Large Vision Model (LVM): An Exploration of Vision Models with Visual Sentences of Images and Videos

Presenter: Muhammad Hamza Zafar
Affiliation: NTNU, Norway
Co-Author: Vijeta Sharma

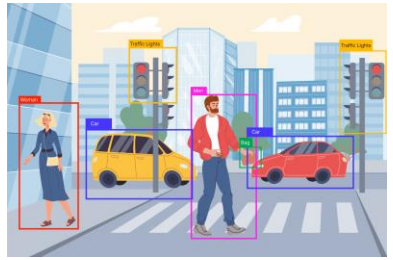
NTNU

Motivation

- Why Video Understanding Needs a New Approach

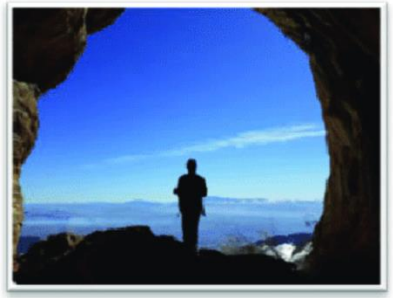
1. Video tasks: complex, dynamic & resource intensive
2. Task-specific models = inefficient for real-time systems
3. We explore **LVMs** as a **unified solution**

1. YOLO → Object Detection



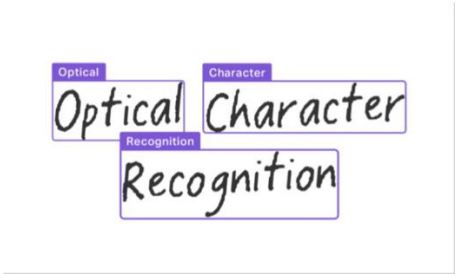
2. SAM → Segmentation

3. GPT4o → Image captioning

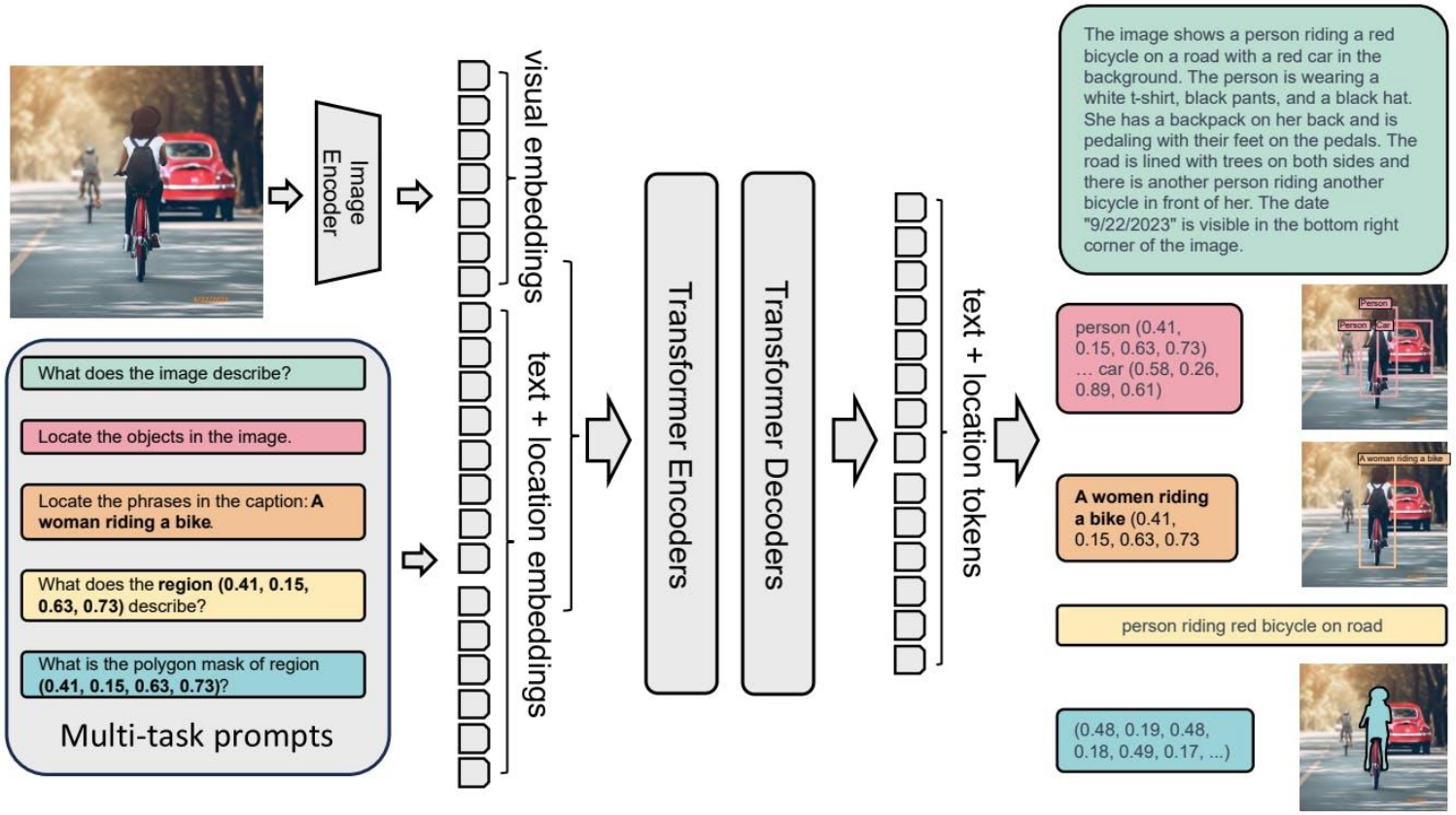


It's a man standing on a rocky hill.

4. Mistral OCR → OCR



Exploring LVMs as a unified solution

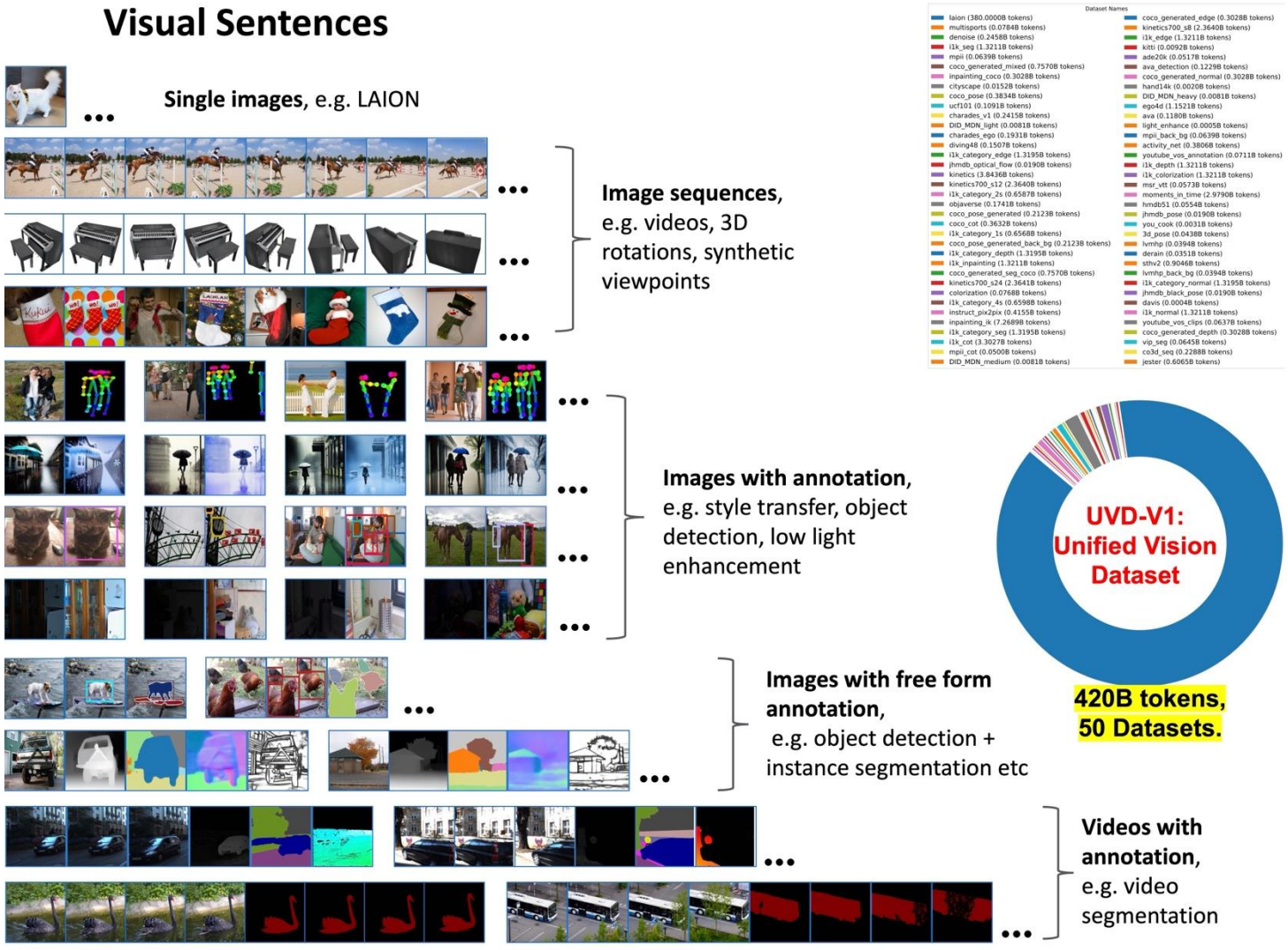


Florence 2

REF: **Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks**
Xiao et al., arXiv:2311.06242, https://arxiv.org/abs/2311.06242

What is a Visual Sentence?

- Visual Sentence = Ordered sequence of visual tokens (patches, queries, masks, etc.)
- Serializes video frames like a paragraph in NLP
- Inspired by the concept of “visual sentences” from Bai et al., 2023



Ref: Bai, Y., et al. (2023). Sequential Modeling Enables Scalable Learning for Large Vision Models. arXiv:2312.00785. https://arxiv.org/abs/2312.00785

Research Goals

- **Evaluate generalization** of LVMs on video tasks
- **No fine-tuning**: Test out-of-the-box models
- **Tasks**:
  1. Detection
  2. Segmentation
  3. Captioning
  4. OCR
- **Models**:
  1. Florence-2
  2. GPT-4o
  3. LLaVA
  4. PaliGemma

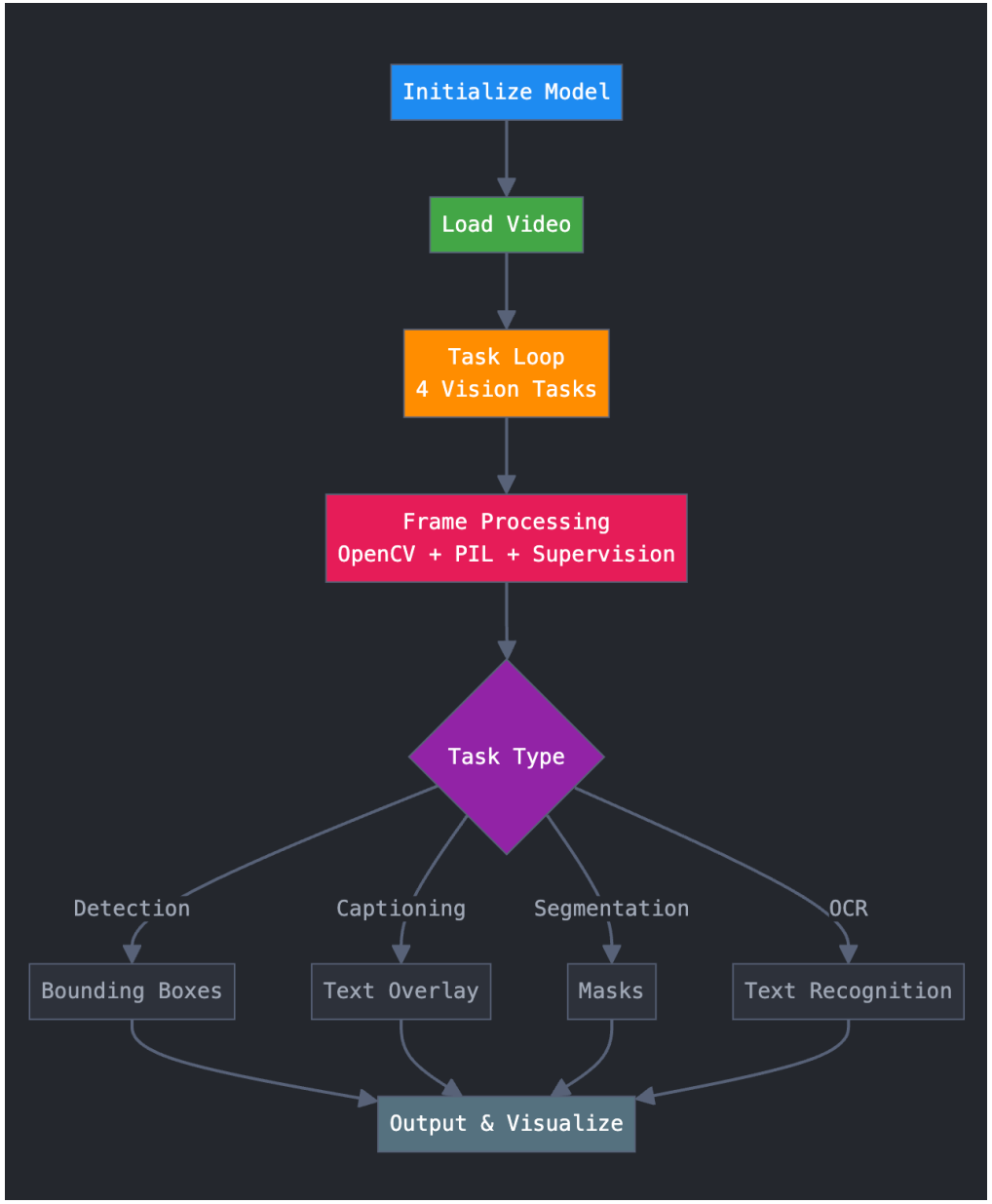
Model	Task	License	Parameters
Florence-2	Captioning, Detection, Segmentation, OCR and others	MIT	230M, 770M
PaliGemma	Captioning, Detection, Segmentation	MIT	3B
LLaVA	Captioning, OCR	Apache-2.0	13B
GPT-4o	Captioning, OCR	Proprietary	X

Dataset

- Two curated **Pexels** videos (controlled)
- 50 samples from **UCF101** (real-world, dynamic)

Processing Pipeline

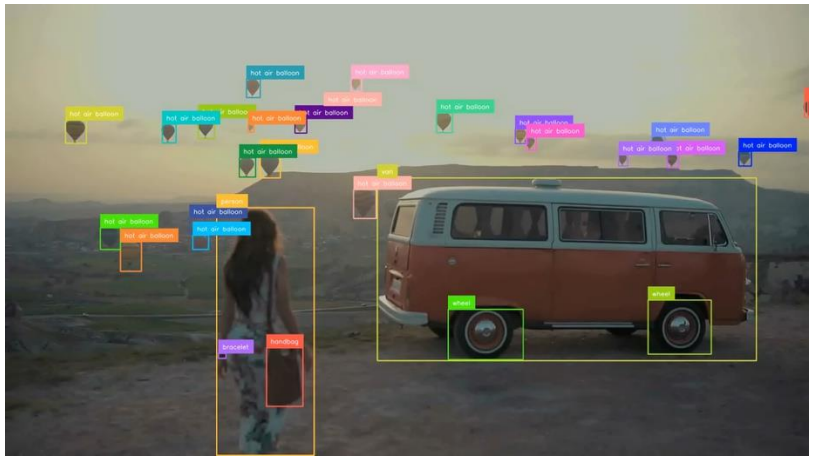
- Extract → Resize → Normalize frames
- Feed to LVMs (no fine-tuning)
- Collect outputs (BBs, masks, text)
- Reassemble annotated video



Results – Pexels (Qualitative)
Florence 2 – Vid 1

- Accurate detection & captioning in clean scenes
- Segmentation is functional but slow
- Unified output from a single model

Image Captioning



Object Detection

Segmentation

Results – Pexels (Qualitative) Florence 2 – Vid 2

- OCR works well on clear text
- Captioning is good
- Segmentation struggles with motion
- Object detection also shaky with motion

OCR

Captioning

Object Detection

Segmentation

Results – Pexels (Qualitative) PaliGemma – Both Vids

- Strong segmentation performance
- Detection is stable but not better than Florence 2
- Captioning remains good

hot_air_balloons: Segmentation

hot_air_balloons: Object Detection

hot_air_balloons: Captioned

november_leaves: Captioned

november_leaves: Object Detection

november_leaves: Segmentation

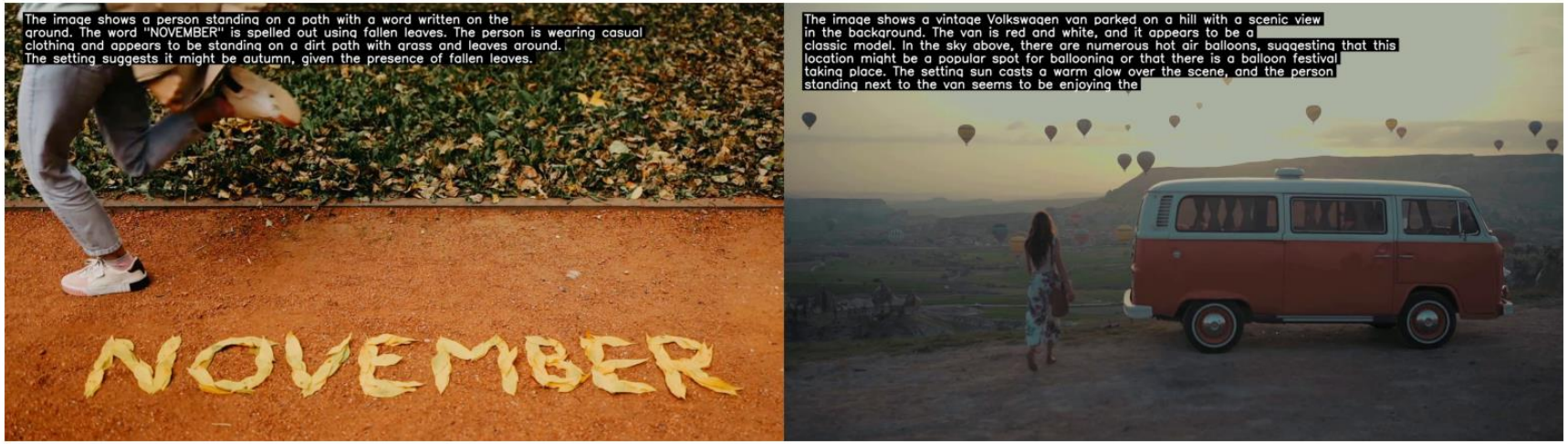
Results – Pexels (Qualitative)
Other Models

- GPT-4o: Fluent captions.
- LLaVA: Detailed Captions.
- Neither model handles all tasks

Middle Frame from Each Video

GPT-4o

Middle Frame from Each Video

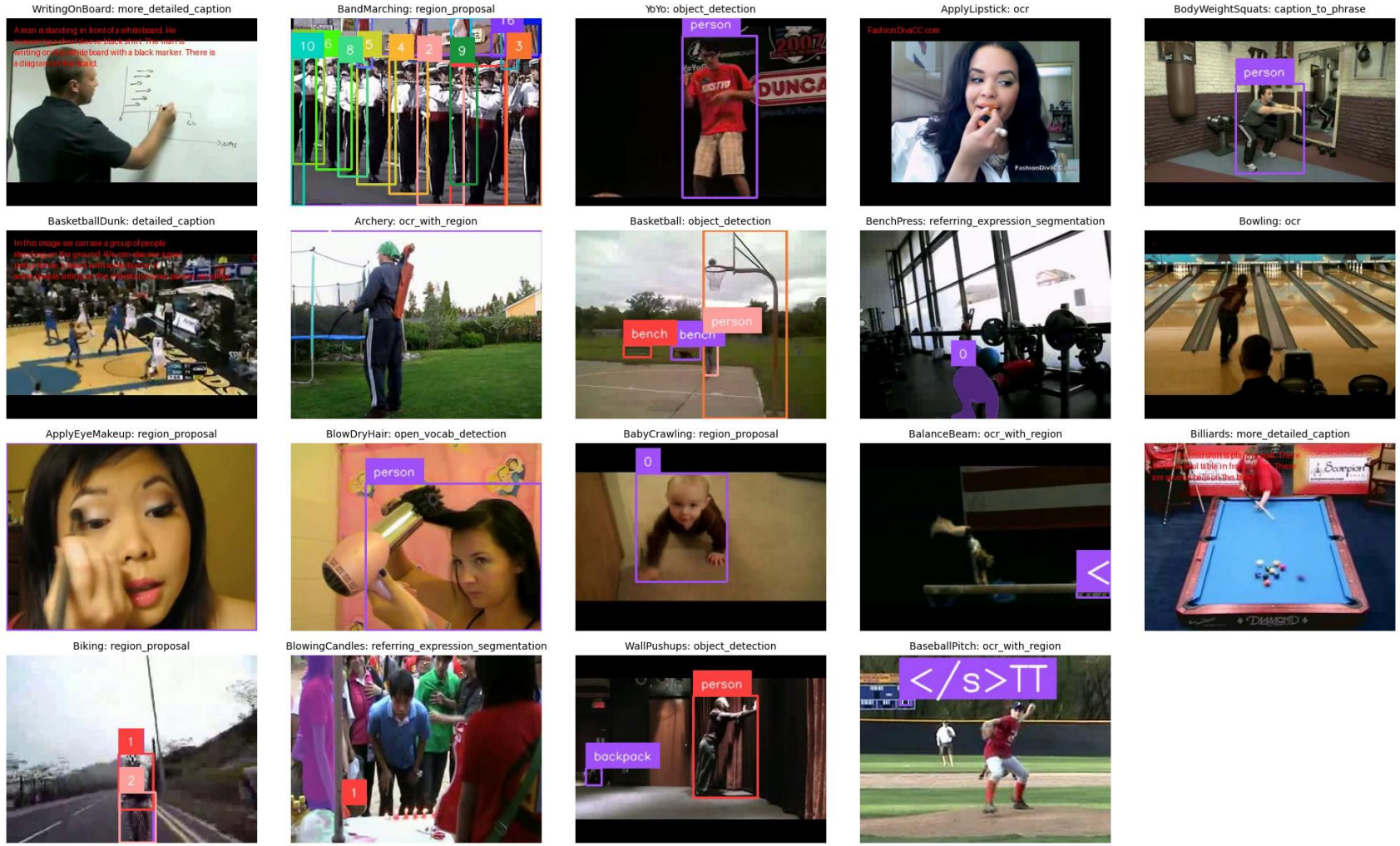


Llava

Results – UCF101: Florence 2

Total processing time for all 54 videos (696.29 minutes)

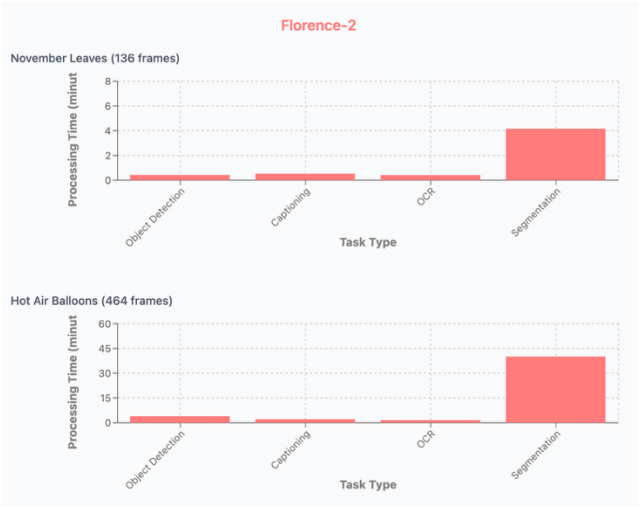
- All 4 tasks
- Full multi-task output on real-world videos
- High scene diversity exposed model limits
- Total processing time: ~11.5 hours

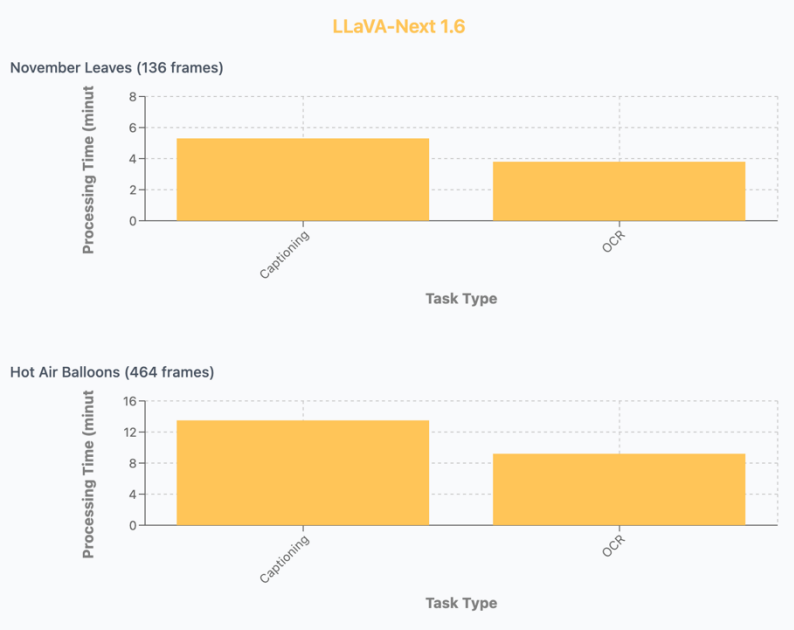


Runtime Breakdown

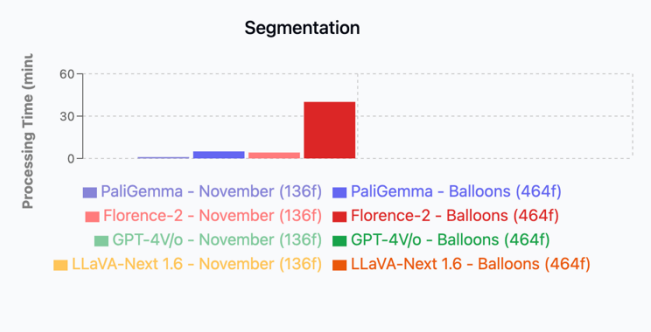
Performance Winners:

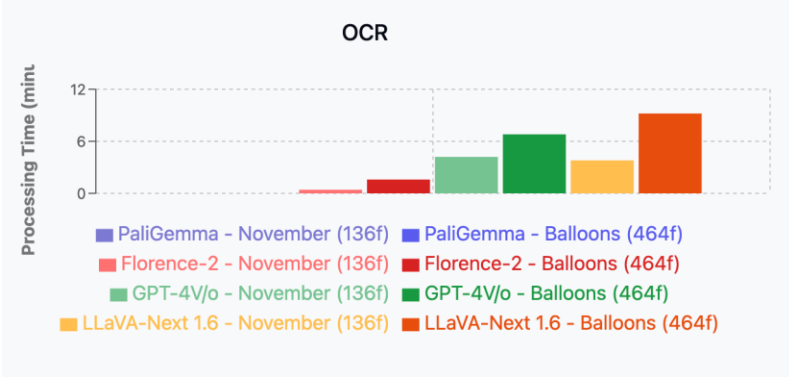
- **Object Detection & OCR**: Florence-2 dominates (0.42-3.94 min)
- **Segmentation**: PaliGemma wins (consistent 1.00→5.00 min scaling)
- **Long Videos**: GPT-4V/o best overall (14.30 min vs 47.71 min Florence-2)





Runtime Breakdown





Critical Insight:

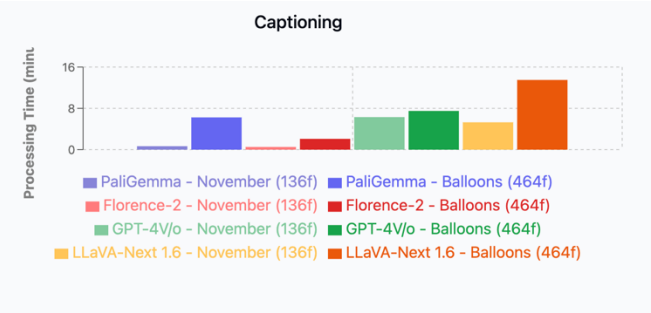
- Florence-2 segmentation completely breaks at scale
- **Short video**:
  - 4.16 min → **Long video**: 40.09 min (10x scaling disaster!)

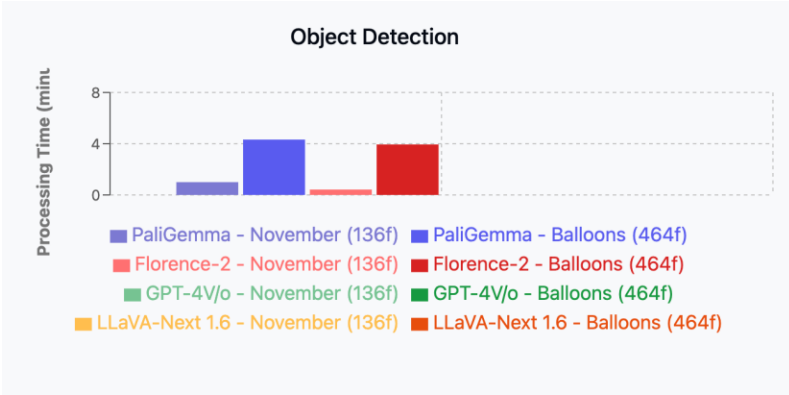
Task Champions:

- **Segmentation**: PaliGemma only viable option (1.00→5.00 min)
- **Object Detection**: Florence-2 fastest (0.42-3.94 min)
- **Captioning & OCR**: Florence-2 dominates when it works

Scaling Winners:

- **Most Predictable**: PaliGemma (linear scaling across all tasks)
- **Best for Long Videos**: GPT-4V/o & LLaVA-Next (flat scaling)
- **Avoid**: Florence-2 for any segmentation on long content





Discussion

- **Florence-2** was the most **balanced performer**, but segmentation scalability was a major weakness
- **PaliGemma** showed **consistent segmentation and predictable scaling**, but lacked strong OCR
- **GPT-4V/o** produced **fluent captions**, but hallucinated and was slow on segmentation
- Real-world videos (UCF101) exposed **generalization gaps** in all models
- Most models **work well on clean frames**, but break under motion blur, occlusion, or clutter
- True **video understanding** still requires model-specific tuning or post-processing
- Runtime and **scaling behavior** are just as important as task accuracy
- Unified models are **promising but not production-ready** yet

Conclusion & Challenges

Contributions

- Unified multi-task pipeline for video
- First Florence-2 application on videos
- Tested on curated + real-world datasets

Limitations

- No quantitative scores yet
- Downsampling may limit performance
- Segmentation is slow

Future Work

- Add mAP, mIoU, CIDEr benchmarks
- Test higher-res + real-time throughput
- Optimize for edge & embedded systems

References

- Xiao, T., Zhang, Y., Huang, Z., Yuan, L., Zhang, Y., Wang, X., & Yu, F. (2023). Florence-2: Advancing Unified Representation Learning in Computer Vision. arXiv:2309.17453
- Beyer, L., Sax, A., Radosavovic, I., Kolesnikov, A., Dohan, D., Misra, I., & Zhai, X. (2024). PaliGemma: A Versatile Vision-Language Model. arXiv:2403.13475
- Li, H., Zhang, P., Li, X., et al. (2024). LLaVA-Next: Tackling Multi-image and Video Understanding with Interleaved Visual Contexts. arXiv:2404.12396
- GPT-4V/o Technical Report: https://arxiv.org/abs/2303.08774
- Segment Anything: https://arxiv.org/abs/2304.02643
- Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. arXiv:1212.0402
- *OpenCV Library Documentation* — https://opencv.org
- *Free Stock Videos Dataset* — https://www.pexels.com/videos/
- https://deeplobe.ai/exploring-object-detection-applications-and-benefits/
- https://www.superannotate.com/blog/image-segmentation-for-machine-learning
- https://www.researchgate.net/figure/Some-examples-of-image-captioning-Each-caption-describes-the-image-above-it-These_fig2_347760884
- https://www.edenai.co/post/optical-character-recognition-ocr-which-solution-to-choose

Thank You

- 📧 muhamhz@stud.ntnu.no
- 📱linkedin.com/in/ihamzafer/
- 🧪 AIAI 2025 — Session 55
- 🤖 Code available on request
- 📱 *(Scan QR code to connect)*

- *Asked one question during the oral presentation:*
  - *What was the hardware used?*
  - *Cloud or in house?*
  - ***GPU 4090 RTX – in* house**

NTNU