WACV DAILY

Sunday

Winter Conference on Applications of Computer Vision

In cooperation with

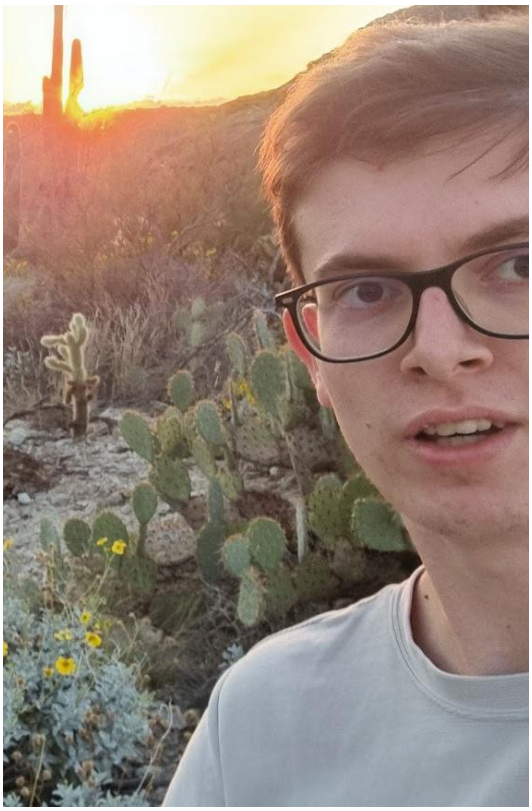
Computer Vision News

The Magazine of The Algorithm community

IEEE COMPUTER SOCIETY

For today, Sunday 8

Giacomo Pacini is a PhD student at the University of Pisa and an associate researcher in Multimodal AI at CNR-ISTI, Italy.



"My research centers on bridging the gap between image and text modalities. My recent work focuses on image captioning, multimodal information retrieval, and finding new applications for unsupervised vision backbones. Currently, I am studying how to better exploit the semantic representations of vision encoders in Large Multimodal Models. When I am not doing research, I love developing apps, tinkering with home automation, and learning new things."

"My work at WACV this year introduces a training-free pipeline for class-agnostic counting. We leverage fully unsupervised vision backbones to extract features and perform zero-shot object counting. It's a simple but highly effective approach that eliminates the huge costs of data collection and parameter tuning for unseen object classes!"

"I hope everyone attending WACV here in sunny Tucson has a fantastic time! Don't forget to check out the amazing sessions today, and feel free to stop by my poster this morning (#77) if you need help counting saguaros!"

Giacomo's picks of the day:

Orals

13:45-14:45 MageBench: Bridging Large Multimodal Models to Agents

13:45-14:45 You May Speak Freely: Improving the Fine-Grained Visual Recognition Capabilities of Multimodal LLMs with Answer Extraction **see full review on page 4**

15:00-16:00 UniCoRN: Latent Diffusion-based Unified Controllable Image ...

Posters

Poster Session 1-37: M4U: Evaluating Multilingual Understanding and Reasoning ...

Poster Session 1-64: Prompt-OT: An Optimal Transport Regularization Paradigm ...

Poster Session 1-69: PaRaChute: Pathology-Radiology Cross-Modal Fusion for ...

Russian Invasion of Ukraine

Our sister conference CVPR condemns in the strongest possible terms the actions of the Russian Federation government in invading the sovereign state of Ukraine and engaging in war against the Ukrainian people. We express our solidarity and support for the people of Ukraine and for all those who have been adversely affected by this war.

WACV Daily

Editor: **Ralph Anzarouth**
Publisher & Copyright: **Computer Vision News**

Our editorial choices are fully independent from IEEE, WACV and the conference organizers.

You May Speak Freely: Improving the Fine-Grained Visual Recognition Capabilities of Multimodal Large Language Models with Answer Extraction

Logan Lawrence is a second-year PhD student at UMass Amherst.

His paper examines how multimodal large language models perform on fine-grained visual recognition tasks such as species identification. Logan speaks to us ahead of his oral and poster presentations this afternoon.

Evaluating large language models is harder than it first appears, because their answers are generated as free-form text rather than selected from a fixed set of labels. **The challenge becomes even greater in highly specialized visual domains where models must distinguish between extremely similar categories.** "*Whether it's birds or flowers or insects, evaluating these niche domains is really, really hard,*" Logan confirms. "*The choice count is in the hundreds of thousands.*"

In this work, Logan explores what happens when multimodal large language models are applied to these fine-grained classification problems. One of the motivations for this came from his advisor, **Grant Van Horn**, who is widely known for the **iNaturalist challenges** and for his work connecting machine learning research with real-world ecological applications. "*He works with ecologists, scientists, and the Cornell Lab of Ornithology to deploy AI and ML to natural world recognition systems,*" Logan reveals. "***Ecologists and scientists have a real need for integrating AI into their products.***"

He traces the origins of the project back to earlier work on vision-language models such as CLIP. As increasingly powerful multimodal systems began attracting attention and investment, he wanted to understand how useful they are in these specialized settings. "*Everyone's putting tons of money into these models,*" Logan says. "*But how well do they perform on these tasks?*"

15x ~

"What is the species of this bird?"

free-form

This bird is an Ivory Gull.

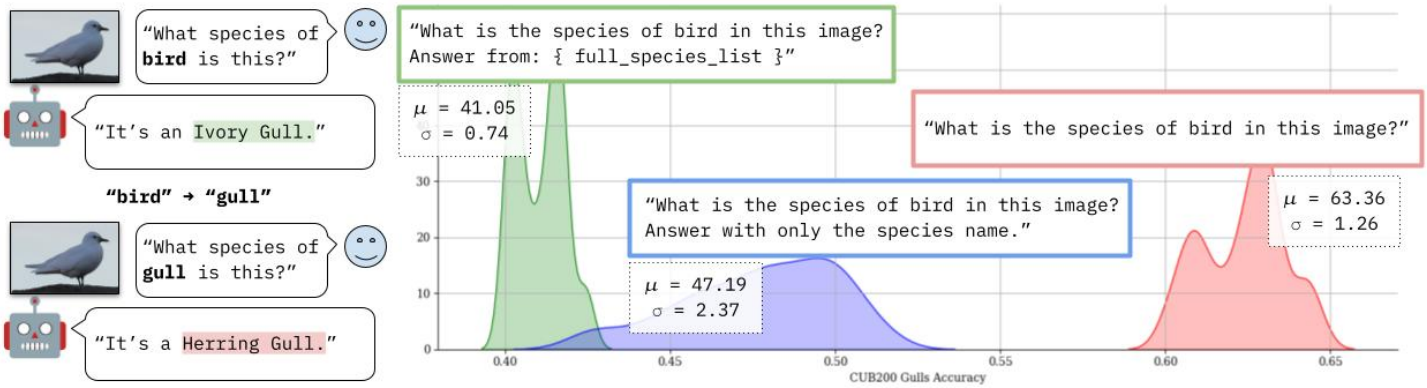
choice

Crested Auklet

"What is the most likely species?"

nlg2choice

Ivory Gull



Through a series of experiments, the team began to narrow down several contributing factors. One issue comes from the scale of the classification space. When a model must choose from hundreds of species, the prompt context can grow very large, making it harder for it to keep track of which option it is selecting. Logan describes this as **a situation in which the model struggles to maintain the connection among the question, the choices, and the final answer.**

Perhaps the most surprising finding was a form of brittleness in the models' responses. "*If you're an ornithologist and you're using an LLM to ask, 'what is the species of this bird?',*" he says. "*Very small changes in that input – for instance, you say, 'identify the species of bird in this image' – can result in very different predictions.*"

He gives an example on the gull subset of the CUB200 dataset, where slight variations in phrasing can shift the model's prediction from one species to another, and affect overall benchmark performance (see image).

Another challenge was computational. Evaluating models across very large sets of candidate labels requires an enormous number of forward passes through the model. For researchers without the resources of large technology companies, this quickly becomes impractical. However, Logan and his collaborators ran their experiments on their own research cluster rather than relying on commercial APIs.

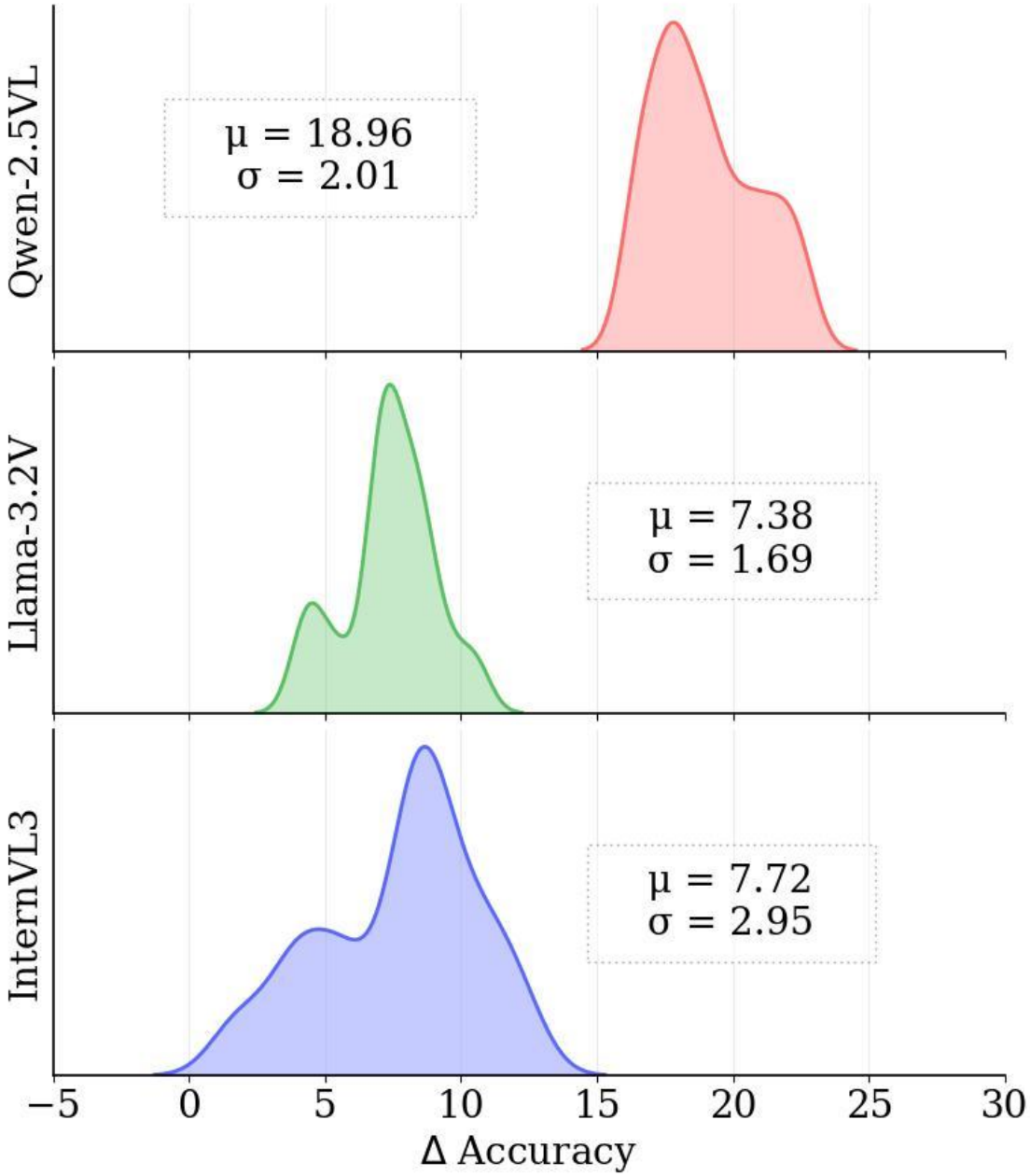
Also, instead of calculating full probability sequences for every possible class name, the team truncated the process once a class could be uniquely distinguished by its early tokens. **By stopping the calculation once a species name becomes uniquely identifiable, the number of required computations can be dramatically reduced.**

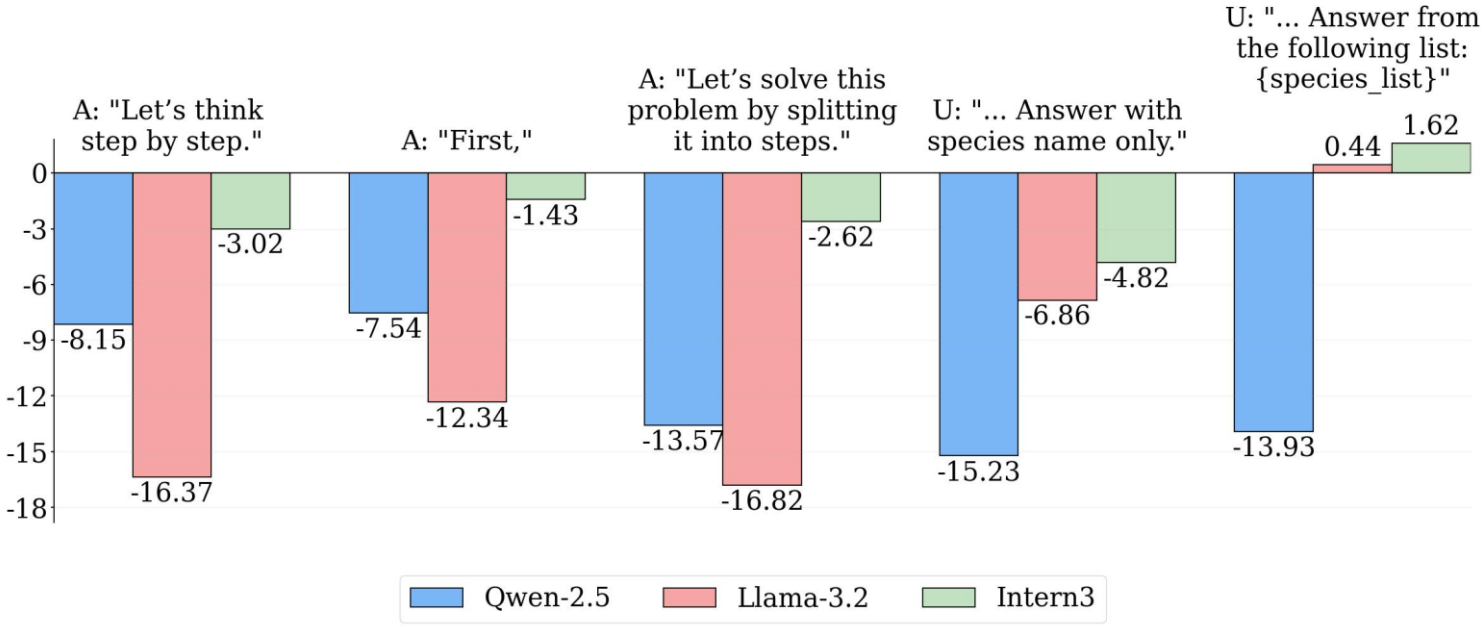
"*CUB200 has 200 bird species, so if you wanted to get the probability that an LLM would say a specific species out of those 200, it would require around 687 forward passes, which is not tenable for us,*" he explains. "***But if you just need to tell which is the most likely, it's much,***

much quicker." For CUB, that meant bringing 687 forward passes down to 47.

Logan believes the importance of the work lies in addressing questions that researchers in applied fields are already asking. "***People are focusing on LLMs, but they're not really focused on this real niche use case for a subset of scientists and ecologists,***" he points out.

He hopes the work will encourage further research into improving these models rather than simply measuring their limitations. One direction he finds particularly promising is fine-tuning multimodal models with stronger domain knowledge. *"We've drilled down closer to how you're supposed to evaluate these models in terms of the specific task,"* he says, *"Now, I'd really like to see some work in terms of what we can leverage to increase the built-in knowledge of these models!"*





For Logan, the project is also part of his broader journey as a researcher. Asked what he hopes to do after completing his PhD, he says he is keeping an open mind. "*To be honest, right now I'm keeping myself open to whoever's doing great work and whoever maintains academic integrity and freedom,*" he tells us. "*Right now, for me, that aligns mostly with academia!*"

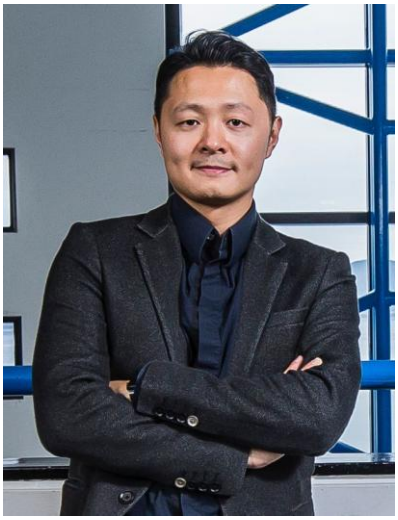
Finally, we asked Logan how he would summarize this work for conference attendees. He puts it plainly: "***LLMs are bad at fine-grained recognition – and here's why.***"

If that has piqued your interest, you can learn more about Logan's work during Oral Session 5B: Vision+Language and Other Modalities I, Sunday 13:45–14:45 in AZ Ballroom 6, and during Poster Session 4, Monday 16:30–18:15 in the Tucson Ballroom and Prefunction Space.

Don't miss today's industry panel, with Gérard Medioni and more great speakers! Moderated by awesome Emanuela Marasco! At 12:30, AZ Ballroom 6


CVPR DAILY
CVPR 2022
Computer Vision and Pattern Recognition
Thursday
Tuesday
Wednesday
Friday
Computer Vision News
The Magazine of The Algorithm community
RSIP VISION
Don't miss the BEST OF WACV 2026
in Computer Vision News of April.
Subscribe for free and get it in your mailbox!
Click here

ORCA: Object Recognition and Comprehension for Archiving Marine Species



Yuk Kwan (David) Wong - left - is a master's student, and Ziqiang Zheng - center - is a postdoctoral researcher at the Hong Kong University of Science and Technology, working with Kit Yeung - right.

Their paper explores how computer vision systems might better support marine science by addressing gaps in existing datasets and task design for ocean research. They speak to us ahead of their oral and poster presentation this afternoon.

Artificial intelligence has achieved strong results across many domains, but its performance is less consistent in specialized scientific settings such as marine research. In response, the team approached the problem from a data perspective rather than focusing primarily on model design. **They identified two major limitations in the datasets commonly used for marine applications**.

The first issue concerns geographic diversity. Many marine datasets are collected in specific locations and reflect only a narrow slice of ocean environments. *"Usually when biologists collect data, they have a dataset of the Red Sea or Hong Kong, for example,"* David tells us. *"They're limited in terms of diversity."* As a result, **models trained on those datasets often struggle to generalize beyond the environments in which the data were collected**.

To address this, the **ORCA dataset** was designed to cover a much broader range of marine species and environments. The dataset contains more than **14,000 images spanning hundreds of species, with tens of thousands of bounding-box annotations and expert-verified captions describing individual organisms**. By expanding both species

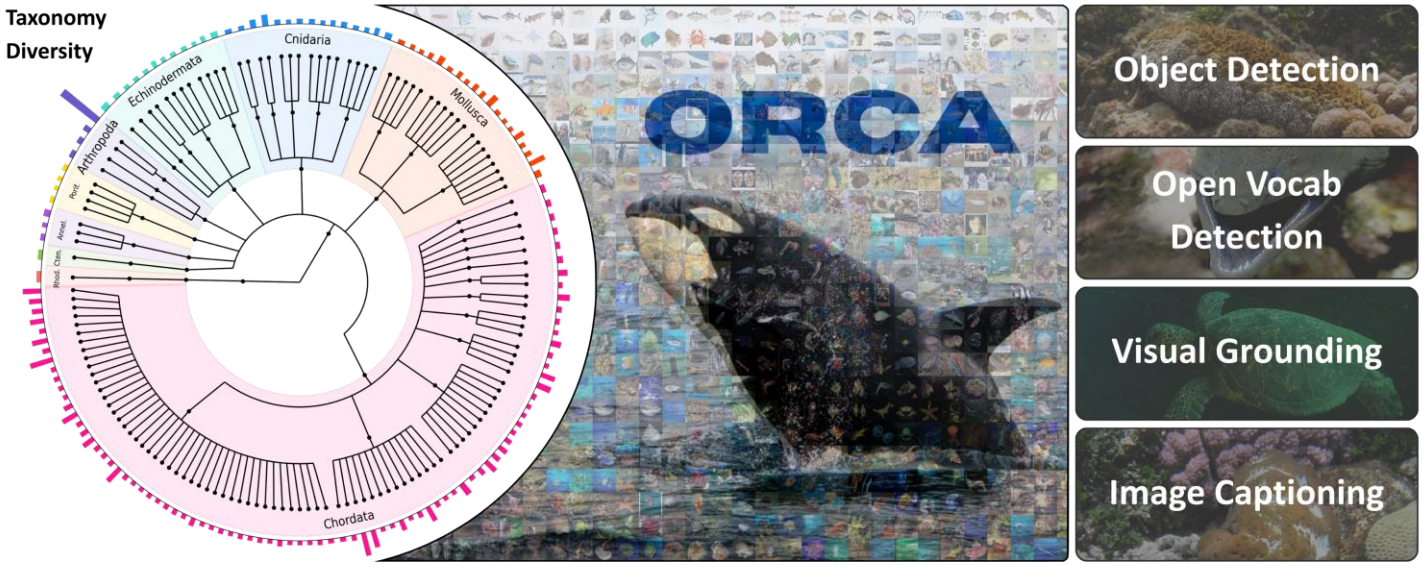
coverage and annotation detail, the dataset aims to reflect the diversity encountered in real-world marine surveys more accurately.

The second challenge relates to how computer vision tasks are typically defined. Many benchmarks rely on simple tasks such as image classification or short captions. For marine scientists, however, those outputs are often too limited to be useful in practice. **Marine ecosystems are complex, and researchers often need richer descriptions of the organisms they observe**. "*Using a single-label task like image classification – this is an image of a fish, this is an image of a dolphin – may not be very helpful,*" David points out. "*They need something more professional.*"

To better match those needs, the team designed a set of tasks tailored to marine research. The dataset supports **object detection, instance-level captioning, and visual grounding,** allowing models to both identify organisms and generate detailed descriptions of them. Crucially, many of the captions describe morphological features, such as body shape, color patterns, or fin structures, that marine biologists use to distinguish between similar species.

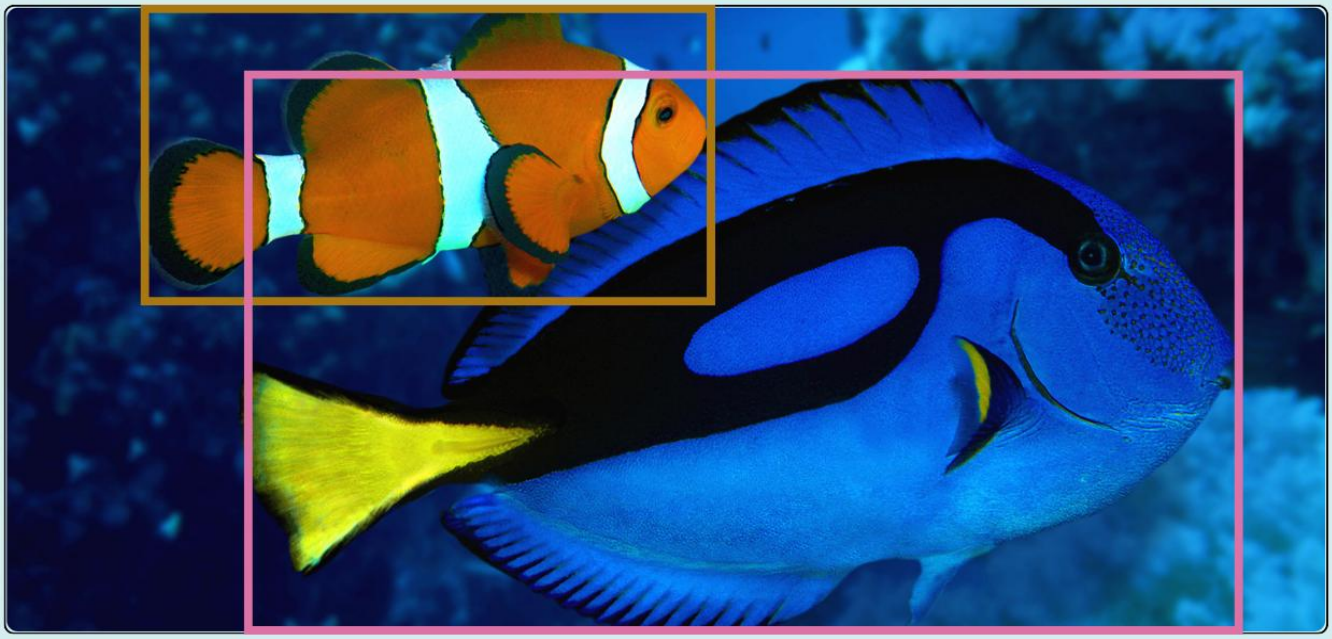
Building those annotations required expertise beyond computer science. The project involved close collaboration between computer vision researchers and marine biologists. "*In the author list, you'll notice there are some people from computer science, and there are biologists as well,*" David says. "*It's a multi-disciplinary paper.*" Marine experts helped annotate images and verify descriptions to ensure that the dataset reflects real biological knowledge.

For Ziqiang, another important aspect of the work is enabling models to recognize species that may not already appear in training data. "*Nowadays, for underwater*



Instance-Level Annotation

Common Name: Clownfish
Caption: The clownfish displays a bold orange body adorned with three vertical white bands outlined in black. Its rounded fins and small, streamlined shape highlight its vibrant, contrasting coloration.



Common Name: Blue Tang Surgeonfish
Caption: The blue tang boasts a sleek, vibrant blue body accented by a bold black pattern resembling a painter's palette. Its bright yellow tail and smooth, oval shape give it a striking, elegant appearance.

exploration, we always need to detect new species," he explains. "*Therefore, in this work, we performed the first attempt to do some open-vocabulary object detection.*" This allows models to identify objects beyond a fixed set of categories. **In domains like marine science, where biodiversity is vast, and many species remain understudied, that flexibility is especially important**.

Developing the dataset also highlighted how interdisciplinary research can introduce new challenges of its own. Kit points out that building effective collaboration with marine scientists requires significant effort and communication. While many researchers in marine science are interested in using AI, integrating it into their workflows requires careful discussion about how the tools can

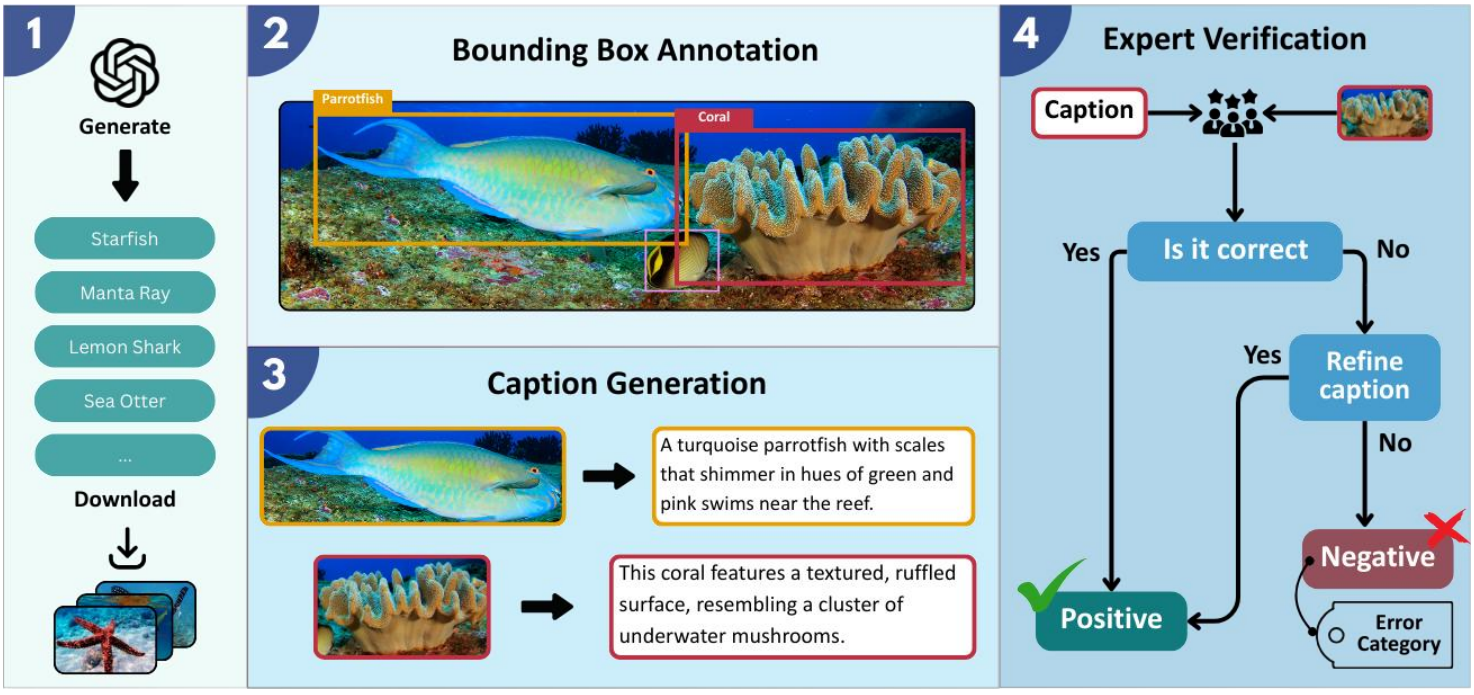
support their research. *"The difficulty is really getting domain expertise to work with us,"* he says. *"We need to spend time communicating with marine biologists so we can understand how these tools could improve their work."*

For Kit, this collaboration reflects a broader goal for the field. *"As computer vision scientists working on AI, we should be more like the first step movers,"* he asserts. *"We should have open arms and just let them collaborate with us because, at the end of the day, we want AI to contribute to scientific study. AI for science is a bigger goal. We need to work closely together for the greater good."*

On the technical side, the team found that current vision-language models still struggle with detailed biological descriptions. Existing captioning systems often produce very short summaries of images. *"Usually, they just say that this is a red fish, this is a green fish,"* David says. *"Marine biologists need a more professional description."* Instead, scientists often rely on fine-grained details such as the shape of the dorsal fin, body proportions, or distinctive scale patterns.

Generating those descriptions requires models to pay attention to specific regions of an organism. In their experiments, the researchers found that many systems focus primarily on the overall scene rather than the fine-grained details needed for species identification. This gap highlights the need for models that integrate global context with localized visual analysis.

For David, the collaboration itself stands out as one of the most rewarding aspects of the project. Working closely with marine biologists



meant learning how to communicate across disciplines and translate scientific requirements into computational tasks. **The result is a dataset that integrates biological expertise directly into computer vision research**. *"We're able to get something that others can't,"* he tells us. *"Since we're collaborating with marine biologists, this is the first dataset that has some domain-specific knowledge. We're bringing new knowledge to the community so others can use it in their work. That's something I'm proud of."*

On that subject, the team hopes the dataset will enable new research directions in marine monitoring. David suggests that **open-vocabulary detection and instance-level captioning remain particularly challenging problems for the field**. Ziqiang highlights another direction: **taxonomy**. Identifying marine species often requires understanding hierarchical relationships between related organisms. *"For species monitoring, there is some hierarchical taxonomy within the marine domain,"* he explains. *"That's very different from recognizing common objects like cars, dogs, or cats."*

Kit says the broader motivation for the work is the increasing reliance on AI systems to interpret the natural world. He recalls taking a photograph of a fish near the university's pier and asking an AI assistant to identify it. He was sure it was a boxfish, but the system suggested it was a pufferfish. Later, a marine expert confirmed the AI tool was mistaken. *"I was correct, actually, but I wasn't certain because I'm not a marine biologist,"* he recalls. *"Imagine if a normal user just thinks this is a pufferfish because they trust the AI tool."* The experience reinforced his desire for deeper domain knowledge in AI systems. *"It gave me more confidence to work harder to really improve things."*

You can learn more about this work during Oral Session 3B: Image Recognition and Understanding I, Sunday 15:00–16:00 in AZ Ballroom 7, and during Poster Session 2, Sunday 16:00–17:45 in the Tucson Ballroom and Prefunction Space.


Can you guess who graduated and got engaged on the same day? Yes, both of us ;-)
read on…

Toward a Virtual Biopsy: Using MRI to Diagnose Cancer Without Pain

Eleftheria Panagiotaki (but everybody calls her Laura) is an Associate Professor at UCL Hawkes Institute and in the Department of Computer Science.

What is your work about, Laura?

My group, which is called CAMI for Computational Cancer Microstructure Imaging, is working on developing computational methods, usually for non-invasive diagnostic imaging. Now we're trying, with the new development of AI neural networks and so on, to add prognostication and some extra elements to help for a speedy diagnosis with the best possible prognostic outcomes for the patients.

How did you get there?

I studied mathematics in Greece at the University of Crete, and then I wanted to do something more applied. I had a very large choice, and I chose the MSc course for vision imaging and virtual environments, which was my introduction to computer vision and computer science at UCL. And after I finished this, I thought I was going to go into computer graphics because they were very cool and shiny and a virtual environment. But then I met Professor Daniel Alexander, who introduced me to the magic of medical imaging, and I never looked back.

Do you attend MICCAI sometimes?

Yes, we just submitted yesterday, 3am. My PhD students worked really hard!

It will be a best paper! How many are you?

I have two PhD students, one research associate, one senior postdoc, and one postdoc. We're a mix, female and male. A very good mix.

Is this on purpose?

No, no, we started like an all-female group, very small. But then we had some wonderful guys joining us. Very smart.

You started studying in Crete. Are you from Crete?

Yes, I am. I'm from Rethymno.

How is it to grow up in Crete? I only visited for 3-4 days, like everybody.

Well, finally, I didn't grow up in Crete because my dad was in the army. We were going there just for holidays, like everybody else, but all of my extended family is there.

Ballet - my other passion

I understand that you toured a lot, instead of growing up in one place. Tell us about this experience.

It was mainly all around Greece. During my primary and secondary school I had to move almost every year. Northern Greece, Cyprus, I made lots of friends everywhere. I think that's why, once I found a suitable place like UCL, I wanted to stay.

Isn't it difficult to leave all the friends every year, kind of forever?

Yes, yes, it was. And back then we were writing letters to each other. It was very sweet because I still have them. And with some of them we can still meet. I bumped into some of them in London. The world is a small place after all.

Tell us something we don't know about UCL.

I can't speak about UCL as a whole: I was very lucky to be in a group which was always within a center that now became an institute. We were CMIC, Center for Medical Image Computing, run by Professor David Hawkes. We were an extended big academic family, which became even bigger, hence the Hawkes Institute named after Professor Hawkes.

Let's talk about the future, Laura. Where are you going?

I don't know if this is a secret or not, but I'm going to tell you: I applied for my professorship. Now I'm waiting for the results in the summer. I'm hoping bigger and better. You know, it is true that with this position you get more power in terms of getting more grants and being more self-efficient. One of the things that is going really well is a method I developed called Verdict MRI. It is specifically for non-invasive imaging of tumors. The first clinical trial was in prostate. We are now in the third clinical trial, which is a multicenter trial that just started. This is very promising and very high impact!

Do you hope that one day it will be used in the clinics?

Yes! The method was never patented. We followed the NHS guideline route to get it out there as soon as possible. And I'm working with UCLH, which is the hospital associated with UCL. It's going really well! Yeah, we're hoping to get it out in the clinical routine soon.

It sounds very exciting!

It is very exciting. It's a method that will potentially reduce significantly the amount of unnecessary biopsies that can cause a lot of side effects. They're painful. And you get the results.

People worry for that. They wait one month to get the result.

Exactly! The histopathology results! We won't have to wait for that. We're hoping to get this virtual biopsy just with the MRI, without

any pain. Just a bit of noise from the scanner.

Is it only your group embarked in this project?

Oh, no! This is huge! The method was invented toward the end of my PhD when I was doing neuroscience. Then I had this idea of translating it to tumors and cancer. It was a prototype, very computational. We made animal experiments first and had positive results. Then David Hawke said, why not? Let's try and do it in humans!

It might become a spinoff of the university.

Who knows? Yes, maybe. My goodness, possibly! We do collaborate with a lot of companies, like Siemens and Philips, to incorporate it in the scanner, because that will be the best, fastest route.

I did not forget your professorship. How does it work, actually, for an Associate Professor who one day says, well, I will apply to professorship?

There is guideline from UCL. For that grade, all the achievements that you have under your belt from research, from teaching, from your EDI contributions, institutional citizenship and all these things. I was a bit hesitant and was going to wait for a few more years, but I have wonderful colleagues who pushed me forward: they told me I was ready and I should try. There's a shortlisting from the department. And then once you get through that, you can apply for the whole thing. And then you need champions, other professors from other universities who know of you and your contribution to the field to vouch for you.

Tell me, Laura, this sounds a little bit bureaucratic. Isn't it something stranger that comes to bother you from your research?

Oh, there's so many things that come into the research. This is nothing. With this, at least, you get a fancy title that comes with power and opportunities. So this is something worth doing and I found it really good for my self-esteem as well. Doing this little exercise, looking back and putting together everything that I have done and presenting it in the best possible way, because sometimes you forget... And many of us suffer sometimes with imposter syndrome. Am I good enough? We compare ourselves with others. So when you have also that external validation, it's a little boost. It's quite a big boost, actually!

But tell me, Laura, how can someone obviously successful like you, someone with your curriculum, still have imposter syndrome? How does it work?

I think quite a few of us have it because we're surrounded by brilliant, really competitive [people]. It's a very competitive field. It's getting even harder to survive. Everyday life comes as an obstacle to your research. I think that's the main thing: trying to balance everything. Of course, you doubt. We're humans. Sometimes we break down.

And this [candidature for professorship] is a combination of having lovely colleagues as well. Because as I said, I wasn't going to go for it this year by myself. And then maybe next year, I wouldn't think that I am quite ready. Everybody in research is quite different. The previous people who got their professorship had a very different CV from me. You can't compare and be certain that, yes, this is the right time for me. So it's good to have academic mentors who really care and can provide independent advice.

"That's my dream: to keep having amazing people surrounding me and trying to improve health care!"

Laura, let's go further into the future. What is your dream?

Oh, I haven't. So far, I've been limited to [dreaming of] getting my professorship and having a group and doing the research that I've always wanted with unlimited funds. That's the dream. I think I'm a little bit more down to earth now. I would like to work with people I love working with, because life is so much better. And I got a little glimpse of that when I created my CAMI group. I had wonderful people working all together, passionate for the research, but also friendly and very easygoing. Our dreams and aspirations matched. It's much easier to work hard with like-minded people. So I think that's my dream: to keep having amazing people surrounding me and trying to improve health care!

What would be a great achievement for you by the end of your career?

Oh, to produce as many brilliant researchers and academics as possible. Definitely, definitely. I see myself as a mentor. I really care about the people. I was very lucky myself. I want to be one of these people that can help others.

"We're hoping to get this virtual biopsy just with the MRI, without any pain. Just a bit of noise from the scanner..."

Can you tell me one impressive thing that you learned from your students?

One impressive thing? Quite a few of them. They are really, really fantastic!

We have time.

I think one of the things that they're proving to me day after day is that you can have everything. You can have your hobbies. You can be brilliant. You can be an excellent computer scientist. Have a life outside academia. Try and get into industry and keep your foot in academia and pursue a family life. And achieve it all with a smile on your face. Several people in my group to do this. I think that's the way forward.

Your final word to the community.

Thank you. Well, I'm looking forward to meeting as many of you as possible. The research that comes out still amazes me. And that's why I enjoy most when I go to conferences, meeting people and seeing what they're passionate about. We're really unique and really lucky. And I hope I remain in this space for a little longer to enjoy it!

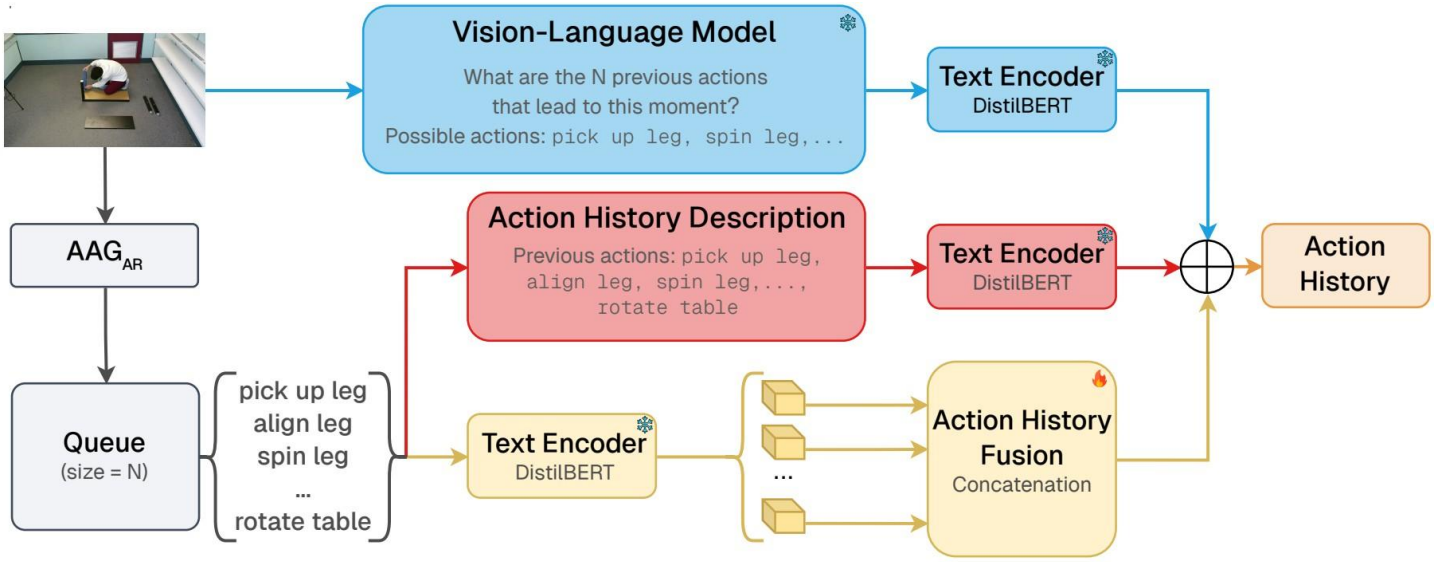
Read 160 FASCINATING interviews with Women in Computer Vision!

Action Anticipation at a Glimpse: To What Extent Can Multimodal Cues Replace Video?

Manuel Benavent-Lledo is currently a postdoc researcher at the University of Alicante.

He speaks to us ahead of his poster presentation later today.

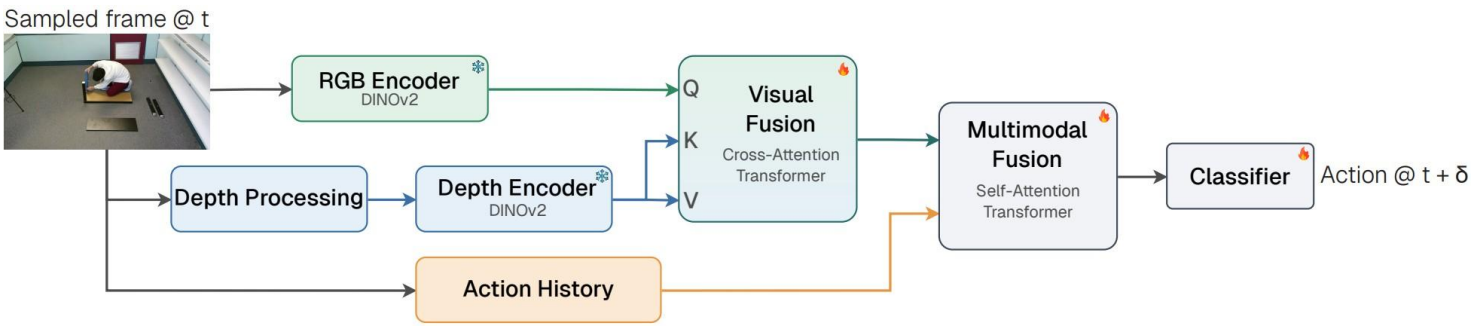
To what extent can multimodal cues replace video? The main goal of Manuel's work is to study whether **the information contained in a single frame is enough to substitute video aggregation in some context**. "*We first analyze what a single frame can provide,*" Manuel explains "*and we study also how multimodal cues can enhance the information provided by this single frame. In particular, we study how depth information can improve this frame and how long-term context extracted by either a visual language model or the action history by previous observations can contribute to enhance the next action anticipation.*"

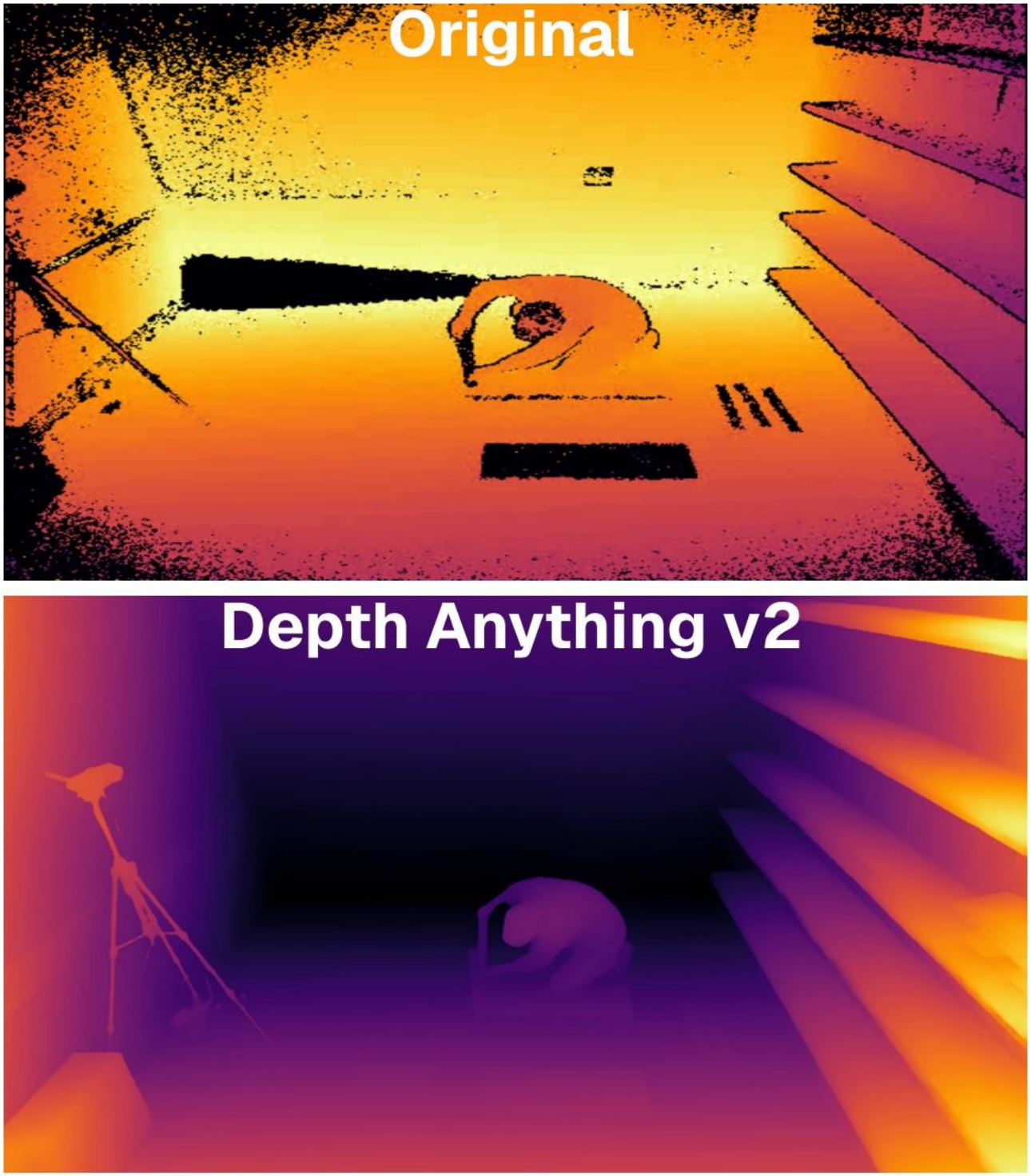


But why do we need this? In some contexts, latency and real-time understanding is critical. Frame aggregation can be very costly and computationally expensive and a single frame in such context can really help in the tasks.

The main difficulty for the authors was the fusion of the different modalities. They had to find the right strategies to combine them and also find the right features. It's not the same to use pre-trained backbones or use self-supervised backbones. It has a critical impact on the final results. But also, and more importantly, they realized that the single frame approach has an enormous impact on the dataset complexity. So on simpler datasets, this method can work really well. But in contrast, in some more complex scenarios where the variability is more complex, such as ADL, we need the information from video in such cases.

How did Manuel and team solve the problem? Their method is based on three different branches: "*First, we have the RGB branch*", he specifies. "*We extract features from a self-supervised transformer, in this case DINOv2, to fuse them with depth features to include spatial information. And for this purpose, we use a cross-attention transformer.*"





He soon realized that the visual information is on its own and is limited, so he needed long-term information, which he got from what the team calls an **action history**. *"We explore three different ways to extract the action history. One, it's from a visual language model, prompting it to obtain what happened in the past. And another way is we reuse the features that we had, the visual features that we had for the anticipation, to perform a recognition, and we start to create an action history. So we use the past information, as in a real-time setting, to build an action history and to encode the actions and fuse*

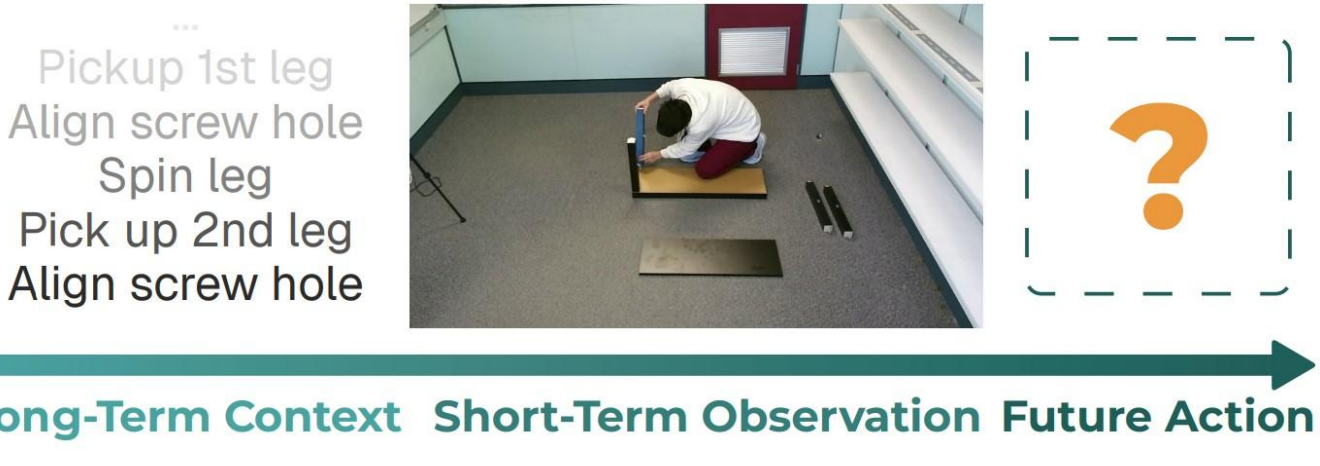
them with the visual information."

Manuel claims that the research in the past involved few works, including single-frame action recognition or action anticipation, but he decided to explore this as a first step into the single-frame anticipation. In fact, the team is turning to future works through which they further develop this work.

The vision part of this work is mostly composed of two main components. One is the visual encoder, whereas the same encoder is used for RGB and depth. Manuel evaluated different encoders, and he realized that a self-supervised encoder provided the best results. That is as expected because self-supervised captures the intrinsic data from the images rather than learn from pre-trained data sets. And also in the depth case, RGB encoders cannot use directly depth information because the depth map is single channel. "*We decided to apply a coloring strategy to convert the depth frame into an RGB one,*" Manuel clarifies. "*And lastly, for obtaining the depth image, we also realized that the ground truth depth can often be noisy. So instead of relying on original depth captures, we relied on depth estimation models. And in this case, it's Depth Anything V2. I wish to add that VLMs were quite limited for this task, rather than work very well as everyone expects them to work!*"

The idea of this work came from discussions with colleagues at ICS-FORTH in Greece, where Manuel was a Visiting Researcher two summers ago. In his regular work aside from this paper, he focuses on action understanding tasks and multimodal action understanding. He has done work related to recognition, online action detection and action dissipation; he also collaborated on different multimodal tasks outside action related tasks, but including computer vision, such as pain assessment and more.

Manuel will be presenting his paper during Poster Session 1, Sunday 11:15–13:00 in the Tucson Ballroom and Prefunction, poster 27.



Katrin Renz has recently completed her PhD at the University of Tübingen and the Max Planck Institute for Intelligent Systems (IMPRS-IS). Supervised by Andreas Geiger, her research focuses on end-to-end autonomous driving, specifically combining vision, language, and action to make self-driving cars smarter and more interpretable.

Katrin is currently building her own startup, continuing her work on embodied AI.

Congrats, Doctor Katrin!

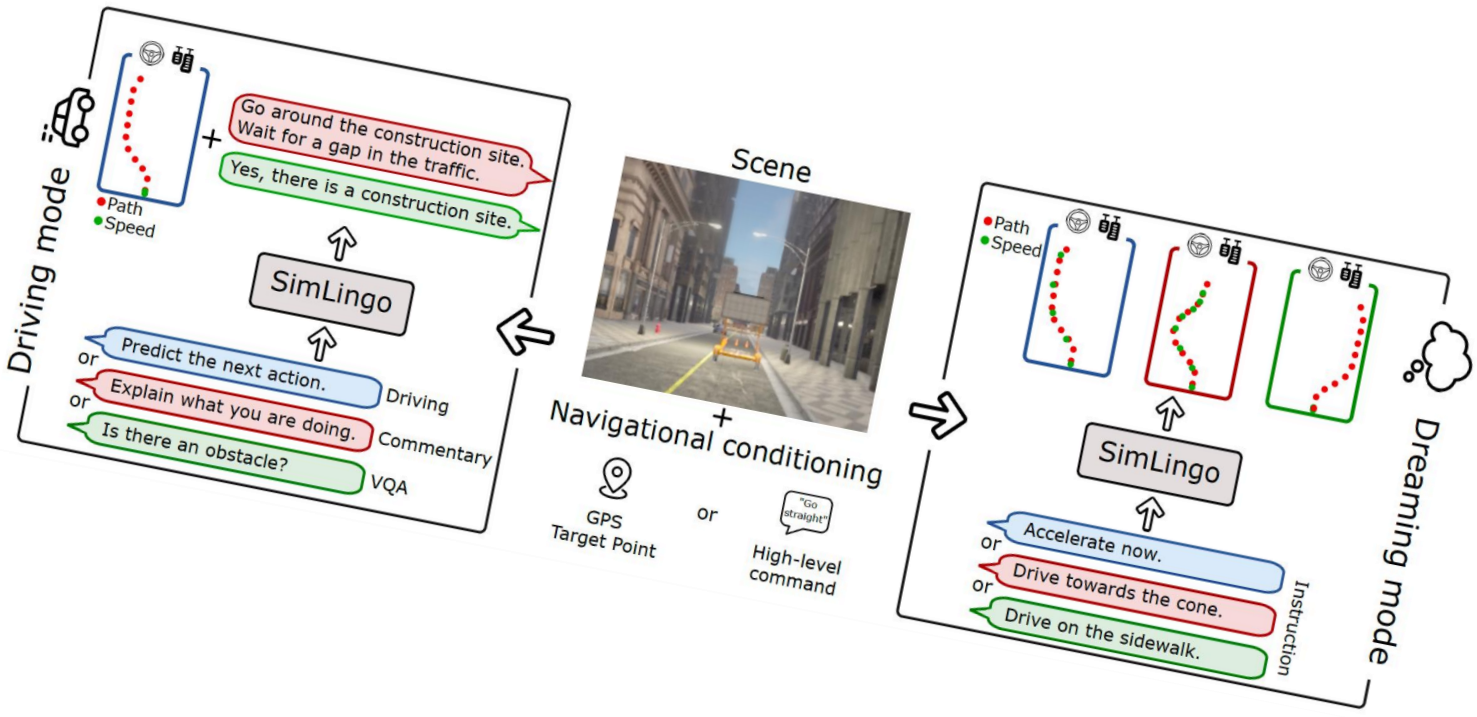
Autonomous driving promises to make roads safer, but building systems that can handle the full complexity of real-world traffic remains a major challenge. In particular, existing systems struggle to generalize to the "long tail" of rare or unusual driving scenarios. These infrequent but safety-critical edge cases are essential to master for reliable, real-world deployment. **Katrin's thesis tackles this gap by leveraging the broad world knowledge of foundation models and grounding their reasoning power in the physical world.** Her goal? To teach cars not just to drive safely, but to reason, explain their actions, and interact in natural language.

In her first major work, Katrin introduces **PlanT, a transformer-based planner**. Moving away from computationally heavy, pixel-level Bird's-Eye View (BEV) images, PlanT uses a compact object-level representation. By leveraging a standard transformer architecture inspired by language modeling, it achieves expert-level driving performance on the CARLA simulator. **Best of all, PlanT's attention weights make the decision-making process highly explainable, clearly highlighting the most relevant objects in a traffic scene without requiring extra manual annotations.**

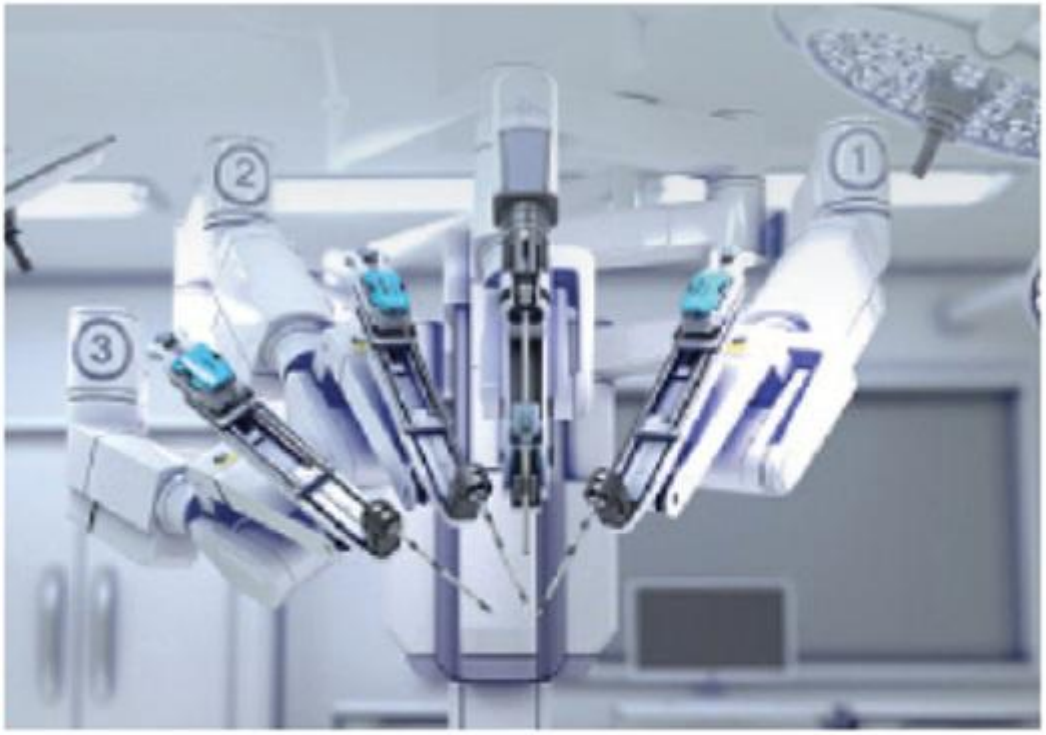
Her second project, DriveLM, directly integrates language as an additional modality, making it one of the pioneering works to explore Vision-Language-Action (VLA) models for autonomous driving. By bringing these components together into a unified architecture, she formulates driving as

a **Graph Visual Question Answering (GVQA) task**. Instead of an end-to-end black box, DriveLM mimics human reasoning step-by-step, linking perception, prediction, and planning through a logical graph. By leveraging vision-language models (VLMs), the car can explicitly answer questions like, "*What are the relevant objects in the scene?*" before executing an action.

In her final work, completed during her internship at Wayve, she **builds a full Vision-Language-Action (VLA) model named SimLingo.** SimLingo achieves state-of-the-art closed-loop driving performance while seamlessly unifying multiple capabilities within a single framework: autonomous vehicle control, visual question answering (VQA), instruction following, and explanations of the driving decisions. To tackle the crucial challenge of language-action alignment, ensuring the model actually bases its driving decisions on language rather than relying solely on visual cues, she introduces **"Action Dreaming," a novel training task utilizing diverse instruction-action pairs. Thanks to these innovations, SimLingo became the winning entry at the 2024 CARLA Autonomous Driving Challenge at CVPR.**



By successfully uniting vision, language, and action, Katrin's research paves the way for generalist autonomous agents that we can actually talk to and understand. For more information, see her website (katrinrenz.de).

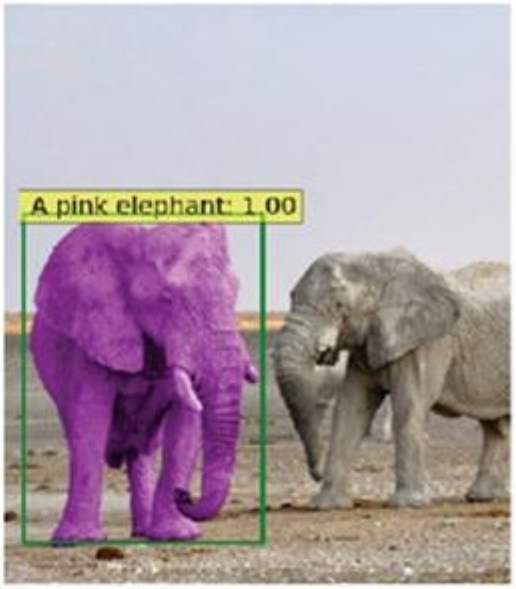


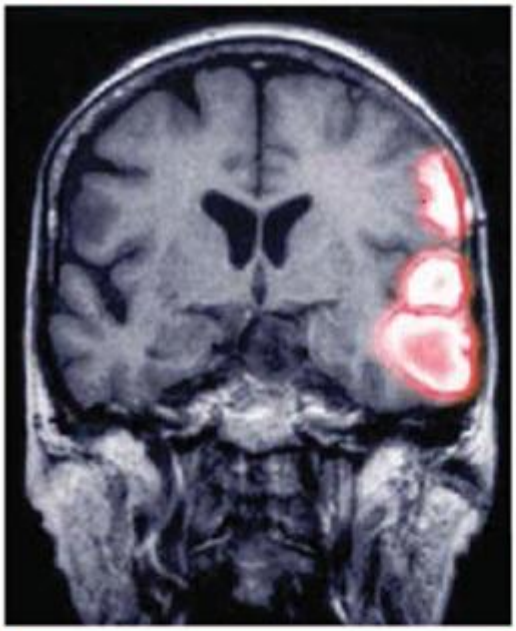
IMPROVE YOUR VISION WITH

Computer Vision News

SUBSCRIBE

to the magazine of the algorithm community





Meet the Scientist behind the Science