Dolphins : Multimodal Language Model for Driving

Yingzi Ma

Dolphins: Multimodal Language Model for Driving

- **We introduce Dolphins, a novel vision-language model architected to imbibe human-like abilities as a conversational driving assistant.**
  - Dolphins excel with abilities mirroring human drivers, from nuanced decision-making to fast adaptation with few-shot demonstrations (in-context learning) and error correction. Engineered for versatility, Dolphins also adeptly handles a wide range of driving tasks.

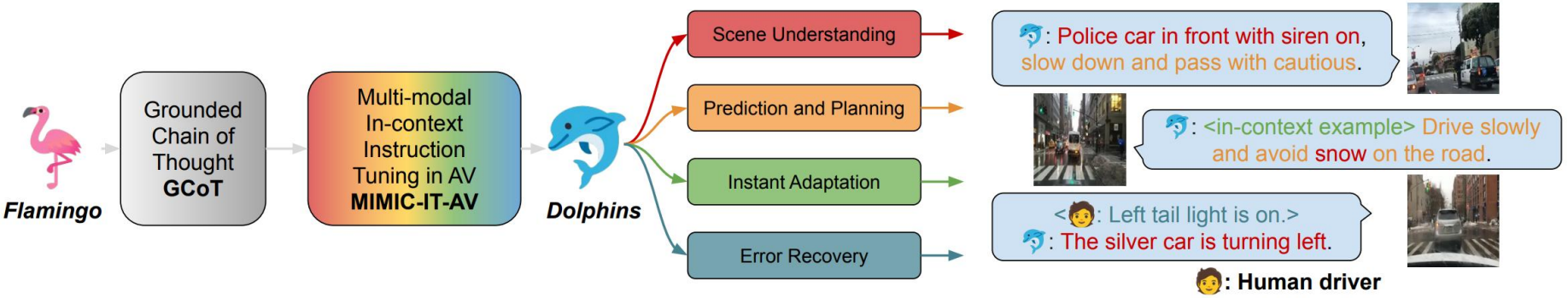


Figure 1. **Dolphins overview.**

Ma, Y., Cao, Y., Sun, J., Pavone, M., & Xiao, C. (2023). Dolphins: Multimodal Language Model for Driving. ArXiv, abs/2312.00438.

Dolphins: Multimodal Language Model for Driving

- 🔶 Featuring the Grounded Chain of Thought process, Dolphins refine its reasoning, aiming for a more nuanced understanding of complex driving contexts, similar to human thought processes. At the same time, it also helps VLM to emerge unseen tasks from the instruction tuning.

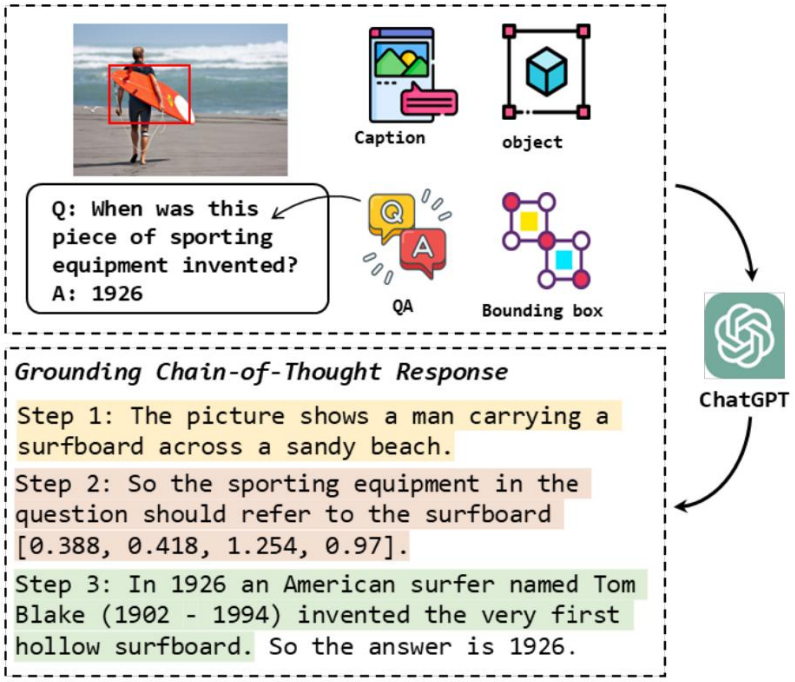
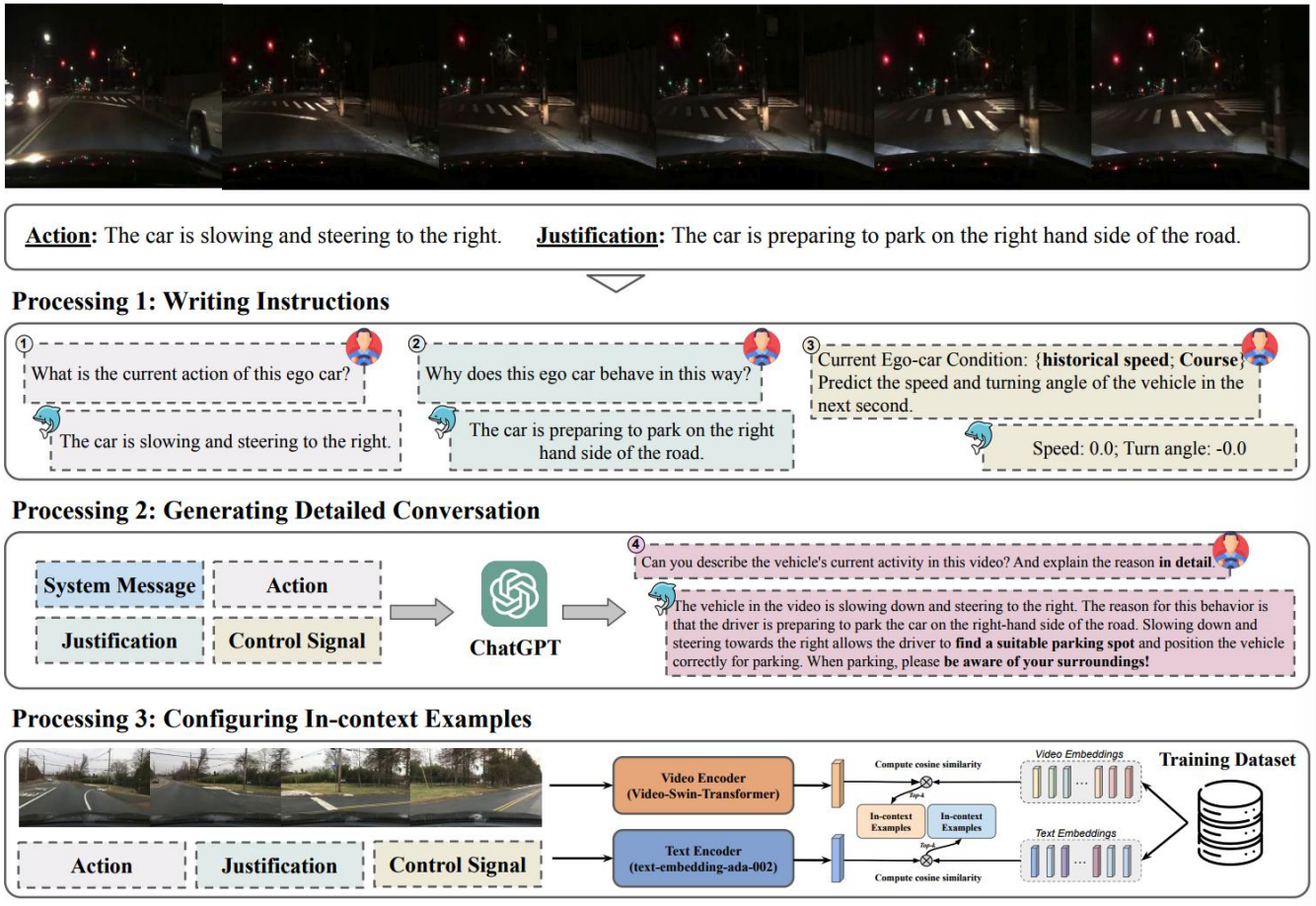


Figure 2. The process of generating GCoT response for VQA tasks to enhance the fined-grained reasoning capability of VLMs. ChatGPT is prompted to generate GCoT step by step from text input.

Ma, Y., Cao, Y., Sun, J., Pavone, M., & Xiao, C. (2023). Dolphins: Multimodal Language Model for Driving. ArXiv, abs/2312.00438.

Dolphins: Multimodal Language Model for Driving

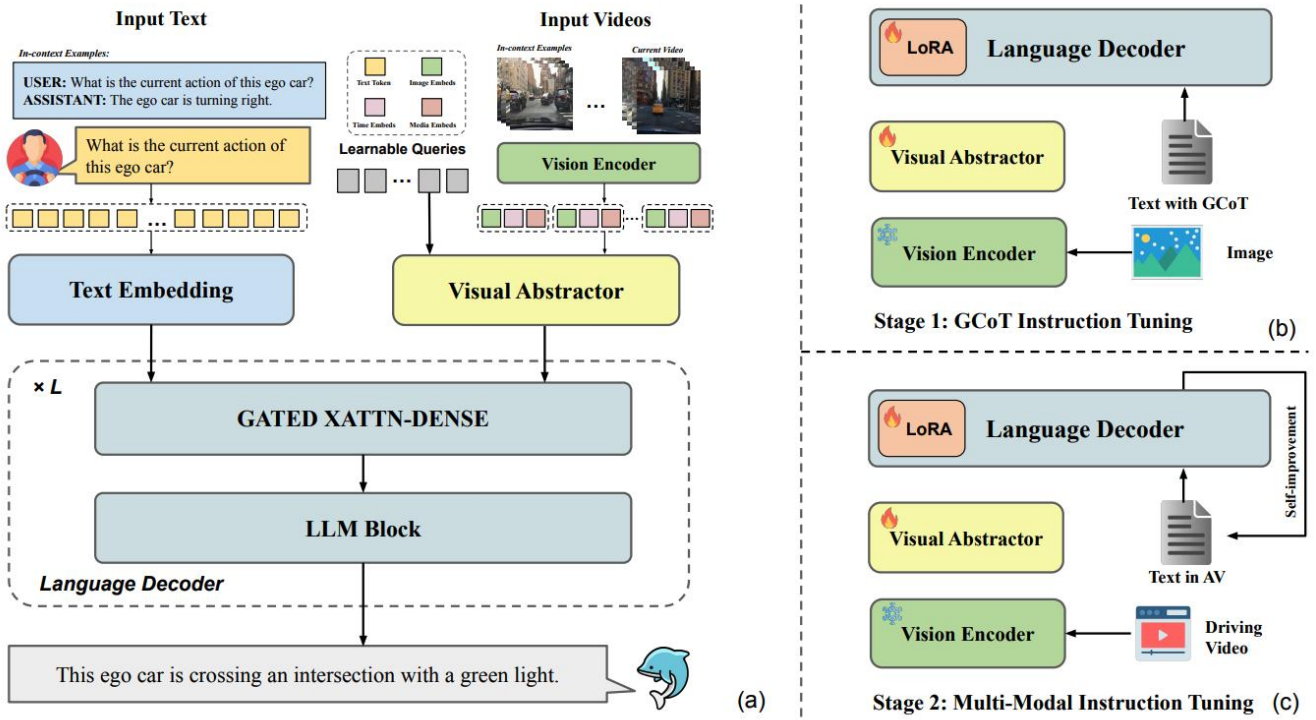
- Grounding the fine-grained understanding and reasoning capabilities of VLM in the traditional image-text pair domain into the field of autonomous driving through in-context instruction tuning on BDD-X dataset.



Ma, Y., Cao, Y., Sun, J., Pavone, M., & Xiao, C. (2023). Dolphins: Multimodal Language Model for Driving. ArXiv, abs/2312.00438.

Dolphins: Multimodal Language Model for Driving

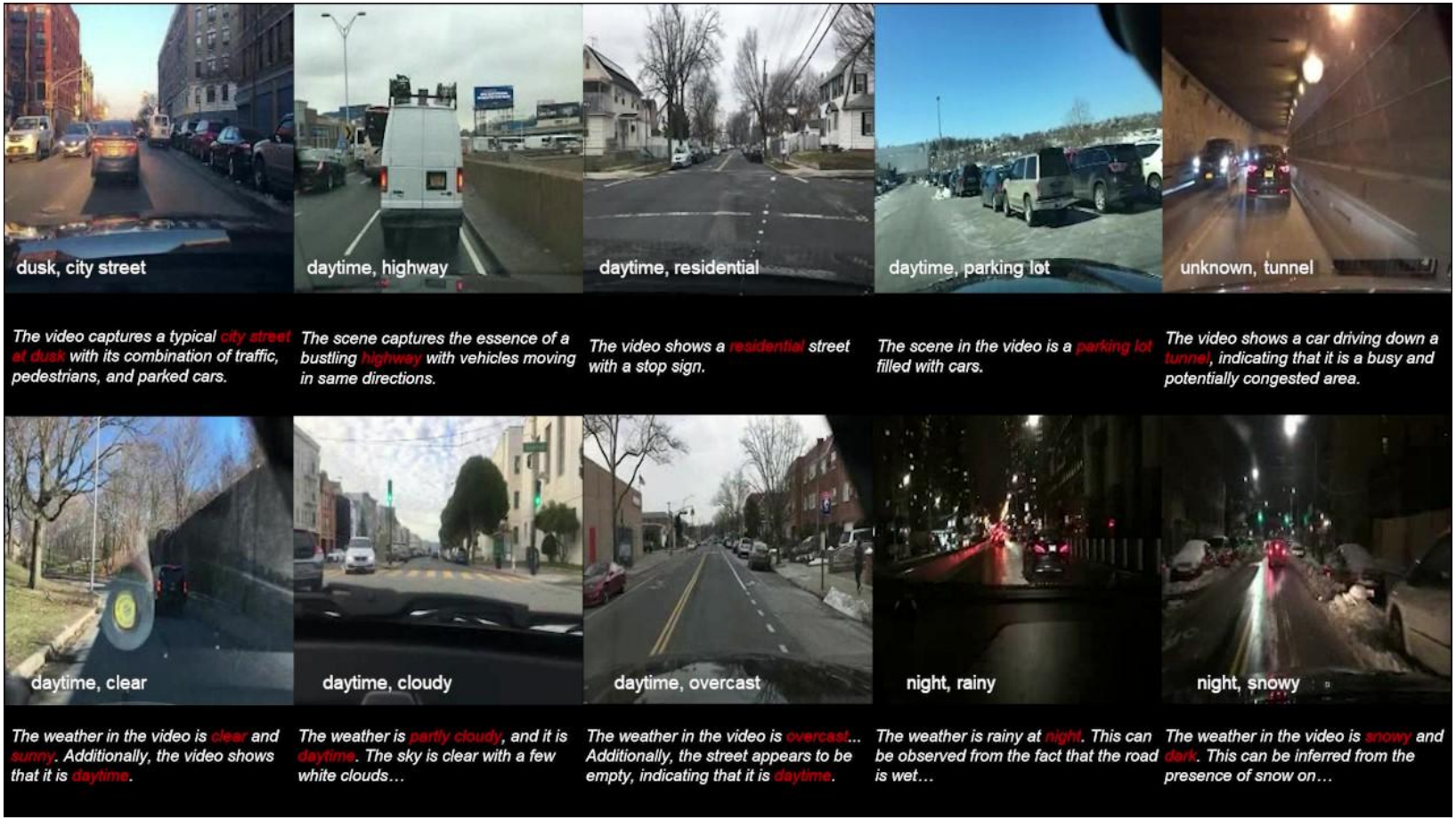
- The training paradigm of Dolphins involves GCoT instruction tuning and multi-model in-context instruction tuning with updates exclusively made to the visual abstractor and LoRA parameters. Furthermore, the model can self-refine by training on its generated pseudo labels.



Ma, Y., Cao, Y., Sun, J., Pavone, M., & Xiao, C. (2023). Dolphins: Multimodal Language Model for Driving. ArXiv, abs/2312.00438.

Dolphins: Multimodal Language Model for Driving

- Versatility in Diverse Scenarios: Designed to tackle various driving-related tasks, Dolphins shows promise in managing different urban and rural driving conditions and interpreting a range of instructions.



Ma, Y., Cao, Y., Sun, J., Pavone, M., & Xiao, C. (2023). Dolphins: Multimodal Language Model for Driving. ArXiv, abs/2312.00438.

Dolphins: Multimodal Language Model for Driving

- Dolphins can have a conversation with the driver as a driving assistant.

Ma, Y., Cao, Y., Sun, J., Pavone, M., & Xiao, C. (2023). Dolphins: Multimodal Language Model for Driving. ArXiv, abs/2312.00438.

Thanks for listening!