Label-Efficient Fine-Tuning of VLMs for Interpretable Autonomous Driving via RLOO Algorithm

Kai-Yuan Jeng [1] Pei-Hsun Wu [2] Chen-Fang Hu [2]

[1]National Tsing Hua University [2]National Yang Ming Chiao Tung University

Abstract

We develop an efficient approach for **fine-tuning Vision-Language Models (VLMs) in End-to-End Autonomous Driving (E2E-AD)**. This work addresses two distinct challenges: heavy labeled data dependency in E2E-AD applications and computational overhead of critic networks in large model RL training. Our solution adapts **REINFORCE Leave-One-Out (RLOO)** from RLHF domains to vision-language tasks. This **critic-free** algorithm enables **label-efficient** VLM fine-tuning using only semantic alignment rewards, eliminating both extensive human annotations (E2E-AD challenge) and expensive critic training (RLHF challenge).

Problem Formulation

Problem Definition: We formulate VLM fine-tuning as a **video understanding** task (extension of the image captioning). The model processes traffic scene image frames and generates natural language text containing both scene descriptions and appropriate AD control actions.

Bandit Setting: We then model this as a bandit problem where

- **State:** Input image frames $\mathbf{I}$
- **Action:** Complete generated text sequence $\mathbf{T}$
- **Reward:** Semantic similarity $\mathcal{R}(\mathbf{I}, \mathbf{T})$, measured by fine-tuned CLIP

RLOO Optimization: Instead of training expensive critic networks, RLOO uses **multiple Monte-Carlo samples as baselines for unbiased policy gradient estimation**, enabling efficient critic-free training. The policy gradient is calculated as:

$$\nabla_\theta \mathcal{J}_{\text{RLOO}}(\theta) = \mathbb{E}_{\mathbf{I}\sim\mathcal{D}}\left[\sum_{i=1}^{k}\left(\mathcal{R}(\mathbf{I},\mathbf{T}^i) - \frac{1}{k-1}\sum_{j\neq i}\mathcal{R}(\mathbf{I},\mathbf{T}^j)\right)\nabla_\theta \log \mathbf{VLM}_\theta(\mathbf{T}^i \mid \mathbf{I})\right].$$

Methodology

Two-Stage Training Framework:

1. **Text Sequence Generation via Reinforcement Learning:** CLIP rewards + RLOO optimization for scene description generation.
2. **Action Selection and Formatting Alignment via Supervised Learning:** Format alignment using instructed prompts and BLEU score.

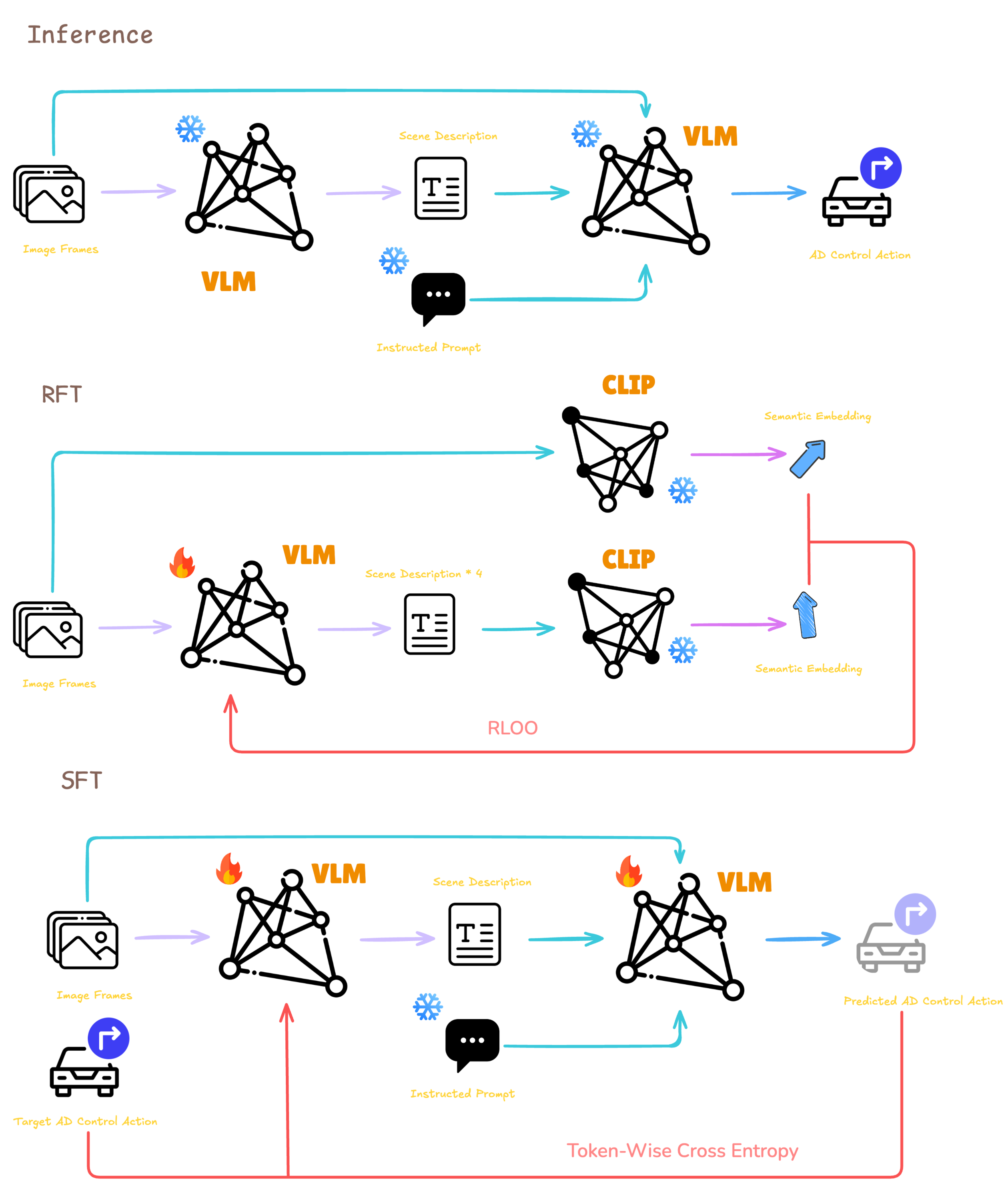


Figure 1. Complete training and inference pipeline.

Experiment

Setup:

- **BDD-OIA** dataset with traffic scenes, actions, and reasoning.
- Fine-tuned **CLIP** as reward model.
- **Token-wise supervised** fine-tuning as baseline vs. our RLOO framework on **SmolVLM-256M**.

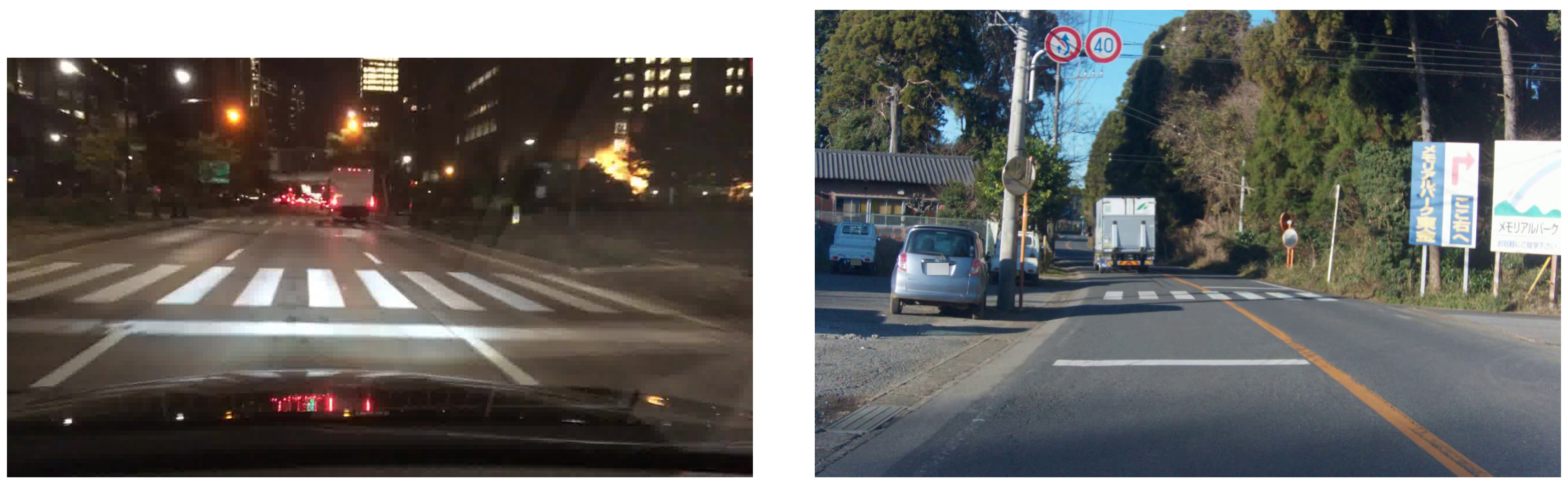
Key Findings:

- Domain gap: Pretrained VLM with 0% performance
- Catastrophic forgetting: CLIP score 91% → 42%
- Model capacity: Cannot balance pretrained knowledge + reward optimization
- Action formatting issue: 84% action F1, but all predictions are `STOP`
- Positive insight: **RLOO enhances scene description capability**

Root Cause: High learning rate (2e-5) + insufficient model capacity (256M) → requires careful hyperparameter tuning.

Table 1. Experimental results.

Method	Action F1	Reason F1	CLIP Score
Ground Truth	100.00%	100.00%	91.32%
Pretrained VLM	0.00%	0.00%	0.00%
Baseline (SFT for 10 Epochs)	80.15%	66.75%	*N/A*
RFT (RLOO for 3 Epochs with $k=4$)	*N/A*	*N/A*	41.88%
RFT+SFT (Each with 3 Epochs)	84.12%	0.00%	*N/A*



(a) BDD-OIA (in-distribution) (b) OOD scenario

Figure 2. Sample model outputs comparison.

Left: (1) **Pretrained:** "Crosswalk." (2) **SFT Baseline:** "Action: stop. Reason: Traffic light is not green." (3) **RFT+SFT:** "Action: stop. Reason: The intersection of an asian city road is fully visible in this photo."

Right: (1) **Pretrained:** "There is a road sign that says 40." (2) **SFT Baseline:** "Action: stop. Reason: Traffic sign." (3) **RFT+SFT:** "Action: stop. Reason: A two way street sign says that there are 40 down."

Conclusion

Contribution: Successfully adapted RLOO from RLHF to vision-language tasks, providing a **critic-free, label-efficient** framework for VLM fine-tuning in autonomous driving.

Future Work: Evaluate on larger VLMs and implement adaptive learning strategies for hyperparameter optimization.