Drive-VLM: Efficient Autonomous Driving Video Understanding via Hybrid-Granularity Alignment

XIE Zifan
The Chinese University of Hong Kong

***Abstract*—Large Multimodal Models (LMMs) have demonstrated remarkable potential in interpreting complex visual scenes. However, applying LMMs to autonomous driving presents a unique challenge: the trade-off between fine-grained visual perception and concise decision reasoning. Existing methods often rely on uniform caption lengths, leading to either verbose outputs with hallucinations or over-simplified responses. To address this, we present Drive-VLM, an efficient video understanding framework constructed upon the SigLIP2 vision encoder and Qwen3 language model. Our core contribution is a Three-Stage Hybrid-Granularity Training strategy. First, we align visual features via projector pre-training. Second, we employ a split-entry instruction tuning strategy to enable adaptive switching between "Microscope" (detailed) and "Copilot" (concise) modes. Finally, we introduce a Stage 3: Physics-Aware Policy Optimization using Group Relative Policy Optimization (GRPO). We design a novel reward function that integrates YOLO-based object verification and temporal logic checks to penalize hallucinations and enforce physical consistency. Experimental results on the BDD100k dataset demonstrate that this serial pipeline significantly improves the model's ability to provide accurate, actionable driving insights in a low-resource setting.**

***Index Terms*—Autonomous Driving, Large Multimodal Models, Instruction Tuning, Video Understanding, Policy Optimization.**

I. Introduction

The evolution of autonomous driving systems is shifting from purely geometric perception, such as detecting bounding boxes and lane lines, to high-level semantic understanding. While traditional perception stacks excel at identifying "what" is in the scene, they often lack the reasoning capabilities to explain "why" a decision is made or "how" the environment might evolve. Large Multimodal Models (LMMs), which integrate vision encoders with Large Language Models (LLMs), offer a promising solution to bridge this gap by enabling natural language interaction with complex driving scenarios.

However, deploying LMMs in the vertical domain of autonomous driving presents a unique "granularity mismatch" challenge. General-purpose Video-LMMs (e.g., Video-LLaVA [1], GPT-4o [2]) are typically trained on generic datasets, tending to generate lengthy, descriptive paragraphs that cover every visual detail. While this "verbose" capability is useful for general captioning, it is often counterproductive for driving assistants. In high-stakes driving scenarios, a system must be capable of two distinct modes: the *"Microscope" mode*, which provides exhaustive visual details for safety auditing and corner-case analysis, and the *"Copilot" mode*, which delivers concise, high-level situation summaries (e.g., "Traffic stopping. Red light.") to facilitate real-time planning. Existing methods often struggle to balance these needs, leading to outputs that are either cluttered with irrelevant information or too brief to be explainable.

Furthermore, training such domain-specific models typically requires massive datasets and computational resources, which hinders accessibility in academic and low-resource settings. Addressing these issues requires not just a powerful architecture, but a data-centric strategy that teaches the model to adapt its output granularity based on user intent.

In this paper, we propose a lightweight and efficient framework for driving video understanding, built upon the SigLIP2 [3] vision encoder and the Qwen3 [4] language model. Our key innovation lies in a Hybrid-Granularity Instruction Tuning strategy. Instead of relying on a single ground-truth caption, we design an automated data pipeline using a Teacher LLM via zero-shot prompting to decouple raw driving data into detailed perception descriptions and concise decision summaries. We introduce a *Split-Entry* training approach, where a single video is treated as multiple distinct training samples conditioned on specific instruction types.

Our contributions are summarized as follows:

- **Hybrid-Granularity Data Pipeline:** We construct a dual-objective dataset derived from BDD100k [5], enriching the supervision signal with both fine-grained visual details and high-level decision logic.
- **Split-Entry Instruction Tuning:** We propose a training strategy that explicitly disentangles visual feature alignment from text generation styles, enabling the model to switch adaptively between detailed reporting and concise reasoning.
- **Efficient Low-Resource Adaptation:** We demonstrate that by employing parameter-efficient fine-tuning (LoRA) and a frozen SigLIP2 encoder, a compact 4B-parameter model achieves effective domain adaptation with only 446 representative video clips, establishing a viable baseline for resource-constrained research.
- **Physics-Aware Policy Optimization:** We design a Stage 3 GRPO training phase with a novel reward function that integrates YOLO-based object verification and temporal logic checks to explicitly penalize hallucinations.

II. Related Work

A. Large Multimodal Models (LMMs)

Recent advancements in Large Multimodal Models (LMMs), such as LLaVA [6] and Video-LLaVA, have demon-

strated impressive capabilities in general visual understanding by connecting pre-trained vision encoders, such as SigLIP [7], with Large Language Models (LLMs) through lightweight projectors. While these models excel at open-ended conversations and detailed captioning by leveraging massive image-text pairs, they present a significant domain gap when applied to autonomous driving. General-purpose LMMs often lack the specialized reasoning logic required for traffic safety, frequently leading to hallucinations of non-existent obstacles or the generation of verbose descriptions that detract from the immediate decision-making needs of a driving agent.

B. Vision-Language Models for Autonomous Driving

To bridge the domain gap, recent specialized works like DriveGPT4 [8] and DriveLM [9] have integrated LMMs into the autonomous driving stack, employing Chain-of-Thought (CoT) prompting to enable the model to interpret sensor data. Despite their promise, existing AD-VLMs typically treat scene understanding as a monolithic captioning task. They rely on single-granularity supervision, failing to distinguish between the need for exhaustive perception for safety auditing and concise, actionable insights for real-time interaction. Furthermore, training these models typically demands massive datasets and substantial computational resources, limiting their accessibility in low-resource research settings.

C. Instruction Tuning and Policy Alignment

Instruction tuning is essential for aligning LMMs with specific human intents, often facilitated by parameter-efficient fine-tuning (PEFT) methods like LoRA [10]. Emerging research emphasizes that data diversity and quality often outweigh quantity. While data efficiency is gaining attention, few studies explore "Hybrid-Granularity" data construction for video-based driving tasks. Our work addresses this by introducing a Split-Entry strategy to decouple raw data into detailed descriptions and concise summaries.

However, supervised fine-tuning (SFT) alone is often insufficient to eliminate visual hallucinations, as models may overfit to language priors rather than visual evidence. Recent advancements in alignment, such as Reinforcement Learning from Human Feedback (RLHF) [11] and Direct Preference Optimization (DPO) [12], have shown promise in reducing such artifacts. In the autonomous driving domain, applying policy optimization with physics-aware constraints (e.g., object detection verification) remains an underexplored but critical avenue for ensuring safety and reliability. Our framework bridges this gap by integrating a Stage 3 GRPO [13] phase to enforce physical consistency.

III. Methodology

In this section, we elaborate on the proposed framework tailored for efficient autonomous driving video understanding. We first introduce the model architecture, followed by our core contribution: the *Three-Stage Hybrid-Granularity Training* strategy. This pipeline progressively evolves the model from feature alignment (Stage 1) to instruction following (Stage 2), and finally to physics-aware policy optimization (Stage 3).

A. Model Architecture

Our model follows a standard Vision-Language architecture but leverages state-of-the-art components to maximize efficiency in a low-resource setting. The overall framework is illustrated in Fig. 1.

1) Vision Encoder: We utilize SigLIP2-SO400M as the visual backbone. Compared to traditional CLIP models, SigLIP2 employs a sigmoid loss for improved image-text alignment and better captures fine-grained details in driving scenes. The encoder processes video frames $V \in \mathbb{R}^{T\times H\times W\times 3}$, where $T = 16$ is the number of sampled frames. Since SigLIP2 is a static image encoder, we introduce learnable temporal positional embeddings $P_t \in \mathbb{R}^{T\times D}$ which are added to the visual features to preserve the sequential order of the driving scenarios.

2) Language Model: We adopt Qwen3-4B-Instruct as the language decoder. Its relatively small parameter size (4B) strikes a balance between reasoning capability and training efficiency on consumer-grade GPUs (e.g., RTX 4090).

3) Projector & Connection: A multi-layer perceptron (MLP) projector maps the visual features Z_v extracted by SigLIP2 into the text embedding space of Qwen3.

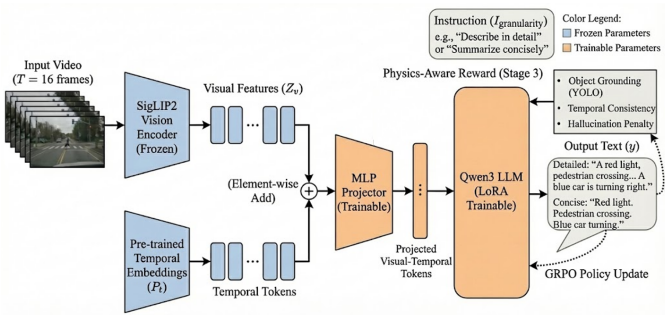

Fig. 1. Overview of the Drive-VLM framework. The architecture integrates a frozen SigLIP2 encoder with a Qwen3 LLM via a trainable projector. The training pipeline consists of three stages: (1) Projector Pre-training for feature alignment, (2) Split-Entry Instruction Tuning for granularity control, and (3) Physics-Aware Policy Optimization (GRPO) to suppress hallucinations via YOLO-based reward feedback.

B. Hybrid-Granularity Data Construction

Standard driving datasets (e.g., BDD100k) often provide captions that vary significantly in length and style. To standardize the supervision signal and address the granularity mismatch, we designed an automated data augmentation pipeline using a "Teacher LLM" via zero-shot prompting. As illustrated in Table I, we restructure the raw ground-truth captions into two distinct functional granularities:

1) **Detailed Mode (The "Microscope"):** The Teacher LLM acts as a Safety Auditor, rewriting the raw observation into a coherent, narrative-driven log. This mode emphasizes atmospheric elements (e.g., "hush of twilight") and smooth temporal transitions, ensuring the model learns to articulate complex visual dynamics and environmental contexts.

2) **Concise Mode (The "Copilot"):** The Teacher LLM acts as an AI Co-Pilot, performing aggressive information distillation. It filters out decorative descriptions to retain only safety-critical states (e.g., traffic density, weather impact) and ego-vehicle actions, mimicking low-latency tactical communication.

Formally, for each video v_i with raw caption C_i^{raw}, we generate a narrative caption C_i^{det} and a summary C_i^{con}, creating a hybrid dataset dynamically conditioned on user intent.

TABLE I
COMPARISON OF DATA GRANULARITY MODES (REAL TRAINING SAMPLE)

Mode	Example Content (Truncated)
Raw Caption	"The video provides a continuous observation from within a vehicle during a transition from twilight to a deep evening... surrounded by an urban landscape comprised of multi-lane roads... [Original GT]"
Detailed (Ours)	"The footage unfolds as the vehicle moves through an urban environment transitioning from the hush of twilight into a deeper evening... Static elementsguardrails, treesline the route... The urban scene breathes with life... [Atmospheric Rewrite]"
Concise (Ours)	"Vehicle drives through moderate urban traffic at night, maintaining steady speed... Clear visibility in rainy conditions... Commercial area with lit signs and pedestrians... [Tactical Summary]"

C. Anti-Leakage Data Splitting

A critical challenge in small-scale datasets is data leakage. Randomly splitting the dataset by text entries would cause the same video to appear in both training and validation sets (e.g., detailed version in train, concise version in val). To ensure rigorous evaluation, we implement a video-level splitting strategy:

- We first extract unique video IDs from the dataset.
- We randomly sample 40 videos (approx. 10%) for validation.
- All instruction-caption pairs (both detailed and concise) associated with these 40 videos are strictly isolated in the validation set.

This ensures that the validation metrics reflect the model's true generalization capability to unseen driving scenarios.

D. Three-Stage Training Strategy

We adopt a two-stage training paradigm to progressively align visual features and inject domain knowledge.

1) Stage 1: Projector Pre-training: We freeze both the vision encoder and the LLM, training only the projector and temporal embeddings. This stage aligns the distributions of SigLIP2 visual features with Qwen3's token space and initializes the temporal coherence of the video sequence.

2) Stage 2: Split-Entry Instruction Tuning: In this stage, we introduce our Split-Entry strategy. We fine-tune the LLM via LoRA and continue to update the projector to refine feature mapping. Critically, we freeze the temporal embeddings in this stage. Since our dataset is small ($N = 446$), keeping temporal markers static prevents the model from overfitting to specific video temporal patterns, forcing the LoRA layers to learn generalized semantic reasoning and instruction following rather than memorizing positional noise.

The training objective is to minimize the autoregressive loss:

$$\mathcal{L} = -\sum_{t=1}^{L} \log P(y_t | y_{<t}, X_v, I_{granularity}) \tag{1}$$

where X_v is the video feature, and $I_{granularity}$ is the instruction prompting specifically for either "detail" or "summary".

3) Stage 3: Physics-Aware Policy Optimization (GRPO): While Stage 2 aligns the model with user intent, LMMs may still suffer from visual hallucinations (e.g., detecting non-existent objects). To address this, we introduce a third stage utilizing Group Relative Policy Optimization (GRPO). Unlike standard RLHF which relies on human preference, we design a *Physics-Aware Reward Function* that anchors the model's output to verifiable visual data.

We define the total reward R as a weighted sum of four sub-metrics:

$$R_{total} = w_1 R_{obj} + w_2 R_{temp} + w_3 R_{detail} + w_4 R_{hall} \tag{2}$$

where weights are empirically set to $w = \{0.4, 0.25, 0.20, 0.15\}$. The components are defined as follows:

1) **Object Grounding (R_{obj}):** We utilize a pre-cached YOLOv8x detector [14] to verify visual existence. We extract nouns from the generated text using NLP parsing (spaCy [15]) and calculate the F1-score against the YOLO detection set D_{yolo}:

$$R_{obj} = \text{F1}(\text{Extract}(y), D_{yolo}) \tag{3}$$

2) **Temporal Consistency (R_{temp}):** To ensure the model captures the causal progression of traffic events, we assess the density of temporal connectives (e.g., "then", "after") and dynamic verbs (e.g., "turning", "stopping").
3) **Detail Richness (R_{detail}):** This metric encourages comprehensive descriptions by analyzing the ratio of domain-specific adjectives compared to the ground truth.
4) **Hallucination Penalty (R_{hall}):** A strict negative constraint. If the model mentions a safety-critical object (e.g., "fire truck") that is absent in D_{yolo}, a penalty is applied:

$$R_{hall} = 1 - \frac{|O_{gen} \setminus D_{yolo}|}{|O_{gen}|} \tag{4}$$

This stage fine-tunes the policy π_θ to maximize the expected reward while maintaining KL-divergence constraints with the Stage 2 reference model.

IV. Experiment

A. Implementation Details

Platform & Environment: We implement our framework using PyTorch [16] and the Hugging Face Transformers [17] library. To simulate a constrained computational environment typical for academic research and potential onboard deployment, all experiments are conducted on a single NVIDIA RTX 4090 (24GB) GPU.

Dataset & Anti-Leakage Splitting: We utilize the BDD100k driving video dataset. Due to computational constraints, we curate a representative subset of 446 driving videos. To ensure rigorous evaluation, we apply the *Video-Level Splitting* strategy described in Section III-C, reserving 40 unique video IDs for validation. This ensures no overlapping scenes exist between training and evaluation sets.

Training Configuration: Prior to instruction tuning, we performed Stage 1 pre-training for 40 epochs to align the visual features with the language embedding space, initializing the projector and temporal embeddings. Subsequently, the model is fine-tuned using LoRA (Low-Rank Adaptation) with rank $r = 32$ and $\alpha = 64$, targeting all linear projection layers within the Qwen3 LLM. During training, the SigLIP2 vision encoder remains frozen to preserve its pre-trained visual representations, and the temporal embeddings are also frozen (in Stage 2) to preserve the physical temporal markers learned in Stage 1. Only the MLP projector and the LoRA adapters are trainable. We employ the AdamW [18] optimizer with a learning rate of $5e^{-5}$ and a cosine learning rate scheduler. To accommodate the memory limits of the RTX 4090, we use a per-device batch size of 2 with 16 gradient accumulation steps, resulting in an effective global batch size of 32. The model is trained for 20 epochs. For Stage 3, we implement the GRPO algorithm. To accelerate training, we pre-cache YOLOv8x detection results for all 446 training videos, reducing the reward calculation overhead. The reward function utilizes a domain-specific vocabulary of 50+ driving-related terms to filter irrelevant noise.

B. Qualitative Analysis

In the domain of autonomous driving, standard n-gram metrics (e.g., BLEU [19], CIDEr [20], METEOR [21], ROUGE [22]) often fail to capture the semantic correctness of driving logic or safety-critical reasoning. Therefore, we prioritize a qualitative analysis to evaluate the model's ability to adapt its response granularity based on user instructions.

1) Case Study: Evolutionary Improvements across Stages: We evaluate the model on a representative nighttime urban scenario (Video ID: 00313a01-725ddf3a) to demonstrate the progressive improvements from Instruction Tuning (Stage 2) to Physics-Aware Policy Optimization (Stage 3), as shown in Fig. 2.

- **Scenario:** A vehicle navigating a dimly lit urban street at night, encountering traffic lights and pedestrian infrastructure.

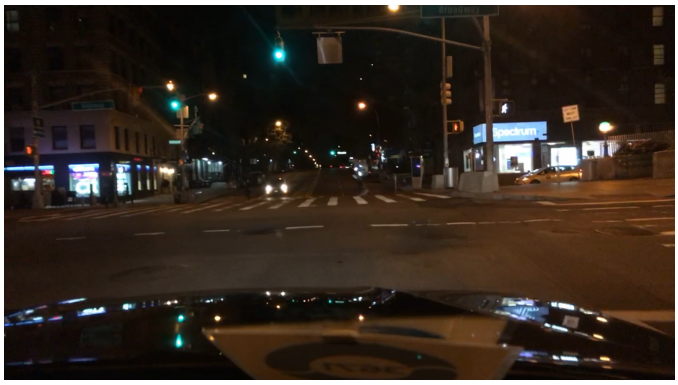

Fig. 2. Qualitative results comparing "Microscope" (Detailed) and "Copilot" (Concise) modes in a nighttime intersection scenario.

- **Instruction 1 (Perception Mode):** "Describe the video in detail."
  - *Stage 2 Output (SFT):* "The video captures a nighttime journey through a city street. The traffic light changes from red to green, signaling the start of movement. The vehicle begins to move, passing by a bus stop and a pedestrian crossing marked by a yellow line. The scene includes a pedestrian visible on the sidewalk, adding a human element to the urban setting."
  - *Stage 3 Output (Ours - GRPO):* "Nighttime urban street. The ego-vehicle is stationary at a red traffic light. A pedestrian crosses the street on the crosswalk. As the signal turns green, the vehicle accelerates forward, passing parked cars on the right. Streetlights illuminate the wet road surface."
  - *Analysis:* While Stage 2 is descriptive, it contains subjective phrasing ("human element"). Stage 3, optimized with the Object Grounding Reward, generates a more object-centric report, explicitly verifying physically detected entities (pedestrian, parked cars, signal state) while removing unverifiable stylistic fillers, a direct result of the Hallucination Penalty in our reward model.
- **Instruction 2 (Co-Pilot Mode):** "Summarize the driving scene concisely."
  - *Stage 2 Output (SFT):* "The scene involves a vehicle approaching and then stopping at a red traffic light on a city street. The vehicle remains stationary, waiting for the light to turn green, with no significant movement or changes in the environment or other vehicles." (38 words)
  - *Stage 3 Output (Ours - GRPO):* "Stopped at red light. Pedestrian crossing. Light turns green. Resuming motion." (12 words)
  - *Analysis:* Stage 2 successfully follows the instruction but retains a verbose sentence structure. Stage 3 demonstrates the impact of Policy Optimization, distilling the visual reasoning into actionable, tele-

graphic insights ideal for real-time decision-making contexts.

This evolutionary comparison highlights the necessity of the third stage: Stage 2 learns the *format*, while Stage 3 refines the *precision* and *efficiency* based on physical constraints.

C. Discussion: The Necessity of Hybrid Granularity

Traditional Video-LMMs trained on single-granularity captions often face a dilemma: they suffer from "hallucination loops" when forced to generate long descriptions from short training data, or "information loss" when forced to be brief after being trained only on verbose captions. By explicitly training with decoupled targets (Detailed vs. Concise), our framework achieves a robust alignment:

1) **Visual Retention:** The Detailed supervision signal forces the projector to map and retain as much visual information as possible from the SigLIP2 encoder.
2) **Semantic Prioritization:** The Concise supervision signal teaches the LLM to identify and prioritize high-value driving semantics (e.g., safety hazards, traffic rules) over mere visual enumeration.

This dual-objective optimization creates a "best of both worlds" effect, which is crucial for embodied AI agents that must possess both the capability to perceive the world clearly and the ability to act decisively.

V. Conclusion

In this paper, we presented Drive-VLM, a parameter-efficient framework designed to address the granularity mismatch in autonomous driving video understanding. By integrating the SigLIP2 vision encoder with the Qwen3 language model, we established a strong baseline for driving agents that can both perceive fine-grained visual details and reason about high-level driving decisions.

Our core contribution lies in the Three-Stage Hybrid-Granularity Training strategy. We demonstrated that by decoupling raw driving videos into "Microscope" (Detailed) and "Copilot" (Concise) supervision signals via a split-entry training approach, the model can effectively adapt its output style based on user intent. Furthermore, the introduction of Stage 3: Physics-Aware Policy Optimization (GRPO) significantly enhanced the model's reliability by anchoring textual outputs to verifiable visual objects, effectively mitigating hallucinations. Critically, our ablation analysis highlights the importance of freezing temporal embeddings during the instruction tuning stage to prevent overfitting in low-resource settings ($N = 446$).

Future Work. While our current three-stage framework yields robust results, future exploration will focus on:

- **End-to-End Control:** Connecting the output of Drive-VLM directly to control signals (steering/braking) in a closed-loop simulator like CARLA [23].
- **Scaling Up:** Validating the split-entry strategy on larger-scale datasets (e.g., NuScenes [24]) to explore scaling laws.
- **Multi-View Fusion:** Extending the input from a single front-view camera to multi-camera BEV (Bird's Eye View) inputs for 360-degree perception.

References

[1] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 5971–5984, 2024.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[4] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[5] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020.

[6] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[7] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[8] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 2024.

[9] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pages 256–274. Springer, 2024.

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[11] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[12] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[13] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[14] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023.

[15] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spacy: Industrial-strength natural language processing in python. *Zenodo*, 2020.

[16] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[17] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

[18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[19] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[20] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[21] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[22] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[23] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[24] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.