-time Training for Matching-based Video Object Segmentation, NeurIPS 2023. /[25]Yao et al. DetCLIPv3: Towards versatile generative open-vocabulary object detection. CVPR, 2024. /[26]Tselentis et al. The usefulness of artificial intelligence for safety assessment of different transport modes. Accid. Anal. Prev., 2024. /[27]Xie eta al. RA-CLIP : Retrieval augmented contrastive language-image pre-training. CVPR, 2023. /[28]Sun et al. Going denser with open-vocabulary part segmentation. ICCV, 2023. /[29]Aoki et al. TAG: Guidance-free open vocabulary semantic segmentation. arXiv. /[30]Caffagni et al. Wiki-LLaVA: Hierarchical retrieval-augmented generation for multimodal LLMs. CVPR, 2024. /[31]Sarto et al. Towards retrieval-augmented architectures for image captioning. ACM TOMM, 2024. /[32]Barsellotti et al. Training-free open-vocabulary segmentation with offline diffusion-augmented prototype generation. CVPR, 2024. /[33]Barsellotti et al. FOSSIL: Free open-vocabulary semantic segmentation through synthetic references retrieval. WACV, 2024. /[34]Venkataramanan et al. It takes two to tango: Mixup for deep metric learning. ICLR, 2022. /[35]Psomas et al. Keep it simpool: Who said supervised tranformers suffer from attention deficit? ICCV, 2023. /[36]Kakogeorgiou et al. What to Hide from Your Students: Attention-Guided Masked Image Modeling. ECCV, 2022. /[37]Psomas et al. Composed Image Retrieval for Remote Sensing. IGARSS, 2024. /[38]themaximalists.substack.com/p/brag.

ended and smaller vocabulary. Other works modify the ViT[14]-based architecture of the VLM by adding segmentation blocks[6,12] or by adding a decoder[8,13], and fine-tune on image-text pairs. Their training requires heavy dataset filtering, causing it to lose its free-form characteristic. A third line of works leverages the inherent capabilities of VLMs and repurposes the trained model without the need of retraining[7,16]. Despite performance improvements in open-vocabulary settings, there is still a large gap compared to the closed-ended settings. Additionally, these methods rely on the vocabulary of the VLM, which although includes many generic concepts, it lacks fine-grained entities[25,28] that are rare or absent from the image-text pre-training dataset.

To develop more *flexible* and *adaptive* models that transcend the vocabulary or, in general, knowledge constraints imposed by the pre-training stage, *retrieval-augmented* methods are proposed. These methods argue that **not all world knowledge can be compressed into a model's parameters**, necessitating the ability to retrieve information from an external *memory*. This concept dates back to the Neural Turing Machines[17], while recently, in the natural language processing (NLP) domain, retrieval-augmented generation (RAG)[18,19] has shown significant promise. It allows models to dynamically access external structured data, facilitating continuous knowledge updates and enabling the seamless integration of domain-specific information. Inspired by these advancements in NLP, recent works in CV leverage a memory to improve classification[21-23], visual question answering (VQA)[30], image captioning[31], and video understanding[24]. Notably, RECO[23] refines the VLM representations with knowledge retrieved from an external memory. However, while RECO boosts performance for classification tasks, it does not extend to dense tasks, like semantic segmentation. Similarly, RA-CLIP[27] and TAG[29] employ a memory mechanism for dense prediction tasks. Nevertheless, they follow the aforementioned two-stage paradigm, decouple recognition from segmentation, which carries the limitations of the first stage. Additionally, their combination of predictions is simplistic and hand-designed. This raises the question of *what the proper way is to leverage a memory along with a VLM in a retrieval-augmented fashion to combine predictions for open-vocabulary segmentation.*

Problem formulation: RAVIOLI aims to develop a sophisticated method for open-vocabulary segmentation that effectively combines and integrates *retrieval-based predictions* from a *memory* with *original predictions* from a VLM. The need for a learnable **fusion function (model)** is identified, termed the *retrieval-augmented predictor* (RAP), which will take as input the within image spatial information, the zero-shot prediction of the VLM, and information retrieved from the memory, such as images, masks, labels, and their corresponding strengths. The fusion model seeks to achieve: **(i) Segmentation enhancement:** To perform segmentation by *leveraging* and *enhancing* the inherent capabilities of the VLM, ensuring that the segmentation is accurate and contextually aware, and **(ii) Granularity flexibility:** To support flexible *granularity*, enabling both detailed part-level spatial segmentation and hierarchical label assignment, reflecting the recognition processes of human perception.

Aim and objectives: We aim to enhance the segmentation capabilities of VLMs by integrating a structured memory. This memory will store examples, including *weakly labeled*, *fully labeled* or *synthetically generated*, to enrich the localization ability and contextual understanding of the VLM. By developing a fusion function on top of the VLM, we will enable the model to **adapt to new or complex domains** that require detailed segmentation **without the need for retraining**. This will be achieved through three measurable and realistic objectives:

- **Objective 1 (O1): Develop the Retrieval-Augmented Predictor**. Design and develop the retrieval-augmented predictor (RAP), focusing on optimizing its inputs and understanding what should be stored in the memory. This involves exploring the trade-off between memory size and the compactness of the stored representations, ensuring effective retrieval for segmentation and label assignment.
- **Objective 2 (O2): Use generative AI to enrich the memory**. Explore the synergy with large-scale text-to-image *generative models*. This approach introduces new challenges and opportunities, enabling multi-modal capabilities and providing flexibility through the creation of synthetic images on demand.
- **Objective 3 (O3): Support granularity flexibility**. The third objective is to ensure that the RAP supports flexible *granularity*, enabling both detailed *part-level spatial segmentation* (e.g., cat's ear) and *hierarchical label assignment* (e.g., Egyptian cat → cat → mammal → animal). This flexibility will allow to adapt to various levels of segmentation detail and align with human-like multi-granular recognition.

Progress beyond the SOTA: RAVIOLI represents a significant leap beyond the current SOTA in open-vocabulary segmentation. By developing the RAP, it will introduce a novel, learnable fusion model that seamlessly integrates retrieval-based knowledge with VLM predictions, **addressing the limitations of existing methods** that rely on *decoupled* and *simplistic* prediction combinations. This approach will not only enhance segmentation accuracy, as demonstrated through rigorous evaluation, but also ensure fine-grained, contextually aware label assignment, maintaining the zero-shot and open-vocabulary capabilities inherent in VLMs. Moreover, by combining *real* and

synthetic images in memory, it will open new avenues for multi-modal adaptability and dynamic knowledge integration. Importantly, **there has been no similar attempt to learn a fusion model with these properties in any open-vocabulary dense task**, such as segmentation, making our approach truly pioneering. The ambitious scope of this project lies in its aim to create a tailored, flexible, robust, and scalable solution that will redefine the capabilities of vision-language models, setting a new standard in the field of open-vocabulary segmentation.

1.2 Soundness of the proposed methodology (including interdisciplinary approaches, consideration of the gender dimension and other diversity aspects if relevant for the research project, and the quality of open science practices).

1.2.1 Methodology

To accomplish the objectives of RAVIOLI, a series of steps will be undertaken, forming WP1, WP2, and WP3:

WP1: Development of the Retrieval-Augmented Predictor: The main goal of WP1 is to design and develop the RAP, a fusion model integrating a structured memory with VLMs to enhance their segmentation capabilities. WP1 **focuses on optimizing the inputs, architecture, training and evaluation of RAP**. The first step involves creating a comprehensive dataset (T1.1) that will serve as the foundation for developing and evaluating RAP. Initially, this will involve using the image-text dataset on which the VLM was originally trained. Starting from the image-text pairs and inspired by NLP techniques, the captions will be broken down into words, filtered and aligned with image patches to create pixel-level pseudo-labels (pseudo-masks). Additionally, the option of incorporating manually segmented images from existing datasets will be explored to further enrich the memory with fully labeled examples. Next, the structured memory will be developed by extracting and storing visual representations of image patches along with their corresponding semantic labels or word descriptions (T1.2). This memory will be organized to facilitate efficient retrieval during segmentation, relying on stored examples to support the RAP in making predictions. The core tasks of WP1 involve designing the fusion model architecture (T1.3), developing a voting mechanism (T1.4), and exploring improved positional encoding techniques (T1.5). The fusion model will be designed to combine visual evidence (patches) retrieved from the memory with the original VLM predictions in a *class-generic* way, enabling it to be applied to any unseen class. *Transformer-based* mechanisms will be leveraged to process the similarities from memory patches collectively and integrate them with the original VLM predictions (T1.3). A voting mechanism will be developed to allow test image patches to *collect votes* from similar patches in the memory, with each vote carrying information about the patch's similarity to memory examples and its spatial layout (T1.4). This mechanism will enable RAP to leverage the contextual information stored in the memory, ensuring informed and accurate predictions. Improved *positional encoding* techniques (T1.5) will be explored to address the spatial complexity inherent in this process, as traditional positional embeddings have shown limited performance gains. These techniques will help the model better integrate spatial and contextual information from both the image and the memory. Following the architectural design, the fusion model will be trained (T1.6) on the created dataset. Various training strategies will be explored, utilizing either the weakly labeled examples generated from grounded image-text pairs, fully labeled examples from benchmark segmentation datasets, or a combination of both. Finally, the RAP will be evaluated (T1.7) in a class-generic way, with a particular emphasis on its ability to maintain segmentation accuracy in *unfamiliar* contexts. This evaluation will ensure that the RAP achieves its goal of enhancing the segmentation capabilities of VLMs, making them more *flexible*, *adaptable*, and *robust*. With extensive **expertise in developing multi-modal models**, particularly in vision-language tasks[37], and **experience in designing models for retrieval**[34-36] **and dense prediction tasks**[35,36], the ER is well-positioned for WP1.

WP2: Generative AI for memory enrichment: The primary goal of WP2 is to leverage text-to-image generative models to *enrich* the structured memory of WP1 with *synthetic* images. WP2 will address the challenges and opportunities associated with **generating images for specific concepts, words, or fine-grained classes**. This will enable the memory to cover a broader range of categories, including those that are underrepresented in real-world datasets. The first task involves synthesizing images for specific concepts using generative models such as Stable Diffusion[10] (T2.1). This will include the exploration of *instance-conditioned generative methods*[15], which can create high-quality images that accurately represent the desired concepts and on which the **ER has experience due to ongoing work**. Following the image synthesis, the next crucial step is to develop pixel-level pseudo-masks for the generated images (T2.2). Inspiration will be drawn from the pseudo-labeling mechanism used in WP1 and will be further expanded by leveraging the visual-language *correspondences* extracted from the generative model. This will ensure that the synthetic images are not only visually accurate but also come with precise pixel-level labels, making them suitable for enriching the memory. The focus will shift to integrating synthetic examples into the memory, creating a hybrid memory that synergizes real and synthetic data (T2.3). *Prompt-tuning* techniques[11] will be explored to enhance the synergy between the real and synthetic memories. Finally, the enriched hybrid memory

will be tested to assess the impact of synthetic data on segmentation performance (T2.4), especially for underrepresented or unseen categories. This evaluation will involve comparing the performance of the RAP using only real data versus the hybrid memory. A secondment with **Prof. Rita Cucchiara** at UNIMORE will directly support WP2, leveraging their expertise in diffusion-augmented VLMs for open-vocabulary segmentation[32,33].

WP3: Granularity flexibility in recognition and segmentation: The primary goal of WP3 is to extend the capabilities of the RAP by **enabling it to support both hierarchical label assignment**[25] and **detailed part-level segmentation**[28]. Objects often need to be recognized and segmented not only as whole entities but also at *finer levels of granularity*, such as distinguishing specific *parts* (e.g., cat's ear) or understanding *hierarchical labels* (e.g., Egyptian cat → cat → mammal → animal). This is essential for practical applications such as robotics, behavior analysis, and other domains requiring detailed visual comprehension. WP3 will first extend the structured memory to incorporate a hierarchical structure of labels, organizing concepts in a tree-like structure where the root represents a generic category, and nodes further down represent more specific, fine-grained categories (T3.1). In parallel, a dedicated part-level memory will be developed to store specific object parts, such as "ear", "nose", or "wheel" (T3.2). This memory will support *"part-transfer"* capabilities, allowing the RAP to apply knowledge of parts from one category (e.g., a "cat's ear") to semantically similar but distinct categories (e.g., an "Egyptian cat's ear"), even if the exact part has not been seen during training. The design of the RAP will be extended to *synergize* with these memory structures, enabling the model to perform both part segmentation and hierarchical label assignment in a seamless manner (T3.3). Additionally, the exploration will include the use of real and synthetic data within these memories, drawing inspiration from the generative approaches developed in WP2, to enhance the robustness and generalizability of the model (T3.4). Finally, these enhancements will be evaluated, focusing on the RAP's ability in part segmentation and hierarchical label assignment across a broad spectrum of categories (T3.5).

1.2.2 Interdisciplinary aspect

RAVIOLI spans distinct fields within Computer Science (CS), including CV, Representation Learning, NLP, and Generative AI, **acting as a bridge** that facilitates the dissemination of innovations across these fields. This enriches the project by integrating diverse perspectives and broadening its impact. This interdisciplinary approach further enhances the project's reach by applying its findings to a wider range of applications within CS.

1.2.3 Gender and other diversity aspects

While RAVIOLI does not directly involve human subjects or the distinction between genders in the recognition tasks, efforts will be made to include *diverse* datasets that represent a wide range of scenarios, ensuring that the developed models are *robust* and *unbiased*. This includes the careful consideration of images across various demographics and environments to avoid any implicit biases in the retrieval-augmented VLMs. Additionally, we will actively seek the *participation* of *women* and *underrepresented groups* in all project activities, such as workshops, lectures, and collaborations, to promote *inclusivity* in the research process. These efforts align with the broader goals of the project, particularly its aim to develop inclusive technologies that can improve the quality of life for individuals experiencing *social isolation*.

1.2.4 Technical Robustness

Given the AI-centric nature of this project, a strong emphasis will be placed on ensuring the technical and social robustness of the models developed. To achieve this, *continuous monitoring* for biases and errors will be implemented, with *feedback loops* for real-time adjustments. The models will undergo rigorous testing across diverse datasets to ensure accuracy, reproducibility, reliability, and explainability. Specific measures include auditing the model's performance on various subsets of data to detect and mitigate biases, and employing explainable AI techniques to make the decision-making process transparent. These steps will minimize unintended outcomes and ensure that the developed technologies are both *technically sound* and *socially responsible*.

1.2.5 Open science

RAVIOLI will adhere to open science practises through the following actions: **(i) Early and transparent sharing:** *Technical reports* and *preprints* will be shared on **Open Access** (OA) platforms such as arXiv and Zenodo, as well as on the personal websites of the ER and the supervisor. This aligns with T6.1, T6.2, T6.4 and T6.5. **(ii) Ensuring reproducibility:** To ensure reproducibility, source code and trained models will be made publicly available and transparently documented through trusted repositories like GitHub, Zenodo, and the EU Open Science Cloud. Repository links will be included in technical reports and publications. This is in line with T6.4 and D6.6. **(iii) Open access to research outputs:** Preprints and published papers will be facilitated through the selection of conferences (e.g., CVPR, ICCV, NeurIPS) and journals (e.g., TPAMI, IJCV, TMLR) that promote OA and encourage the distribution of source code, as outlined in T6.4 and D6.6. **(iv) Public engagement:** The project will promote engagement with public and end-users from the early stages by offering an *interactive demo* on the project's website, providing hands-on experience with the research outputs. This aligns with T6.4 and T6.9.

1.2.6 Research Data Management (RDM)

RAVIOLI will implement and deliver a comprehensive **Data Management Plan** (DMP) (T7.2 and D7.2), ensuring data management aligns with FAIR principles (Findable, Accessible, Interoperable, and Reusable). This includes: (i) using persistent identifiers like DOIs for all research outputs, (ii) providing comprehensive dataset descriptions, (iii) using version control, and (iv) hosting large datasets/models on CTU servers for permanent accessibility. All practices will comply with GDPR to ensure data security and privacy.

1.3 Quality of the supervision, training and of the two-way transfer of knowledge between the researcher and the host

1.3.1 Qualifications and experience of the host and the supervisor

The supervisor, associate **Prof. Giorgos Tolias** leads a team supported by a 5-year Junior Star grant from the Czech Science Foundation. He **supervises active MSCA-PF projects** (one CZ and one EU). He has published over 50 papers in top conferences (e.g., CVPR, NeurIPS) and high-impact journals (e.g., TPAMI, IJCV), with **over 7,000 citations**. He received an "honourable mention – best science paper award" at BMVC 2017, and has co-organized workshops at CVPR, ICCV, and ECCV. His research focuses on *visual representation learning*, *large-scale instance-level recognition*, and *cross-modal recognition*. He currently advises two post-doctoral researchers, three Ph.D. students, and two undergraduates. His work is open-sourced on GitHub, including the very impactful cnnimageretrieval-pytorch package with **over 1,400 stars**. His deep knowledge and hands-on experience with *instance-level* and *open-vocabulary recognition* will directly address the challenges in optimizing the RAP, ensuring RAVIOLI's innovative goals are met. The host organization, CTU, hosts leading CV groups such as VRG and Impact (part of **ELLIS**), led by **Prof. Jiri Matas** and **Prof. Josef Sivic**, respectively. VRG includes 6 faculty members, over 15 Ph.D. students, and 8 post-doctoral researchers. The group has a strong track record in EU research projects (e.g., Darwin, MASH, DIPLECS) and collaborates with industrial giants like Toyota, Samsung, Boeing, Google, and Facebook.

1.3.2 Qualifications and experience of the secondment supervisor, rationale of the secondment

Prof. Rita Cucchiara is director of AImageLab, of the Modena ELLIS Unit, of the Artificial intelligence Research and Innovation (AIRI) and fellow in ML and CV Program. She is Full Professor since 2005 at UNIMORE and responsible for the UNIMORE Unit in the National PhD School in AI for Society. She is in the Board of Directors of IIT, in the Advisory Board of the Max Planck Institute of Intelligent Systems (Germany) and of CVF (Spain). Fellow IAPR, she has been awarded the "Maria Petrou" prize in '18. She has been PI of many EU and National projects. She has published more than **400 papers** (**28872 Cit.**, **h-index 67**). She is **General Chair** of CVPR 2024, ECCV 2022, ICPR 2020, ACM Multimedia (2020), **Program Chair** of ICCV 2017; several times Area Chair of NeurIPS, CVPR, and others. She is Associate Editor of IEEE TPAMI. The rationale for the secondment at AImageLab-UNIMORE is rooted in Prof. Cucchiara's expertise in generative AI, which is crucial for WP2. This expertise will be pivotal in enhancing the integration of synthetic data into the structured memory. This collaboration equips the ER with crucial skills and strengthens the partnership between CTU and UNIMORE.

1.3.3 Two-way transfer of knowledge from host to researcher

Training activities for the ER: A comprehensive training program has been designed to enhance the ER's skills and support the project's success. In collaboration with the supervisors, the ER will design a **Career Development Plan** (CDP) (T5.1) aimed at significantly broadening his skill set. This plan includes both *scientific training* (**WP4**) at CTU and UNIMORE (secondment) and a wide range of *complementary skills development* (**WP5**). Concerning scientific training (**WP4**), the ER will receive training in SOTA CV and DL methodologies (T4.1), covering topics related to RAVIOLI like instance-level recognition, cross-/multi-modal recognition, open-vocabulary recognition and segmentation, retrieval-augmented classification and generation, and generative AI. At CTU, the ER will collaborate with leading experts in these fields, including **Prof. Tolias** (instance-level recognition, cross-/multi-modal recognition, open-vocabulary recognition, retrieval-augmented classification, self-supervised learning), **Prof. Chum** (deep local descriptors, graph-based learning, visual localization), and **Prof. Matas** (CV). During the secondment at UNIMORE, the ER will work with renowned experts like **Prof. Cucchiara** (generative AI, retrieval-augmented generation) and **Prof. Baraldi** (image captioning, explainable AI). The ER will also enhance his writing skills through close collaboration with the supervisor on each scientific publication (T4.2). Additionally, participation in conferences (e.g. CVPR, ICCV, ICLR) will provide the ER with valuable *soft skills* and *networking opportunities* (T4.3). Concerning complementary skills training (WP5), the ER will participate in various courses offered by CTU, such as Intellectual Property Protection, Scientific Work Methodology, Scientific Writing, and Everyday Science and Technology (T5.3). This training will provide the ER with critical skills in *scientific communication*, including academic writing and presentation skills tailored for both technical and non-technical audiences. Additionally, the ER will gain expertise in *IPR management*, *leadership*, *networking*, *knowledge transfer*, and *project management*. To develop leadership and mentorship abilities and integrate into the academic community, the ER will *supervise* undergraduate and postgraduate students (T5.2). CTU allows formal supervision as a co-advisor with post-doc status. The ER will use CTU's system for M.Sc. thesis supervision to post and

promote project-related topics. He will collaborate and engage with Prof. Tolias' team, including three Ph.D. students and two postdoctoral researchers and will also have the opportunity to *lecture* in the supervisor's course.
Host environment and opportunities: CTU is home to **internationally acclaimed researchers** who have made significant contributions to CV and DL. The supervisor regularly publishes in esteemed journals and conferences and maintains strong connections with industry leaders such as Facebook, Naver Labs, and Google, facilitating collaboration and knowledge exchange. Consequently, the ER will have the opportunity to *engage* with accomplished researchers in both academia and industry. VRG consists of individuals from various nationalities and diverse backgrounds, creating an enriching international environment conducive to the *seamless integration* of ER. **Regular weekly discussions** within reading group will enable the ER to present ongoing work, receive feedback, refine his research ideas and tackle ongoing challenges. Lastly, the alumni network of CTU and VRG includes numerous distinguished Ph.D. holders, many of whom work in leading research labs at Facebook and Google or have established their own research groups across Europe. The ER will have opportunities to connect with these alumni through organized seminars, workshops, and alumni events, creating opportunities for collaboration, directly benefiting RAVIOLI.

1.3.4 Two-way transfer of knowledge from researcher to host

The host organization will benefit on multiple levels from the ER. **(i) Research expertise:** The ER's extensive background in CV and DL, honed through extensive Ph.D. research and participation in high-impact research projects, along with numerous top-tier publications, will enhance VRG's research capabilities. He will actively contribute to the university and the team by **conducting workshops and talks at least once per quarter**, specifically tailored for students or team members, including Ph.D. students, postdoctoral researchers, and faculty. These sessions will focus on advancing knowledge in areas such as recognition, detection, and localization. **(ii) Software engineering skills:** The ER's proficiency in DL frameworks and coding will be shared with VRG members and students, increasing their productivity and helping them achieve research goals. His skills in creating webpages and interactive demos will also strengthen VRG's outreach. **(iii) Network and collaboration:** The ER will leverage his extensive network, built during his Ph.D. and through his involvement with ELLIS, to foster long-term relationships between VRG at CTU and key institutions and companies like the National Technical University of Athens, ETH Zurich, University of Alicante, University of Crete, EURECOM, and Inria Rennes. This network will **enhance collaborative opportunities and increase VRG's visibility** within the research community. The ER will lead efforts in collaborative research, facilitate academic exchanges, organize joint presentations, and pursue joint research proposals, thereby enriching VRG's research environment and broadening its impact.

1.4 Quality and appropriateness of the researcher's professional experience, competences and skills

ER is an early-career researcher with a **solid background** in CV and DL, focusing on class- and instance-level recognition, retrieval, detection and localization. He completed his Ph.D. in 2024 from the National Technical University of Athens (NTUA) and holds a M.Sc. in Data Science and Information Technologies from the National Kapodistrian Univ. of Athens (NKUA). Throughout his Ph.D., ER developed innovative methods directly relevant to RAVIOLI, including SimPool[35] for weakly-supervised localization, AttMask[36] for transformer pre-training on dense tasks, Metrix[34] for recognizing unseen categories in dynamic environments, and WeiCom[37] for enhancing VLMs' ability to handle new domains. These contributions, published in top conferences (ICCV, ECCV, ICLR, NeurIPS) and high-impact journals, have garnered **over 190 citations** on Google Scholar in under 4 years. ER's **commitment to open-source research** is evident through his active sharing of code, models, and datasets on platforms like GitHub, which have garnered **over 250 stars**. ER has substantial research experience, having held positions at Inria Rennes, IARAI, Athena RC, and NTUA. He has **contributed to European and international projects**, including H2020 "iTOBOS" (melanoma detection) and KAUST-funded "eOSD" (oil spill detection) – both relevant to RAVIOLI. Beyond research, ER has co-organized the 2nd ELLIS Doctoral Symposium, served as session co-chair and peer reviewer for top conferences and journals, and contributed to teaching at NTUA. This blend of experience shows his **readiness to lead and execute the proposed project**.

2. Impact #@IMP-ACT-IA@#

2.1 Credibility of the measures to enhance the career perspectives and employability of the researcher and contribution to his/her skills development

ER aims to make impactful contributions to CV and DL through innovative ideas that will lead to *major advancements* towards general AI. Throughout the project, ER will explore novel concepts, expand his academic network, and deepen his expertise, ultimately establishing himself as **scientifically mature**, **independent** and **internationally recognized**. Collaborating with Prof. Tolias and VRG, ER will deepen his knowledge in instance-level recognition, cross-/multi-modal recognition, and retrieval-augmented classification. His secondment with Prof. Cucchiara and AImageLab will enhance his skills in generative AI, retrieval-augmented generation, image

captioning, and explainable AI, providing *invaluable insights*. RAVIOLI is expected to yield high-quality, impactful results, disseminated in top-tier journals and conferences, advancing the SOTA and **solidifying ER's publication record**, contributing significantly to his academic development. The teaching and training activities, such as student supervision and lecturing (T5.2), will also **enhance ER's teaching portfolio**, providing valuable experience in academic leadership and mentorship.

Looking ahead, ER **envisions establishing his own research lab**, focusing on the intersection of CV and DL. He plans to apply for major EU grants, such as ERC Starting Grants, to support his research initiatives and further his academic career. To summarize, this project will enable ER to: (i) **enhance science communication skills**, improving his ability to effectively convey scientific concepts to diverse audiences; (ii) **achieve international recognition** through high-impact publications and conference presentations; (iii) **strengthen his academic record**, advancing his career and employability in roles like Lecturer or Assistant Professor; (iv) **expand his international network** through collaboration with leading experts; and (v) **acquire essential transferable and entrepreneurial skills** for managing research projects and engaging in impactful initiatives. These measures will enhance ER's academic career prospects.

2.2 Suitability and quality of the measures to maximise expected outcomes and impacts, as set out in the dissemination and exploitation plan, including communication activities #@COM-DIS-VIS-CDV@#

Instrument	Activities	Target Group	Expected Outcomes
Scientific publications	Research papers in top-tier conferences (CVPR, ICCV, ICLR, NeurIPS) and extensions in high-impact journals (TPAMI, IJCV, TMLR).	Scientific community, other R&I stakeholders	Enhanced visibility, increased citations, and collaboration invitations.
Workshops and Tutorials	Workshops and tutorials on the topics of RAVIOLI, providing a platform to bring together community, raise awareness, discuss project outcomes, share insights, engage with peers, and acquire feedback.	Scientific community, industry professionals.	Networking opportunities, feedback from peers, and potential for new research collaborations.
Technical blog posts	Accessible descriptions and targeted extracts of project results disseminated via the project's website and social media (LinkedIn, X).	Public, technology and scientific groups	Increased public engagement, attracting diverse audiences.
Multimedia material	Slideshows, videos, demos, and webcasts presenting project information, communicated via the project website and social media (YouTube, Hugging Face, Facebook, Medium, LinkedIn, and X).	Public, research groups, technology providers	Broadened communication, enhanced visibility, and public engagement.
Science Outreach Events	Communication of the project to science outreach events like Czech European Researchers' Night, Maker Faire, schools and universities.	Public, innovators, students	Increased awareness, inspiring future generations.

Dissemination Plan: The dissemination activities aim to share the project's outcomes with the scientific community and other stakeholders (T6.4). These activities, detailed in the table above, include *scientific publications* (D6.6), *workshops* and *tutorials* (T6.6), and *technical blog posts* (T6.2, T6.3). The scientific findings and advances will be disseminated through **three high-quality publications and presentations at top-tier conferences** like CVPR, ICCV, and ICLR (D6.6). These conferences, ranked among the top 15 across all scientific fields (Google Scholar), are known for their high impact and broad reach, ensuring the findings reach researchers in both academia and industry. Extensions of the works will be published in high-impact, open-access journals like TPAMI, IJCV, and TMLR (D6.6). Leveraging the experience of Prof. Tolias and Prof. Cucchiara in organizing workshops and tutorials, **a workshop will be organized** during project's second year, collocated with CVPR (T6.6). This workshop will include *invited talks* and *calls for papers* on the project's research topic, with a target of attracting 30-50 participants. Additionally, technical blog posts will present accessible descriptions and key results on the project's website and social media to engage a broader audience and attract diverse stakeholders, with a goal of reaching at least 1000 unique visitors per blog post. This involves creating a *project website* and *social media accounts* like X (T6.2), which will be maintained and updated (T6.3) throughout the project's duration. Given the ER's extensive social media network and demonstrated reach (e.g., previous LinkedIn (X) post garnering 18k (23k) impressions), the project is expected to achieve significant online engagement.

Communication Plan: The communication activities aim to inform and promote the project among the public, researchers, industry professionals, and stakeholders, ensuring *broad awareness* and *engagement* (T6.4), with

support from CTU's Office of Public Relations. The communication activities, detailed in the table above, include development of *multimedia material* and communication via the project website and social media channels (T6.3), as also participation in *fairs*, *exhibitions*, and *science outreach events* (T6.9). Slideshows, videos, and webcasts will be shared through the project's website and social media channels, with an expected outreach to 5,000 social media connections across platforms. An **interactive open-vocabulary segmentation demo** (T6.7) will be hosted both on Hugging Face and the project's website, enhancing public outreach and communication through hands-on experience, targeting 200 demo interactions in the first year. As part of CTU's school outreach programs, regular school visits provide an opportunity to present this demo and project information to high school students (T6.10), with an aim to engage 100 students annually. A similar presentation will be given at the Czech European Researchers' Night (T6.9), where the target is to reach at least 150 participants. Both ER and the supervisor have experience in such activities. ER will lead these efforts with support from student researchers, allowing them to deepen their scientific understanding and develop valuable soft and hard skills.

Exploitation Plan: To ensure practical application and utilization of research outcomes, the project will adopt a comprehensive exploitation strategy (T6.5). This involves making frameworks, models, and results available as OA, ensuring they are *well-documented* for easy adoption. **Key Exploitation Results** (KER) will be identified as critical outputs with potential for development and commercialization. A robust **Intellectual Property Rights** (IPR) strategy will be established under the guidance of Prof. Tolias and CTU experts (T7.1), with initial development within the first month (D7.1) and reviews every three-months. Leveraging Prof. Tolias' experience with patent applications, this strategy will protect significant innovations through patents, design rights, and copyrights ER will collaborate with IPR management experts at CTU to implement the IPR strategy, ensuring protection of all significant outputs. Intellectual property agreements with CTU will define ownership, rights, and responsibilities, including open licensing options like Creative Commons, MIT, GPL, or Apache 2.0. This exploitation plan (T6.5, D6.5) is integrated with the dissemination and communication strategies (T6.1, D6.1), ensuring that all activities are *aligned* and *mutually reinforcing*.

2.3. The magnitude and importance of the project's contribution to the expected scientific, societal and economic impacts

Scientific Impact: RAVIOLI advances CV and DL by addressing a key limitation of current VLMs, which excel at image-level but struggle with pixel-level tasks. This project introduces a novel approach that integrates a memory with a VLM, marking a first in the field. This memory, operating alongside a pre-trained and frozen VLM, enhances the VLM's ability to adapt to new concepts, domains, and classes without retraining. This framework bridges the gap between VLMs and segmentation tasks, enabling their application in specialized domains like satellite and medical imagery, where models like CLIP currently underperform. The project's impact extends beyond CV and DL researchers; it offers new tools and methodologies applicable across disciplines, enabling scientists in fields like geoscience to perform precise segmentation for environmental monitoring or disaster response without advanced coding skills. The commitment to open-source and open-access principles ensures that all models and frameworks developed are accessible, fostering transparency, reproducibility, and collaboration. By **introducing a new paradigm in retrieval-augmented VLMs**, the project will make lasting contributions to scientific progress, innovation, and interdisciplinary collaboration.

Economic/technological Impact: RAVIOLI is set to deliver significant economic and technological advancements by enhancing VLMs with a memory, making them more *efficient*, *adaptable* and *interpretable* for various industries. This innovation will enable retrieval-augmented VLMs to perform dense tasks without retraining, potentially reducing computational demands by up to 50%, leading to *energy savings* and significant *cost reductions*[38]. For example, in the automotive industry, autonomous vehicles equipped with memory-augmented models could quickly adapt to new environments, improving safety and reliability and reducing accident rates, as suggested by similar advancements[26], resulting in substantial savings for insurance companies and manufacturers. In retail, these adaptable models could enhance inventory management and customer service, boosting sales[5] and reducing operational costs. Additionally, RAVIOLI's enhanced VLMs can *optimize processes* in fields like geospatial analysis, where more accurate segmentation of satellite images can improve environmental monitoring and disaster response. These improvements will lead to better decision-making and more effective resource management, benefiting both public and private sectors. The open-source nature of RAVIOLI's frameworks will also drive *innovation* and *commercialization*, enabling start-ups and established companies to develop new AI-driven products and services, contributing to economic growth and technological progress across multiple sectors.

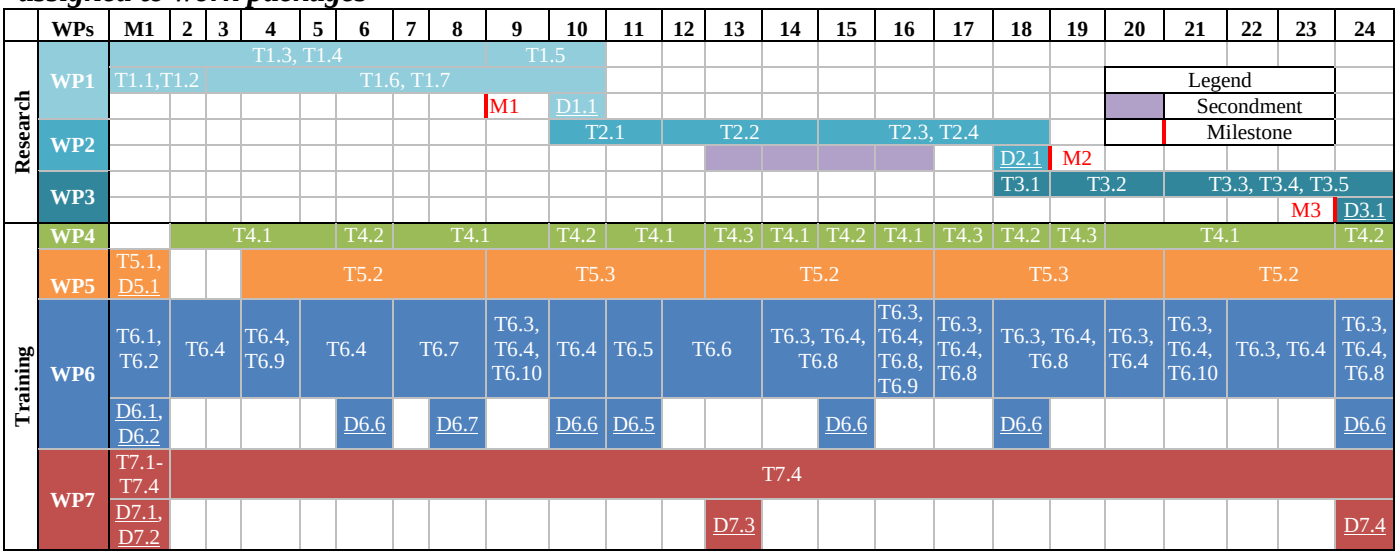
Societal Impact: RAVIOLI will make substantial societal contributions by promoting Green AI, enhancing safety, accessibility, and environmental stewardship. By reducing computational demands[38] with retrieval-augmented

VLMs, the project promotes *energy efficiency* and *cost savings*, **aligning with global sustainability goals**. The interpretability of memory-based predictions also offers *transparency*, building trust in AI applications. For public safety, our framework improves autonomous vehicle performance in complex, densely populated areas, potentially *reducing traffic accidents*[26] and *saving lives*. In assistive technology, RAVIOLI enables smart devices like canes and glasses to adapt to the specific needs of visually impaired individuals, *promoting independence* and *reducing healthcare burdens*. The framework's adaptability extends to environmental monitoring, providing precise segmentation for disaster response and conservation efforts, crucial for *managing natural resources* and *responding* to *crises*. Additionally, in fields like precision agriculture, cultural heritage preservation, and security, RAVIOLI's ability to tailor AI applications to specific contexts enhances *efficiency* and *safety*, offering societal benefits.

#§COM-DIS-VIS-CDV§# #§IMP-ACT-IA§#

3. Quality and Efficiency of the Implementation #@QUA-LIT-QL@# #@WRK-PLA-WP@# #@CON-SOR-CS@# #@PRJ-MGT-PM@#

3.1 Quality and effectiveness of the work plan, assessment of risks and appropriateness of the effort assigned to work packages



Work Plan: RAVIOLI is organized in **7 Work Packages** (WPs) according to the Gantt chart above and the sequence below (NOTE: more details in Sections 1.2, 1.3 and 2.2; M1: 1st month). **WP1: Development of the Retrieval-Augmented Predictor** [M1-M10]: T1.1: Create dataset; T1.2: Create a structured memory; T1.3: Design the fusion model architecture; T1.4: Design the voting mechanism; T1.5: Explore improved positional encoding techniques; T1.6: Train the fusion model; T1.7: Evaluate in a class-generic way; D1.1: OA source code and trained models; M1: Retrieval-Augmented Predictor (RAP). **WP2: Generative AI for memory enrichment** [M10 – M18]: T2.1: Synthesize images for specific concepts; T2.2: Develop and explore grounding techniques; T2.3: Integrate hybrid memory into the RAP; T2.4: Evaluate the RAP with hybrid memory; D2.1: OA source code and trained models; M2: RAP with enriched hybrid memory. **WP3: Granularity flexibility in recognition and segmentation** [M18-M24]: **T3.1**: Develop a hierarchical label structure in memory; **T3.2**: Create a dedicated part-level memory for specific object parts; **T3.3**: Extend the RAP to support hierarchical labels and part segmentation; **T3.4**: Integrate real and synthetic examples into the hierarchical and part-level memories; **T3.5**: Evaluate the RAP in hierarchical label assignment and part segmentation; **D3.1**: OA source code and trained models; M3: RAP with extended granularity support. **WP4: Scientific training** [M1-M24]: T4.1: Acquire expertise in SOTA CV and DL; T4.2: Acquire skills in writing scientific publications; T4.3: Participate in conferences. **WP5: Complementary training** [M1-M24]: T5.1: Design Career Development Plan (CDP); T5.2: Develop skills in lecturing and supervising; T5.3: Develop skills in IPR management; scientific writing, etc.; D5.1: CDP. **WP6: Dissemination, Communication, and Exploitation** [M1-M24]: T6.1: Design optimal Dissemination and Communication Plan; T6.2: Develop project website, social media, etc.; T6.3: Maintain and update project website, social media, etc. T6.4: Communicate (disseminate) project (results) to public, researchers, stakeholders, etc.; T6.5: Design Exploitation Plan; T6.6: Organize a workshop; T6.7: Develop interactive demo; T6.8: Maintain and update interactive demo; T6.9: Participate in science outreach events (e.g., Czech European Researchers' Night); T6.10: Participate in CVUT-VRG school outreach programs (e.g., Open Day); D6.1: Dissemination and Communication Plan; D6.2: Project website, social media pages; D6.5: Exploitation Plan; D6.6: Scientific publications in

conferences and journals; D6.7: Interactive demo. **WP7: Project Management** [M1-M24]: T7.1: Design Intellectual Property Rights (IPR) strategy; T7.2: Design Data Management Plan (DMP); T7.3: Design Impact Action Plan (IMP); T7.4: Plan project activities and manage tasks (weekly meetings, use of funds, risk assessment, etc.); D7.1: IPR strategy; D7.2: DMP; D7.3: Interim report to EC; D7.4: Final report to EC.

Appropriateness of the effort: The work plan allocates ten months for WP1, focusing on the foundational development of the RAP. This extended period allows for careful optimization of inputs, architecture, and training, ensuring a robust model generalizing well across unseen classes. Eight months are allocated for WP2, focusing on the complex integration of generative AI for memory enrichment, including a secondment at UNIMORE with Prof. Cucchiara to leverage expertise in diffusion-augmented VLMs. WP3 spans six months, emphasizing the development of granularity flexibility for detailed part-level understanding and hierarchical label assignment. Overlapping work periods allow flexibility for adjustments without impacting the timeline. WP4, WP5, WP6, and WP7 run concurrently throughout the project, ensuring that training, dissemination, communication, and management activities are fully integrated, maximizing impact and ensuring high-quality outcomes.

Risk Management and Management Structure: RAVIOLI will be primarily managed by the ER, with guidance and supervision provided by the host supervisor. To ensure the *smooth progress* of the project, biweekly meetings between the ER and the supervisor will be scheduled to discuss ongoing tasks, track progress, and promptly address any emerging risks. Quarterly meetings with the supervisor and UNIMORE collaborators, including Prof. Cucchiara, will focus on WP2's generative AI components, exploitation strategies, and timely delivery of results. CTU's administration will support financial management, with a dedicated EU Financial Officer assisting in reporting, financial planning, and resolving any potential issues related to project implementation. Potential risks will be reviewed at each meeting, ensuring proactive identification and mitigation. These risk mitigation strategies have been specifically designed to align with the project's phased timeline, ensuring that potential issues are addressed promptly without disrupting the overall progress. Detailed measures are outlined in the following table.

Risk	Description	WPs	Chance	Impact	Mitigation Measures
Model Performance	RAP may not achieve SotA performance or generalize well.	WP1, WP2, WP3	Low	High	Buffer periods for model adjustments; diverse memory with synthetic examples; continuous fine-tuning.
Computational Resources	Limited resources may delay training and evaluation.	WP1, WP2, WP3	Low	Medium	Schedule cluster use; backup setup available; reallocate tasks if needed.
Delay in Synthetic Data Integration	Delays may impact WP2 timeline.	WP2	Low	Medium	Align secondment with WP2, providing necessary expertise; parallel tasks to mitigate delays.
Publication Rejection	Papers may be rejected by top-tier venues.	WP6	Medium	Low	Revise/resubmit quickly; use preprints on arXiv.
Coordination Challenges	Communication issues during secondment may hinder coordination.	WP2, WP6	Low	Medium	Regular virtual meetings; clear responsibility plan; contingency plans in place.

3.2 Quality and capacity of the host institutions and participating organisations, including hosting arrangements

CTU-VRG is globally renowned for its work in CV and DL, offering the ER the opportunity to collaborate with leading scientists and deepen his expertise. His prior engagement with CTU-VRG ensures seamless integration, allowing him to contribute from day one through regular meetings, collaborative projects, and mentorship activities. The ER will also benefit from the group's frequent interactions with distinguished visiting researchers, such as those participating in "The Colloquium in Pattern Recognition and Computer Vision." During his secondment at AImageLab-UNIMORE, he will work closely with Prof. Cucchiara, benefiting from a supportive environment that maximizes the secondment's impact. Both CTU-VRG and AImageLab-UNIMORE have a strong track record in MSCA and European projects, providing robust support, high-performance computing, and advanced deep learning hardware. CTU-VRG offers multiple GPU servers, access to a supercomputing cluster, ample storage, IT infrastructure for website development, and a vast library of online journals. Similarly, AImageLab-UNIMORE provides substantial computational resources, including multiple servers and access to university and national clusters. Both institutions offer the necessary infrastructure for communication, outreach, and data storage, ensuring the project's success. The host organization, CTU, will adhere to MSCA Guidelines on Supervision, aligned with the European Charter and Code for Researchers. More details in Section 5.2 of Part B-2.

#§CON-SOR-CS§# #§PRJ-MGT-PM§# #§QUA-LIT-QL§# #§WRK-PLA-WP§#