EPFL

Mathis Group for Computational Neuroscience and AI

Benchmarking VLMs for Animal Behavior Analysis with LMMs-Eval

Extending LMMs-Eval to wildlife datasets.

Master student:

Lucie Huamani-Cantrelle

Supervisors:

Sepideh Mamooler
Prof. Alexander Mathis

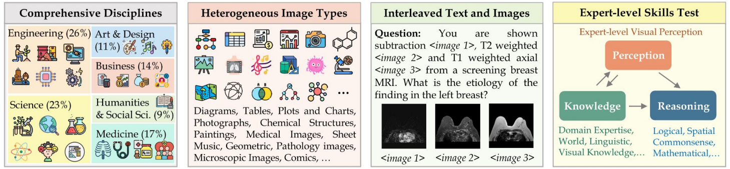
Summer Internship 2025

Lucie Huamani-Cantrelle

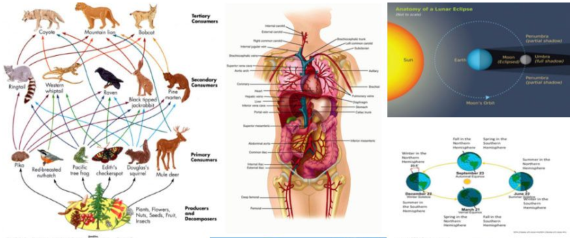
Motivation: Why this project?

Current benchmarks:

Common Objects in Context (COCO)



Massive Multi-discipline Multimodal Understanding (MMMU)



AI2D

- **Importance:** Animal behavior analysis underpins **ethology**, **ecology**, and **neuroscience**.
- **Gap:** Most VLM benchmarks are **human-centric**; ecological datasets remain largely unused.
- **Challenge:** Closed-vocabulary classifiers fail to generalize to unseen species and behaviors.

EPFL

Current problems

- **Vision-only baselines:** narrow, task-specific, non-generalizable.
- **Existing benchmarks:** lack standardization, suffer contamination, and are static, limiting reproducibility.

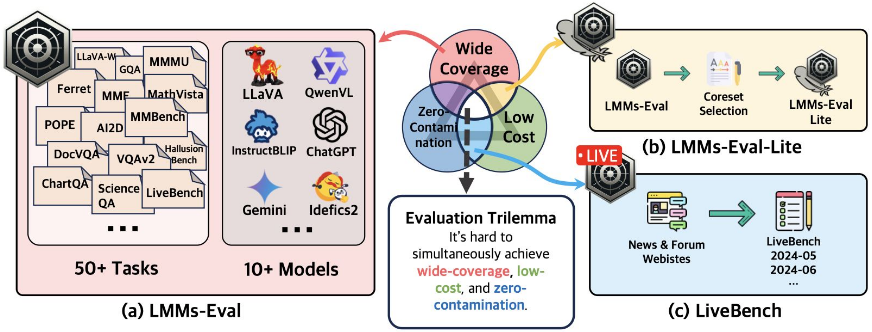
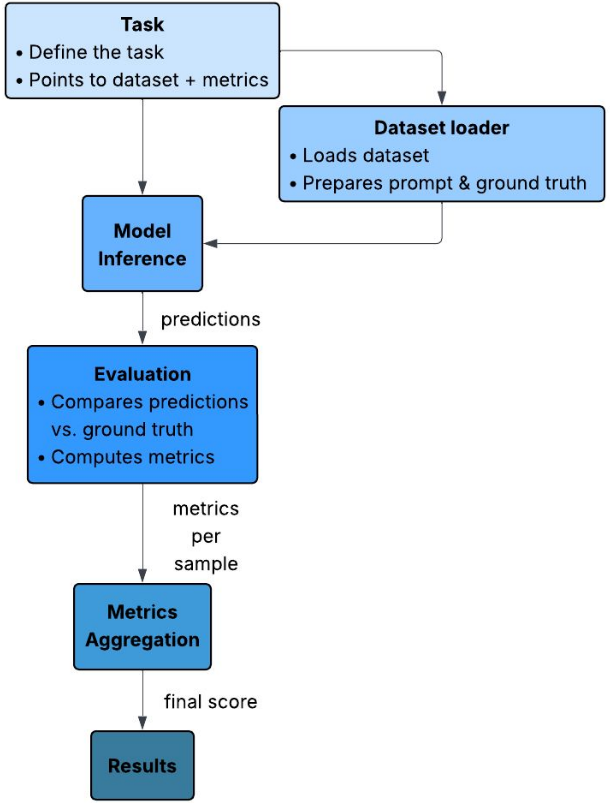


Figure 1. Overview of the LMMs-Eval framework combining 50+ tasks, 10+ models, a lite version, and LiveBench, addressing the evaluation trilemma (*adapted from Zhang et al., 2024*)

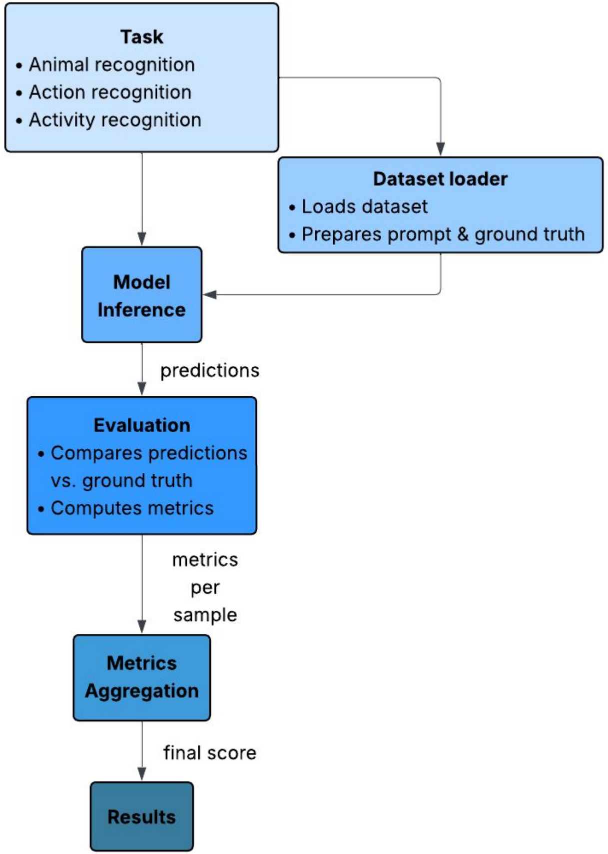
- What is the project?
  - Implement a species recognition and an animal behavior-oriented benchmark

Why



- Provides unified, transparent evaluation (50+ tasks, standardized metrics).
- Inspired by LM-Eval-Harness, extended to multimodal tasks.
- Enables reproducibility and comparability.

Project goals



→ Extend LMMS-Eval with **animal-specific benchmarks**:

1. animal recognition
2. action recognition
3. activity recognition

Dataset: MammAlps

- Collected with camera traps in **Swiss Alps**.
- 8.5 hours annotated video.
- **5 species**: red deer, roe deer, fox, wolf, mountain hare.
- **19 actions**, **11 activities**

Dataset: AnimalKingdom

- 50 hours of YouTube videos.
- **850 species**, **140 actions**.
- Large taxonomy spanning mammals, birds, reptiles.

Dataset builder

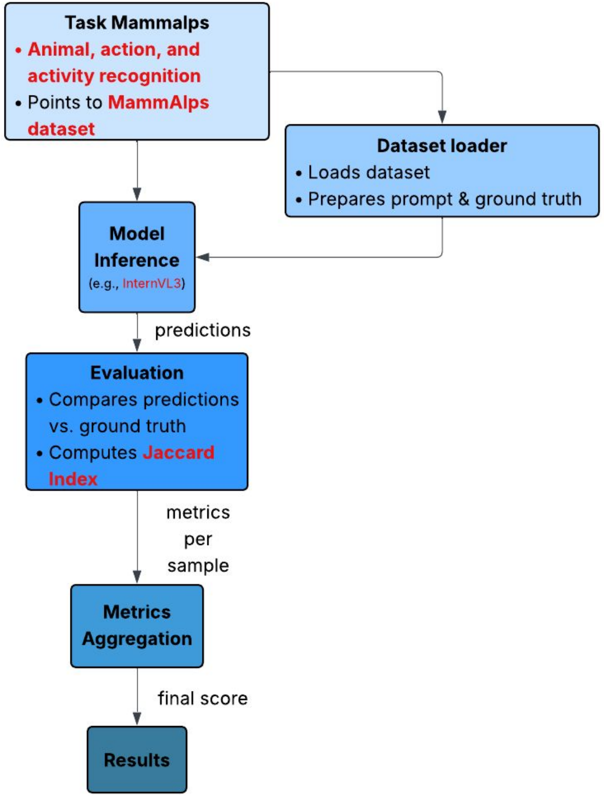
Example for MammAlps dataset:

```
mammalps/
├── mammalps_train_dataset.json
├── mammalps_test_dataset.json
├── clips/
│   ├── S1_C1_E4_V0016_ID1_T1/
│   │   ├── S1_C1_E4_V0016_ID1_T1_c0.mp4
│   │   └── S1_C1_E4_V0016_ID1_T1_c1.mp4
│   ├── S2_C7_E154_V0066_ID2_T2/
│   │   ├── S2_C7_E154_V0066_ID2_T2_c0.mp4
│   │   └── S2_C7_E154_V0066_ID2_T2_c1.mp4
│   └── ...
└── README.md
```

```
{
  "id": 42,
  "clip": "clips/S1_C3_E154_V0066_ID1_T1_c0.mp4",
  "video_id": "S1_C3_E154_V0066_ID1_T1_c0",
  "action": {
    "prompt": full prompt with <video> token and
output example,
    "answer": ["walking"]
  }
}
```

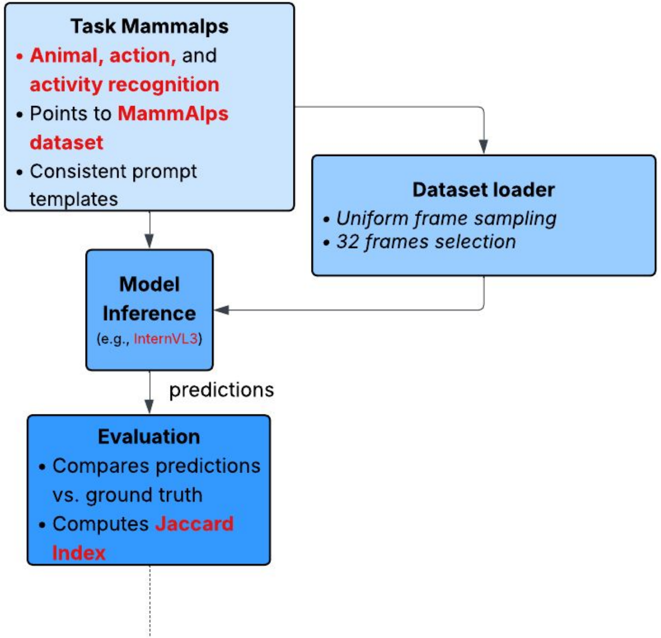
- Converts JSONL annotations into HuggingFace-ready datasets.
- Standardized folder structure.
- Ensures reproducibility across datasets.

Implementation in LMMs-Eval



- Defined tasks for animal, action, and activity.
- Registered strict Jaccard Index as evaluation metric.
- Outputs JSONL with id, prompt, full answer, extracted answer, ground truth, and scores.

Ensuring reproducibility



- Used **identical prompt templates** as workshop baselines.
- Controlled frame sampling:
  - Number of frames
  - **Uniform sampling**
- Ensures **fair comparison** and **reproducible** evaluations across datasets and models.

Results: MammAlps

Task	Workshop Jaccard	LMMs-Eval Jaccard	Δ Jaccard
Action Recognition	0.1843	0.2618	0.0775
Activity Recognition	0.3034	0.2716	-0.0318
Species Recognition	0.0878	0.1187	0.0309

Table 1: Results for MammAlps dataset on 1,244 test split records on InternVL3-8B in zero shot mode.

Results: MammAlps

Task	Workshop Jaccard	LMMs-Eval Jaccard	Δ Jaccard
Action Recognition	0.4678	0.3778	-0.09
Activity Recognition	0.6109	0.3955	-0.2154
Species Recognition	0.2122	0.1695	-0.0427

Table 2: Results for MammAlps dataset on 311 test split records on InternVL3-8B in zero shot mode.

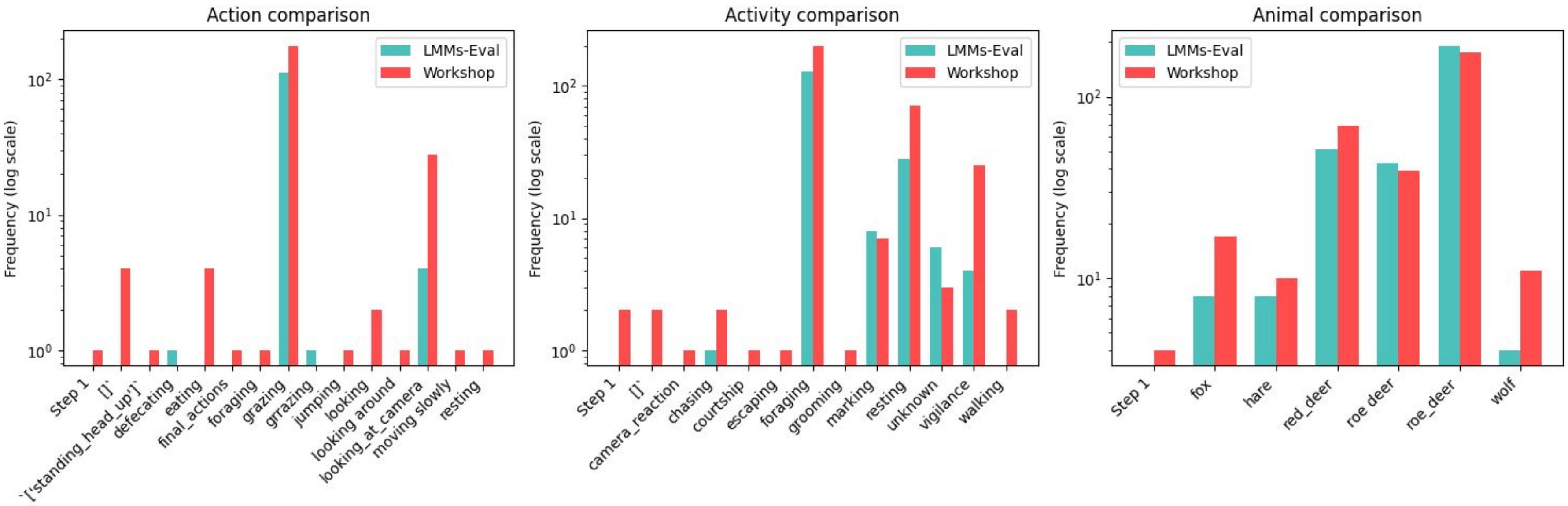


Figure 2: Comparison of Prediction Distributions: LMMs-Eval vs. Workshop Baseline on MammAlps Tasks.

Results: AnimalKingdom

Task	LMMs-Eval Jaccard
Action Recognition	0.3301
Animal Recognition	0.4197
Activity Recognition	0.1044

Table 3: Results for AnimalKingdom dataset on 6,096 test split records on InternVL3-8B in zero shot mode.

Additional experiments

- Adaptive frame sampling

Dataset	Standard Jaccard	Adaptive Jaccard	Δ (Adaptive − Standard)
MammAlps Full Test	0.2618	0.2494	-0.0124
MammAlps (311 samples)	0.3778	0.3961	0.0183

Table 4: Comparison results for MammAlps action recognition task.

Additional experiments

- Parsing improvements

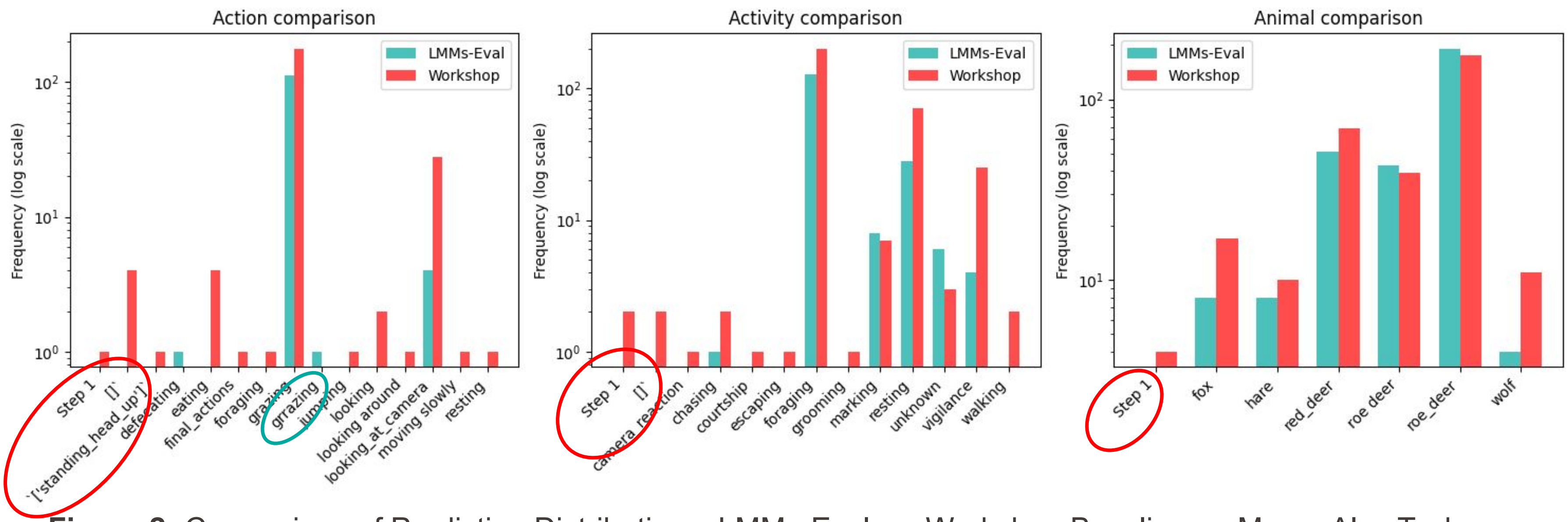
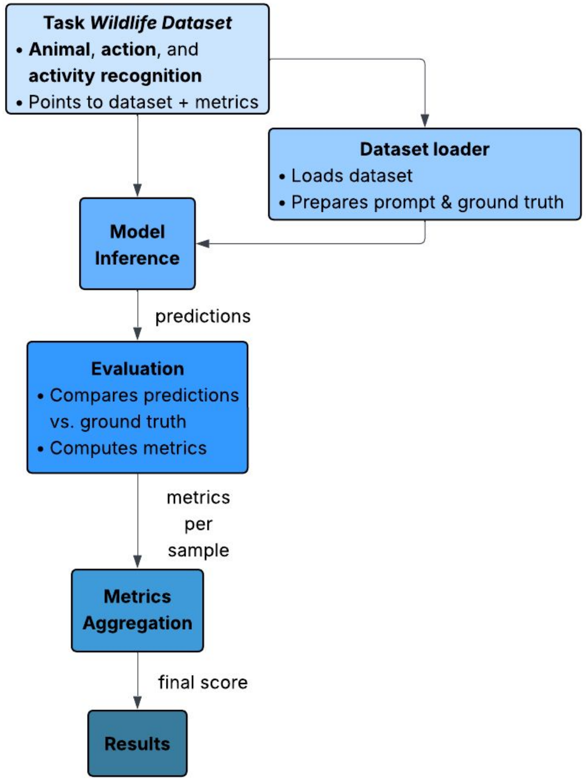


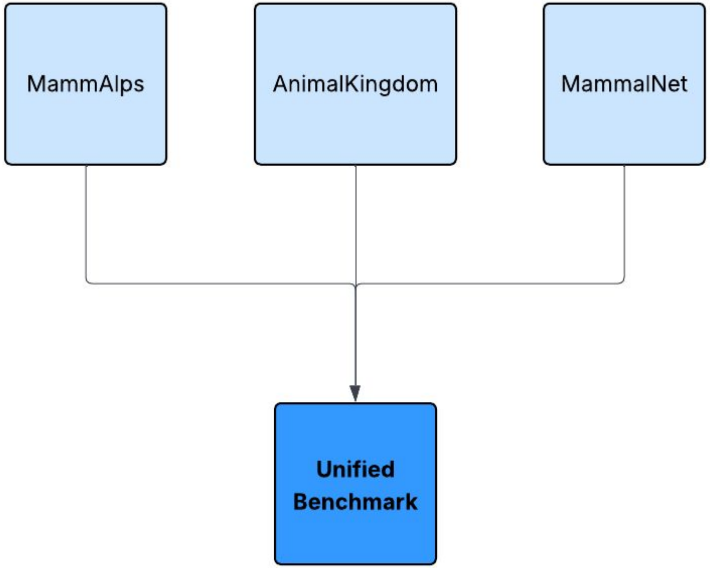
Figure 2: Comparison of Prediction Distributions: LMMs-Eval vs. Workshop Baseline on MammAlps Tasks.

Main takeaways



- Extended LMMs-Eval with 2 wildlife datasets.
- Registered Jaccard Index in LMMs-Eval
- Registered InternVL3-8B in LMMs-Eval
- Achieved standardized, reproducible evaluation pipeline.

Outlook



- **Next steps**
  - Investigate gaps.
  - Complete MammalNet integration.
  - Compare broader range of VLMs.


THANK YOU I SAY
GRATEFUL I AM

References

- Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. International Conference on Learning Representations (ICLR). arXiv:1405.0312. https://arxiv.org/abs/1405.0312
- Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., … Chen, W. (2024). MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://arxiv.org/abs/2311.16502
- Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., & Farhadi, A. (2016). A diagram is worth a dozen images. arXiv preprint arXiv:1603.07396. https://arxiv.org/abs/1603.07396
- Gabeff, V., Qi, H., Flaherty, B., Sumbül, G., Mathis, A., & Tuia, D. (2025). MammAlps: A multi-view video behavior monitoring dataset of wild mammals in the Swiss Alps. Conference on Computer Vision and Pattern Recognition (CVPR). arXiv:2503.18223. https://arxiv.org/abs/2503.18223
- Ng, X. L., Ong, K. E., Zheng, Q., Ni, Y., Yeo, S. Y., & Liu, J. (2022). Animal Kingdom: A large and diverse dataset for animal behavior understanding. Conference on Computer Vision and Pattern Recognition (CVPR). arXiv:2204.08129. https://arxiv.org/abs/2204.08129
- Zhang, K., Li, B., Zhang, P., Pu, F., Cahyono, J. A., Hu, K., Liu, S., Zhang, Y., Yang, J., Li, C., & Liu, Z. (2025). LMMs-Eval: Reality check on the evaluation of large multimodal models. arXiv:2407.12772. https://arxiv.org/abs/2407.12772
- Duan, H., Fang, X., Yang, J., Zhao, X., Qiao, Y., Li, M., Agarwal, A., Chen, Z., Chen, L., Liu, Y., Ma, Y., Sun, H., Zhang, Y., Lu, S., Wong, T. H., Wang, W., Zhou, P., Li, X., Fu, C., … Chen, K. (2025). VLMEvalKit: An open-source toolkit for evaluating large multi-modality models. arXiv:2407.11691. https://arxiv.org/abs/2407.11691
- Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Duan, Y., Tian, H., Su, W., Shao, J., Gao, Z., Cui, E., Cao, Y., Liu, Y., Xu, W., Li, H., Wang, J., Lv, H., Chen, D., … Wang, W. (2025). InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv:2504.10479. https://arxiv.org/abs/2504.10479