2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

VL-DPO: Vision-Language-Guided Finetuning for Preference-Aligned Autonomous Driving

Zhefan Xu[1], Ghassen Jerfel[2], Marina Haliem[2], Qi Zhao[2], Jeonhyung Kang[2], and Khaled S. Refaat[2]

***Abstract*— The rapid growth of autonomous driving datasets has enabled the scaling of powerful motion forecasting models. While large-scale pretraining provides strong performance, the standard imitation objective may not fully capture the complex nuances of human driving preferences. Meanwhile, recent advances in vision-language models (VLMs) have demonstrated impressive reasoning and commonsense understanding. Building on these capabilities, this paper presents VL-DPO, a vision-language-guided framework that aligns ego-vehicle motion forecasting models with human preferences. Our approach leverages a VLM as a zero-shot reasoner to automatically generate preference pairs from a pretrained model's rollouts, which are then used to finetune the model via Direct Preference Optimization (DPO). We finetune our models on the Waymo Open End-to-End Driving Dataset (WOD-E2E) and evaluate performance against held-out human preference annotations using rater feedback score (RFS) and average displacement error (ADE). Our experiments confirm that the VLM's trajectory selection is a high-quality proxy for human preference. Our final model, VL-DPO, yields an 11.94% increase in RFS and a 10.01% reduction in ADE over the pretrained model.**

I. Introduction

Deep neural networks, empowered by the proliferation of large-scale datasets, have demonstrated strong performance in motion forecasting for autonomous driving [1][2][3][4][5]. Despite their effectiveness, the standard imitation learning objectives prioritize local geometric accuracy, often via next-token prediction accuracy. This fails to capture the holistic preferences that characterize human driving behavior, creating an alignment gap.

The recent emergence of Vision-Language Models (VLMs), with their strong commonsense reasoning and contextual understanding, offers a promising path to bridge this gap. We hypothesize that a VLM's vast world knowledge can serve as a high-quality proxy for nuanced human preferences.

However, the prevailing trend for leveraging this capability has been to adopt VLMs as monolithic end-to-end backbones for ego-vehicle motion prediction [6][7][8][9][10]. While these approaches benefit from the large-scale pretrained knowledge embedded in VLMs and have shown promising results, they still face several key challenges. First, these models often require large-scale, high-quality, and diverse datasets with aligned language and action (as exemplified by the data curation process in [11]), which can be expensive and time-consuming to collect and annotate. Second, without careful data curation, finetuning large pretrained models on domain-specific driving datasets can lead to catastrophic forgetting [12], eroding the general world knowledge that initially underpins their effectiveness [13]. Third, the black-box nature of these end-to-end models makes it difficult to interpret individual decisions, posing a critical barrier for safety-sensitive applications such as autonomous driving. Fourth, the computational demands of large VLMs often lead to high inference latency, making them impractical for real-time onboard deployment in autonomous vehicles.

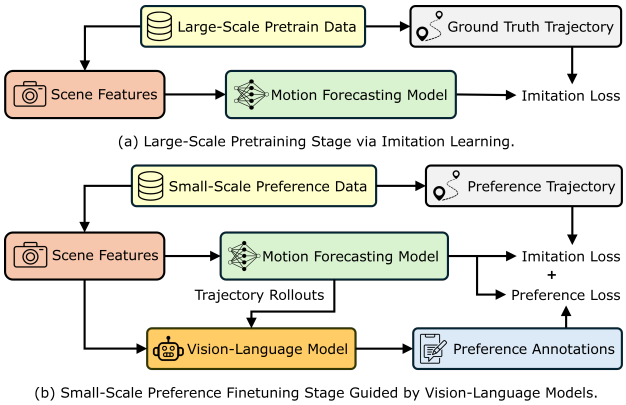

Fig. 1. Illustration of the proposed training stages of the motion forecasting model. (a) Multi-agent pretraining via imitation learning on large-scale data. (b) Single-agent (ego-vehicle) preference finetuning using VLM-generated pairs, which replace human annotation.

To address the limitations of finetuning VLMs as Vision-Language-Action models, we propose a more modular paradigm that decouples high-level reasoning from low-level motion forecasting by leveraging the VLM not as an integrated component, but as a frozen zero-shot reasoner and critic. This design preserves the VLM's world knowledge and avoids catastrophic forgetting while enabling a more interpretable and efficient finetuning process. In this role, the VLM generates supervisory signals, including fine-grained trajectory preference pairs—our main contribution—and coarser discrete High-Level Actions (HLAs), which we also investigate as a simpler supervision method.

This culminates in VL-DPO: a vision-language-guided framework that aligns a pretrained motion forecasting model with human preferences, as illustrated in Fig. 1. The framework builds on Direct Preference Optimization (DPO) [14], where the VLM evaluates candidate trajectories from model rollouts to automatically generate preference pairs for finetuning. Our empirical study on the Waymo Open End-to-End Driving Dataset (WOD-E2E) shows that VL-DPO consistently outperforms the coarser HLA-based supervision and

[1] zhefanx@andrew.cmu.edu
[2] {ghassen,mhaliem,zhaqi,jeonhyung,krefaat}@waymo.com

achieves state-of-the-art alignment with human preferences. The main contributions of this work are:

- **VLM as a Zero-Shot Preference Annotator:** We propose the novel use of a VLM as a zero-shot reasoner to generate preference pairs for supervising a separate motion forecasting model. We validate this approach by demonstrating that the VLM's trajectory selections are more human-aligned than the model's own most-likely predictions, establishing it as an effective and scalable proxy for human feedback.
- **VL-DPO for Motion Forecasting:** We present the first framework that leverages vision-language-guided preferences within direct preference optimization to significantly improve the alignment of ego-vehicle motion forecasting models with human driving preferences.
- **Comprehensive Empirical Study:** On WOD-E2E, our proposed VL-DPO framework demonstrates significant improvements over both a pretrained model and a strong imitation learning baseline (finetuned on the single highest-rated human preference trajectory), while consistently outperforming alternative VLM-guidance strategies like HLA-based supervision.

The rest of this paper is structured as follows: we review related work in Section II, detail our methodology in Section III, present our experimental results in Section IV, and finally conclude with a discussion of our findings.

II. Related Work

Motion forecasting, crucial for autonomous driving, has advanced from CNNs on rasterized scenes [15][16][17], and GNNs on scene graphs [18][19], to powerful Transformer-based architectures [3][20][21]. Paradigms like MotionLM [2], reformulate it as a language modeling for motion tokens, effectively capturing complex interactions and multimodal motion distributions. While achieving high accuracy, its standard next-token imitation training objective is primarily optimized for replicating expert-demonstrated trajectories. This focus on local fidelity means it remains misaligned with the long-term consistency and holistic nature of human driving behavior, as it does not explicitly account for nuanced preferences or explore alternative, equally valid or even preferred paths that may deviate from the exact expert trace. This motivates our work on human preference alignment in motion forecasting.

VLM-based Approaches: The dominant approach in leveraging VLMs for autonomous driving involves finetuning them as monolithic end-to-end models [22][23][24]; for instance, DriveLM [6] trains a VLM for trajectory prediction via a VQA framework, WiseAD [8] enriches this with explicit driving knowledge, and EMMA [9] finetunes a VLM to output both reasoning traces and trajectory predictions. Others, like VDT-Auto [25], use a VLM to output embeddings for a policy module while DiMa [26] relies on the joint-training of a VLM and a vision-based policy model. All of these approaches share the common trait of directly modifying the VLM's weights for the driving task which can erode its world knowledge and reasoning capabilities.

In contrast, our work adopts the less-explored modular paradigm of using VLMs as frozen, zero-shot reasoners. Prior work in this area focused on auxiliary tasks like safety-aware decision verification for perception and planning modules [27] or hard-example mining [28]. The most closely related work to ours is VLM-AD [29] which leverages VLM-generated reasoning text to construct an auxiliary classification loss. Another line of work explores aligning BEV features with LM-derived expectation embeddings (ALP) and extending this alignment to ego-vehicle query features for planning (SLP) via contrastive learning [30]. While this paradigm enriches perception and planning with semantically meaningful representations, it remains an auxiliary training signal for feature refinement, rather than a direct mechanism for aligning autonomous driving decisions with human preferences. OmniDrive [31] generates a counterfactual Q&A dataset by simulating alternative trajectories and prompting GPT-4 for safety reasoning, offering denser supervision for perception and planning but focusing on dataset quality rather than preference alignment. Our work explores a novel and complementary direction focusing on simpler and more interpretable forms of supervision: we investigate direct model conditioning on discrete HLAs and, most critically, the generation of explicit preference pairs for human preference alignment. This approach is simpler, as it avoids hand-crafting intricate loss functions, and more interpretable, as the quality of the VLM's discrete outputs—a chosen action or a preference—can be directly evaluated.

Preference Alignment: The standard paradigm for aligning models with human preferences is Reinforcement Learning from Human Feedback (RLHF) [32][33]. While powerful, this multi-stage process of reward modeling and reinforcement learning is notoriously complex and can suffer from training instability. To avoid these issues, our work employs direct preference optimization, a more direct and stable offline method that optimizes a policy via a simple classification loss on preference pairs [14]. DPO has recently been applied to embodied agent tasks, such as motion forecasting with implicit preferences from heuristics [34] and robotic manipulation [35]. Our work introduces a fundamentally different approach for generating the supervisory signal. We leverage a VLM in a zero-shot capacity to generate explicit, semantically-grounded preference pairs based on its commonsense understanding of the scene.

III. Methodology

A. Generative Motion Forecasting Model

The motion forecasting model in our framework is pretrained on a joint autoregressive prediction task for 8 agents: the ego vehicle and its 7 nearest neighbors. Following the MotionLM paradigm [2], we adopt an encoder–decoder transformer architecture and formulate motion forecasting as a sequence generation task conditioned on a rich scene context, $\mathbf{S}$, comprising the road graph, traffic light states, and historical agent trajectories (see Fig. 2). The model generates fine-grained sequences of discrete action tokens for each agent from a finite vocabulary. These tokens represent

quantized changes in the agent's position, which can be deterministically converted back into a continuous trajectory. By sampling and aggregating multiple trajectories, the model produces a multi-modal set of 12 predictions for each agent.

The pretraining objective is imitation learning, where parameters are optimized to maximize the log-probability of expert-demonstrated trajectories. This allows the model to learn a rich implicit distribution over possible futures $P(a_1^1, a_1^2, ..., a_1^8, \ldots, a_T^1, a_T^2, ..., a_T^8 | \mathbf{S})$ capturing complex multi-agent interactions and yielding a foundational representation suitable for diverse downstream tasks. Importantly, this generative formulation enables the model to score candidate trajectories by computing their log-probabilities.

During preference finetuning, we restrict the task to single-agent (ego-vehicle) trajectory prediction since the autonomous vehicle is the only agent whose perception data can be directly processed by the VLM.

B. VLM as a Zero-Shot Driving Reasoner

1) Multimodal Scene Representation

To enable the VLM to perform robust context-aware reasoning, we construct a rich multimodal scene representation that provides a holistic view of the driving environment by combining three modalities as illustrated in Fig. 3.

- **Egocentric and Temporal Vision.** A panoramic image is constructed by stitching the front, front-left, and front-right camera feeds into a wide egocentric view. To capture scene dynamics, we represent this view as a sequence comprising the current frame and four historical frames sampled at 1 Hz.
- **Spatial Reasoning and Visual Grounding.** We generate a top-down, bird's-eye-view (BEV) visualization that renders the local road graph, the positions of other agents, and, most importantly, candidate ego-trajectory predictions from motion forecasting model rollouts. By presenting the candidate trajectories visually within a geometrically precise ego-centric frame, we transform the abstract preference selection task into a more direct visual comparison problem, enabling the VLM to reason about spatial conflicts and path suitability. To present all options in a single input, we generate a BEV image for each of the 12 candidate trajectories and combine them into a composite visualization, then the VLM selects the preferred trajectory by image index. Sec. III-C provides details on using this for preference finetuning.
- **Textual Grounding.** To ground the VLM's reasoning in both long-term goals and immediate dynamics, we add structured text describing key scene features, including the route planner's navigational command and the ego vehicle's current speed.

2) Chain-of-Thought Prompting for Driving Supervision

To generate high-quality supervisory signals for preference finetuning, we apply Chain-of-Thought prompting, which elicits a structured and interpretable reasoning trace from the VLM. As illustrated in Fig. 3, this multi-step process guides the model through a hierarchical reasoning cascade—progressing from scene understanding to the final high-level action output and preference trajectory selection.

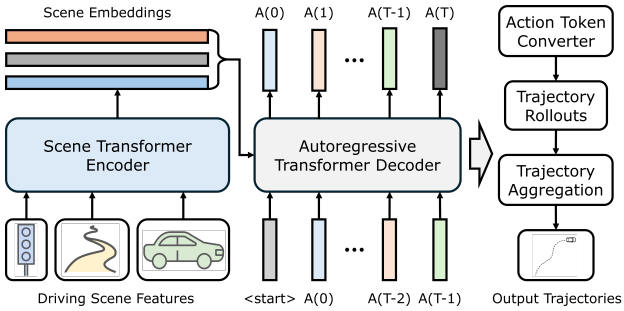

Fig. 2. The architecture of MotionLM [2]. It adopts an encoder–decoder transformer that takes scene features as input and autoregressively generates discrete action tokens from a predefined vocabulary, followed by postprocessing to convert these tokens into trajectories across multiple modalities.

1) **Scene Understanding:** The designed prompt begins by establishing general context, with the VLM describing the overall driving environment, including road type, prevailing weather, and time of day.
2) **Critical Object Perception:** The next step identifies critical objects in the camera image together with their relative positions to the ego vehicle.
3) **Object-Scene Relation Identification:** For each critical object, the reasoning process determines interactions with other objects and environmental conditions.
4) **Dynamic Object Behavior Prediction:** For each critical moving object detected in previous steps, high-level future actions are predicted from a predefined set.
5) **Potential Risk Analysis:** The VLM analyzes the scene to identify potential risks or hazards to the ego vehicle and explains why each poses a threat.
6) **Driving Decision Prediction:** Based on the preceding analysis, a high-level action is recommended for the ego vehicle from a predefined set (see Section III-C.1).
7) **Preference Trajectory Selection:** Finally, using the full reasoning trace together with the stitched top-down visualization of candidate trajectories, the framework selects the most preferred trajectory for the scenario.

C. VLM-Guided Finetuning Framework

To leverage the VLM's supervisory signals, we propose and evaluate two distinct finetuning methodologies: direct supervision on discrete High-Level Actions in Sec. III-C.1, and a more holistic alignment on trajectory-level preferences via direct preference optimization in Sec. III-C.2.

1) Supervision with High-Level Actions (HLAs)

To transform the VLM's free-form reasoning into a structured format, we experiment with three categories of high-level action vocabularies defined as follows:

- **Maneuver Action:** This action set provides a holistic semantic label for complete driving behaviors, combining both directional and speed intent. It includes: [Move to stop, Stop to move, Left turn, Right turn, Backup, Left lane change, Right lane change, Remain stopped, U-turn, Pullover, Lane following].

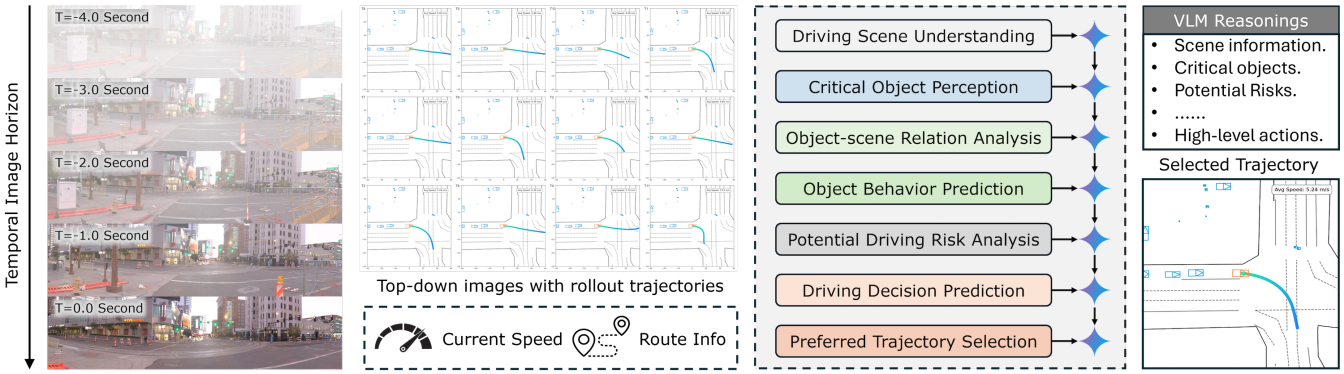

Fig. 3. Illustration of the VLM's Chain-of-Thought (CoT) reasoning process. The VLM takes as input the sequence of image history, top-down view images with sampled trajectories, along with speed and route information. The reasoning flows from perception to driving decision prediction and ultimately to preference-based trajectory selection. The model outputs both the high-level action and the selected trajectory, which serve as finetuning signals.

- **Direction Action:** This action set specifies the ego vehicle's directional intent while omitting explicit speed information. It includes: [Stop, Go straight, Turn left, Turn right, Left lane change, Right lane change].
- **Speed Action:** This action set captures the dynamic aspect of motion by describing the intended change in the ego vehicle's speed profile. It includes: [Maintain speed, Accelerate, Decelerate, Hard brake].

In our framework, the **maneuver** category serves as the default action definition. We also explore an alternative formulation where **direction** and **speed** actions are combined to represent driving intent. These different definitions are primarily evaluated to investigate whether the choice of action representation influences finetuning performance.

With the HLAs from the VLM, we propose two integration strategies: (i) incorporating them in an auxiliary loss, and (ii) conditioning the motion forecasting model on them as input.

HLAs as an Auxiliary Loss. The first approach treats HLA prediction as an auxiliary task within a multi-task learning framework. The underlying hypothesis is that enforcing the model's latent representations to be predictive of the VLM's high-level semantic intent encourages a more structured and context-aware representation space, thereby improving the quality of trajectory prediction. To implement this, we attach a lightweight multi-layer perceptron (MLP) head to the decoder's final hidden state embeddings. This classification head, followed by a softmax layer, is trained to predict the VLM-provided HLA token using cross-entropy loss. The resulting auxiliary loss is combined with the primary imitation learning loss during finetuning.

HLAs as Conditional Input. The second approach leverages the VLM-provided HLA as a conditional input, framing the problem as goal-conditioned trajectory generation. The hypothesis is that explicitly conditioning on the high-level goal reduces the size of the output space, simplifying prediction and improving accuracy. Concretely, the VLM-generated HLA is encoded as a one-hot vector, projected into the model's embedding dimension, and concatenated with other scene context features before entering the encoder. This design ensures that the entire model is conditioned on the semantic maneuver it is expected to execute.

Algorithm 1: VL-DPO Loss Computation

$\theta_{ref} \leftarrow$ pretrained motion forecasting model;
$\theta_{target} \leftarrow$ target motion forecasting model;
$\mathcal{S}^{\mathcal{T}}_{\text{rollout}} \leftarrow$ **sampleRollouts**$(\theta_{ref}, \text{inputs}, N_{\text{samples}})$;
$\mathcal{T}_{\text{select}} \leftarrow$ **getVLMSelectedTraj**$(\mathcal{S}^{\mathcal{T}}_{\text{rollout}}, \text{prompts})$;
$\mathcal{S}^{\mathcal{T}}_{\text{unselect}} \leftarrow \mathcal{S}^{\mathcal{T}}_{\text{rollout}} \setminus \mathcal{T}_{\text{select}}$;
$\pi^{w}_{ref} \leftarrow$ **computeProb**$(\theta_{ref}, \mathcal{T}_{\text{select}})$;
$\pi^{w} \leftarrow$ **computeProb**$(\theta_{target}, \mathcal{T}_{\text{select}})$;
$\mathcal{L}_{\text{VL-DPO}} \leftarrow 0$;
for $\mathcal{T}^{i}_{\text{unselect}}$ **in** $\mathcal{S}^{\mathcal{T}}_{\text{unselect}}$ **do**
  $\pi^{l}_{ref} \leftarrow$ **computeProb**$(\theta_{ref}, \mathcal{T}^{i}_{\text{unselect}})$;
  $\pi^{l} \leftarrow$ **computeProb**$(\theta_{target}, \mathcal{T}^{i}_{\text{unselect}})$;
  $\mathcal{L}_{\text{DPO}} \leftarrow$ **DPOLoss**$(\pi^{w}_{ref}, \pi^{w}, \pi^{l}_{ref}, \pi^{l})$;
  $\mathcal{L}_{\text{VL-DPO}} \leftarrow \mathcal{L}_{\text{VL-DPO}} + \mathcal{L}_{\text{DPO}}$;
$N_{\text{pairs}} \leftarrow N_{\text{samples}} - 1$;
$\mathcal{L}_{\text{VL-DPO}} \leftarrow \mathcal{L}_{\text{VL-DPO}} \;/\; N_{\text{pairs}}$;
return $\mathcal{L}_{\text{VL-DPO}}$;

2) Supervision with VLM Trajectory Selections

The standard next-token prediction objective—central to paradigms such as MotionLM [2]—is effective at capturing locally coherent behaviors but does not guarantee long-term consistency. This results in a misalignment, as the model may optimize for trajectory prediction accuracy while neglecting aspects humans value, such as yielding to pedestrians, maintaining safe gaps, or avoiding overly aggressive maneuvers. Direct preference optimization [14] addresses this limitation by training on preference pairs, each consisting of a preferred and an unpreferred behavior. The objective increases the likelihood of the preferred example (in our case, the VLM-selected trajectory) while decreasing the likelihood of the unpreferred ones. Building on this idea, we propose a vision–language-guided DPO (VL-DPO) loss that leverages VLM-generated preference pairs to align motion forecasting models with human preferences as detailed in Alg. 1.

To compute the VL-DPO loss, we use a pretrained reference motion forecasting model, θ_{ref}, which remains frozen during finetuning, along with a target model, θ_{target},

initialized from the same pretrained weights (Lines 1-2). Specifically, we first sample $N_{\text{samples}} = 12$ rollout trajectories from the reference model, render the corresponding top-down BEV representations with surrounding agents and the roadgraph, and then apply the CoT method described in Sec. III-B.2 to select the most preferred trajectory (Lines Alg. 3–4). This yields a single selected trajectory, $\mathcal{T}_{\text{select}}$, and a set of 11 unselected trajectories, $\mathcal{S}^{\mathcal{T}}_{\text{unselect}}$ (Line 5). From this, we construct 11 preference pairs, each comparing the VLM-selected trajectory against one of the unselected trajectories:

$$\{(\mathcal{T}_{\text{select}}, \mathcal{T}^{i}_{\text{unselect}}) \mid \mathcal{T}_{\text{unselect}} = \mathcal{T}_{\text{rollout}} \setminus \mathcal{T}_{\text{select}}\}. \quad (1)$$

For each preference pair, the DPO loss requires computing the probabilities, π, of both the selected and unselected trajectories under the reference and target models (Lines 6–8 and 10–11 in Alg. 1). DPO loss for a single pair is:

$$\mathcal{L}_{\text{DPO}} = -\log\sigma\left(\beta\log\frac{\pi^{w}}{\pi^{w}_{ref}} - \beta\log\frac{\pi^{l}}{\pi^{l}_{ref}}\right), \quad (2)$$

where σ denotes the sigmoid function and β is a scaling parameter that regulates how much the target model is permitted to deviate from the reference model. The final VL-DPO loss for each example is obtained by averaging the DPO losses over all 11 preference pairs (Lines 13–15).

IV. Result and Discussion

A. Dataset and Model Configurations

The motion forecasting model is pretrained on a large-scale internal dataset. Our model adopts the encoder–decoder architecture of MotionLM [2], pretrained for multi-agent prediction and finetuned for single-agent ego forecasting. For finetuning and evaluation, we use the Waymo Open End-to-End Driving Dataset (WOD-E2E)[1], which contains 2,037 training and 479 validation examples. In each example, our model is given a 4-second observation window and tasked with predicting the ego-vehicle trajectory over the next 5 seconds. To ensure compatibility with the pretrained backbone, we augment WOD-E2E with features extracted using proprietary systems, including road graph information, traffic light states, and historical agent trajectories. For the VLM, we use Gemini 2.5 Pro [36] in its zero-shot setting. Performance is assessed on the post-aggregation (12) trajectories using the human-annotated preference labels based on:

- **Rater Feedback Score (RFS):** The primary metric for human preference alignment. Each example in WOD-E2E provides three human-annotated trajectories ranked from most to least preferred, with raters assigning scores from 0 (worst) to 10 (best). The model's predictions are then evaluated against these annotated preferences, where higher scores indicate better alignment. Our motion forecasting model outputs 12 trajectories, so we report three RFS metrics: RFS (central-mode trajectory[2]), avgRFS (average over all trajectories), and mlRFS (most-likely trajectory).

[1]https://waymo.com/open/challenges/2025/e2e-driving/

[2]centrally located trajectory among all predicted trajectories.

Method	RFS ↑	ADE (5S) [m] ↓
Most Likely Selection	7.2279	3.1948
VLM Selection	7.2970	2.9460
VLM Selection Gain	+0.96%	-7.79%

TABLE I

Comparison of VLM-selected trajectory with the most likely trajectory from baseline model rollouts.

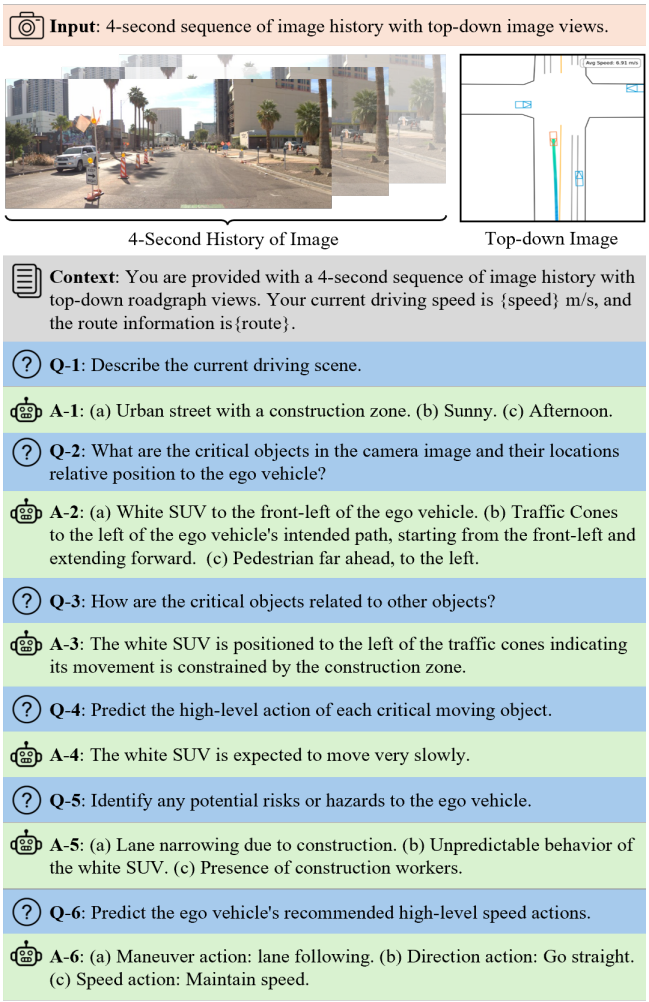

Fig. 4. Example of VLM Chain-of-Thought reasoning. For clarity, the full VLM response is simplified, and the trajectory selection step is omitted.

- **Average Displacement Error (ADE):** A geometric accuracy metric, computed as the mean L2 distance between the predicted trajectory and the target ego-vehicle trajectory at each timestep over a predefined prediction horizon. Lower values indicate higher performance.

B. VLM Reasoning Evaluation

To evaluate the VLM-generated supervisory signals, we perform both qualitative and quantitative analyses.

Qualitative Reasoning Analysis. Fig. 4 provides a representative example of the VLM's Chain-of-Thought reasoning with simplified prompt questions, intermediate reasoning steps, and final high-level action outputs. For brevity, the

Method	RFS ↑	avgRFS ↑	mlRFS ↑	ADE (5s) ↓
MotionLM [2]	7.292	7.083	7.227	3.195
Imitation Learning	7.844	7.861	7.855	3.075
IL+VL-HLA (Loss)	7.841	7.799	7.779	2.825
IL+VL-HLA (Input)	8.060	8.032	8.096	**2.715**
IL+VL-DPO	**8.163**	**8.069**	**8.138**	2.875

TABLE II

PERFORMANCE COMPARISON OF MOTIONLM [2] BASELINE AND FINETUNED VARIANTS ON RFS, AVGRFS, MLRFS AND 5-S ADE.

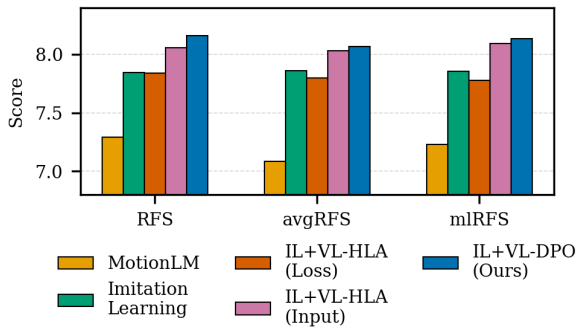

Fig. 5. Comparison of RFS, avgRFS, mlRFS across the finetuning methods.

trajectory selection step (as in Fig. 3) is omitted. In the complex construction-zone scenario of Fig. 4, the VLM correctly identifies the street condition and analyzes critical objects in the scene. It successfully reasons about object relations and forecasts the motion of moving agents. Importantly, it highlights potential risks such as lane narrowing due to construction, the possibility of unexpected motion from an oncoming SUV, and the presence of construction workers. Based on this reasoning, the VLM predicts a reasonable maneuver: lane following while going straight and maintaining speed.

Quantitative Selection Performance. We compare the quality of the VLM's final trajectory choice against the motion prediction model's own probabilistic ranking (i.e., its most likely rollout). As shown in Table I, the trajectory selected by the VLM is demonstrably better aligned with human preferences, yielding an improved RFS (+0.96%) and a significant reduction in ADE (–7.79%). We note that the zero-shot VLM selection in Table I yields a much larger improvement in ADE than in RFS. This is a direct consequence of how the two metrics are defined. ADE is computed against the single most-preferred human trajectory, so choosing smoother and safer rollouts translates directly into lower error. RFS, however, is assigned by matching to one of three annotated trajectories with their scores, using trust regions. Once the prediction falls inside the trust region of a given rater trajectory, further geometric improvements (which strongly reduce ADE) do not necessarily change the RFS score, making the gain appear small. Importantly, our finetuning results (Table II) demonstrate that this limitation is overcome when VLM selections are turned into preference pairs for DPO training: the pairwise supervision signal targets ranking consistency across all annotated trajectories, leading to much stronger RFS improvements.

C. Finetuning Results

Having validated our supervisory signals, we now present the main finetuning results, building from a simple baseline to our final proposed method as summarized in Table II.

1) Imitation Learning (IL)

The pretrained MotionLM model serves as our initial baseline. As a first step, we finetune it using Imitation Learning (IL) on the single, highest-rated human trajectory in each example. This establishes a very strong baseline, dramatically improving both RFS (avgRFS increases from 7.08 to 7.86) and ADE (from 3.19 to 3.07) and confirming the need for preference alignment.

2) Finetuning with High-Level Actions (IL+VL-HLA)

Next, we explore augmenting the IL baseline with VLM-generated HLAs. We find that using HLAs as a model input (IL+VL-HLA(input)) provides a significant boost, particularly to ADE, achieving our best geometric accuracy of 2.715 in table II. This shows that providing a coarse semantic anchor answering the question "WHAT is the correct maneuver?" helps regularize the model and simplifies the prediction task. In contrast, the auxiliary loss approach (IL+VL-HLA (Loss)) yielded no notable improvement, likely because the gradients from a small classification head are insufficient to meaningfully influence the model's primary representations.

3) Vision-Language Guided Preference Tuning (VL-DPO)

Finally, we apply our primary contribution, augmenting the IL baseline with VLM-generated preference pairs via DPO. This method, IL+VL-DPO, achieves the highest RFS[3] (Fig. 5) and competitive ADE. It demonstrates improvements of 11.94% and 4.07% in RFS, and reductions of 10.01% and 6.5% in ADE as compared to the baseline and imitation learning–finetuned models, respectively. This superiority demonstrates that the greatest gains are unlocked by providing a fine-grained, trajectory-level comparative signal that answers the nuanced question: "HOW should this maneuver be executed safely and comfortably?" This approach also holds a critical practical advantage: While the IL+VL-HLA (Input) model achieves a slightly lower ADE, it requires the VLM at inference time, whereas our VL-DPO framework only requires the VLM for offline dataset creation.

To demonstrate the effect of RFS improvements, Fig. 6 shows an example of central-mode trajectory predictions from the MotionLM baseline, imitation learning–finetuned models, and our IL+VL-DPO. In this scenario, a jaywalking pedestrian has nearly finished crossing the street. While baseline models exhibit conservative behavior, our IL+VL-DPO model generates a more efficient and human-like trajectory, demonstrating its superior situational understanding.

D. Discussion and Ablation Studies

1) Impact of HLA Representation

The effectiveness of HLA supervision is highly dependent on both the training strategy (auxiliary loss vs. conditional

[3]As of August 2025, the leading RFS score is 7.986. These scores are not comparable to WOD-E2E leaderboard scores due to differences in pretraining data and finetuning features.

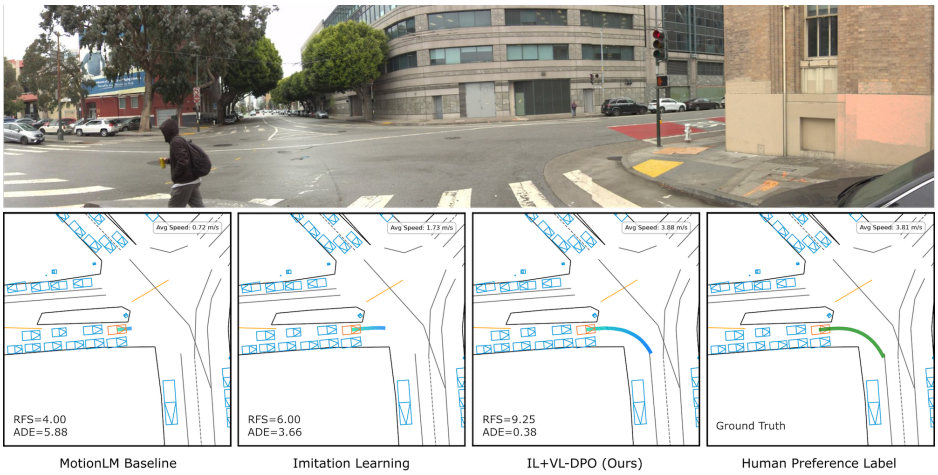

Fig. 6. Comparison of central-mode trajectory prediction plots in top-down images from MotionLM [2], the imitation learning–finetuned model, and our proposed IL+VL-DPO finetuned model, shown alongside the ground-truth human-annotated preference trajectory in an example driving scenario.

Method	RFS ↑	avgRFS ↑	ADE (5s) ↓
Imitation Learning	7.8446	7.8613	3.0752
Maneuver Loss	7.8160	7.7994	2.8284
Direction+Speed Loss	7.8915	7.8699	2.8655
Maneuver+Direction+Speed Loss	7.9240	7.8178	2.8634
Maneuver Input	**8.0604**	**8.0324**	**2.7232**
Direction+Speed Input	7.8743	7.8962	2.9185
Maneuver+Direction+Speed Input	7.8841	7.8733	2.8626

TABLE III

ABLATION STUDY ON VLM HLAS WITH DIFFERENT DEFINITIONS AND TRAINING STRATEGIES, EVALUATED BY RFS AND 5-S ADE.

Method	RFS ↑	avgRFS ↑	ADE (5S) [m] ↓
MotionLM [2]	7.2926	7.0839	3.1121
VL-DPO only	7.4432	7.3681	3.7895
Imitation Learning	7.8446	7.8613	3.0752
IL+Preference-DPO	8.0140	7.9784	2.6689
IL+VL-DPO (Ours)	8.1637	8.0694	2.8756

TABLE IV

ABLATION STUDY OF DPO VARIANTS IN TERMS OF RFS AND ADE.

input) and the semantic representation (Maneuver vs. Direction+Speed). Our ablation (Table III) reveals a clear conclusion: using HLAs as an auxiliary loss provides an insufficient signal, regardless of the definition. In contrast, when used as a conditional input, the holistic Maneuver action provides a powerful inductive bias, substantially improving both RFS and ADE. This finding suggests that for conditioning, a single, comprehensive semantic token is a more effective and direct signal than multiple, disentangled ones.

To analyze the sources of performance gains, we conduct ablation studies on our two finetuning strategies.

2) Dissecting the Efficacy of DPO

We now analyze the key components of our best-performing model, IL+VL-DPO, with results in Table IV.

First, we investigate the relationship between preference optimization and simple imitation learning. When evaluated as a standalone method, VL-DPO only improves RFS over the MotionLM baseline, confirming that the VLM's preference signal is effective for alignment even without any human data. However, this comes at the cost of a regression in geometric accuracy (ADE). In contrast, we observe that pure Imitation Learning yields higher RFS gains while improving ADE over VL-DPO. This is expected: a direct supervised signal on the top human preference trajectory provides a powerful gradient.

This leads to our central finding: the combination, IL+VL-DPO, surpasses both standalone methods. This underscores that DPO's strength is not in replacing imitation learning, but in augmenting it. While IL learns from the single "best" trajectory, DPO leverages all 11 comparisons, teaching the model a nuanced understanding of failure modes that is critical for improving human preference scores.

Next, we compare the data sources for the DPO signal. In Table IV, IL+VL-DPO model outperforms IL+Preference-DPO (which uses human oracle preferences) in RFS. This suggests that the VLM, guided by its structured CoT process, provides a more diverse and informative set of "what not to do" signals than the human annotator. Human preference data is limited to 3 pairs and can be noisy, whereas our VLM generates up to 11 systematic comparisons per scene, offering a richer and more effective optimization signal.

V. CONCLUSION AND FUTURE WORK

This work introduces VL-DPO, a vision-language-guided preference finetuning framework, and establishes a clear hierarchy for leveraging VLM-generated signals. While imitation learning on a human-preferred trajectory provides a strong baseline, we show that augmenting it with VLM-generated High-Level Actions yields further improvements by providing a coarse semantic anchor on "WHAT" maneuver to execute. However, our primary finding is that the greatest gains are unlocked by the fine-grained, comparative feedback

of VL-DPO, which addresses the critical question of "HOW" a maneuver should be performed safely and comfortably.

These findings lead to two powerful conclusions. **Methodologically**, the most effective approach is a combination of imitation learning on high-quality positive examples and DPO on a rich set of comparative examples. This dual signal teaches the model both what to do and, critically, the nuanced details of what to avoid. **Architecturally**, our results validate a modular paradigm where a frozen zero-shot VLM provides state-of-the-art alignment for a specialized motion model.

ACKNOWLEDGMENT

We thank Kate Tolstaya, Kratarth Goel, and Neerja Thakkar for their valuable contributions to this work.

REFERENCES

[1] X. Huang, E. M. Wolff, P. Vernaza, T. Phan-Minh, H. Chen, D. S. Hayden, M. Edmonds, B. Pierce, X. Chen, P. E. Jacob, X. Chen, C. Tairbekov, P. Agarwal, T. Gao, Y. Chai, and S. Srinivasa, "DriveGPT: Scaling autoregressive behavior models for driving," in *Forty-second International Conference on Machine Learning*, 2025.

[2] A. Seff, B. Cera, D. Chen, M. Ng, A. Zhou, N. Nayakanti, K. S. Refaat, R. Al-Rfou, and B. Sapp, "Motionlm: Multi-agent motion forecasting as language modeling," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8579–8590.

[3] N. Nayakanti, R. Al-Rfou, A. Zhou, K. Goel, K. S. Refaat, and B. Sapp, "Wayformer: Motion forecasting via simple & efficient attention networks," in *ICRA*, 2023.

[4] B. Varadarajan, A. Hefny, A. Srivastava, K. S. Refaat, N. Nayakanti, A. Cornman, K. Chen, B. Douillard, C. Lam, D. Anguelov, and B. Sapp, "Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction," in *ICRA*, 2022.

[5] M. Baniodeh, K. Goel, S. Ettinger, C. Fuertes, A. Seff, T. Shen, C. Gulino, C. Yang, G. Jerfel, D. Choe *et al.*, "Scaling laws of motion forecasting and planning–a technical report," *arXiv preprint arXiv:2506.08228*, 2025.

[6] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, J. Beißwenger, P. Luo, A. Geiger, and H. Li, "Drivelm: Driving with graph visual question answering," in *European conference on computer vision*. Springer, 2024.

[7] X. Tian, J. Gu, B. Li, Y. Liu, Y. Wang, Z. Zhao, K. Zhan, P. Jia, X. Lang, and H. Zhao, "DriveVLM: The convergence of autonomous driving and large vision-language models," in *8th Annual Conference on Robot Learning*, 2024.

[8] S. Zhang, W. Huang, Z. Gao, H. Chen, and C. Lv, "Wisead: Knowledge augmented end-to-end autonomous driving with vision-language model," *arXiv preprint arXiv:2412.09951*, 2024.

[9] J.-J. Hwang, R. Xu, H. Lin, W.-C. Hung, J. Ji, K. Choi, D. Huang, T. He, P. Covington, B. Sapp, Y. Zhou, J. Guo, D. Anguelov, and M. Tan, "EMMA: End-to-end multimodal model for autonomous driving," *Transactions on Machine Learning Research*, 2025.

[10] X. Zhou, X. Han, F. Yang, Y. Ma, and A. C. Knoll, "Opendrivevla: Towards end-to-end autonomous driving with large vision language action model," *arXiv preprint arXiv:2503.23463*, 2025.

[11] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning*, 2023.

[12] S. Yang, H. Li, Y. Chen, B. Wang, Y. Tian, T. Wang, H. Wang, F. Zhao, Y. Liao, and J. Pang, "Instructvla: Vision-language-action instruction tuning from understanding to manipulation," *arXiv preprint arXiv:2507.17520*, 2025.

[13] Z. Zhou, Y. Zhu, M. Zhu, J. Wen, N. Liu, Z. Xu, W. Meng, R. Cheng, Y. Peng, C. Shen *et al.*, "Chatvla: Unified multimodal understanding and robot control with vision-language-action model," *arXiv preprint arXiv:2502.14420*, 2025.

[14] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[15] H. Cui, V. Radosavljevic, F.-C. Chou, T.-H. Lin, T. Nguyen, T.-K. Huang, J. Schneider, and N. Djuric, "Multimodal trajectory predictions for autonomous driving using deep convolutional networks," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[16] J. Hong, B. Sapp, and J. Philbin, "Rules of the road: Predicting driving behavior with a convolutional model of semantic interactions," in *CVPR*, 2019.

[17] Y. Chai, B. Sapp, M. Bansal, and D. Anguelov, "Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction," in *Conference on Robot Learning*, 2020, pp. 86–99.

[18] S. Casas, C. Gulino, R. Liao, and R. Urtasun, "Spagnn: Spatially-aware graph neural networks for relational behavior forecasting from sensor data," in *ICRA*, 2020.

[19] M. Liang, B. Yang, R. Hu, Y. Chen, R. Liao, S. Feng, and R. Urtasun, "Learning lane graph representations for motion forecasting," in *European Conference on Computer Vision*, 2020.

[20] S. Shi, L. Jiang, D. Dai, and B. Schiele, "Motion transformer with global intention localization and local movement refinement," *Advances in Neural Information Processing Systems*, 2022.

[21] X. Jia, P. Wu, L. Chen, H. Li, Y. Liu, and J. Yan, "Hdgt: Heterogeneous driving graph transformer for multi-agent trajectory prediction via scene encoding," *CoRL*, 2022.

[22] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[23] B. Jiang, S. Chen, Q. Xu, B. Liao, J. Chen, H. Zhou, Q. Zhang, W. Liu, C. Huang, and X. Wang, "Vad: Vectorized scene representation for efficient autonomous driving," in *ICCV*, 2023, pp. 8306–8316.

[24] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang *et al.*, "Planning-oriented autonomous driving," in *CVPR*, 2023.

[25] Z. Guo, K. Gubernatorov, S. Asfaw, Z. Yagudin, and D. Tsetserukou, "Vdt-auto: End-to-end autonomous driving with vlm-guided diffusion transformers," *arXiv preprint arXiv:2502.20108*, 2025.

[26] D. Hegde, R. Yasarla, H. Cai, S. Han, A. Bhattacharyya, S. Mahajan, L. Liu, R. Garrepalli, V. M. Patel, and F. Porikli, "Distilling multimodal large language models for autonomous driving," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

[27] Y. Wang, R. Jiao, S. S. Zhan, C. Lang, C. Huang, Z. Wang, Z. Yang, and Q. Zhu, "Empowering autonomous driving with large language models: A safety perspective," in *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.

[28] Y. Yang, Q. Zhang, K. Ikemura, N. Batool, and J. Folkesson, "Hard cases detection in motion prediction by vision-language foundation models," in *2024 IEEE Intelligent Vehicles Symposium (IV)*, 2024.

[29] Y. Xu, Y. Hu, Z. Zhang, G. P. Meyer, S. K. Mustikovela, S. Srinivasa, E. M. Wolff, and X. Huang, "Vlm-ad: End-to-end autonomous driving through vision-language model supervision," *arXiv preprint arXiv:2412.14446*, 2024.

[30] C. Pan, B. Yaman, T. Nesti, A. Mallik, A. G. Allievi, S. Velipasalar, and L. Ren, "Vlp: Vision language planning for autonomous driving," in *CVPR*, 2024.

[31] S. Wang, Z. Yu, X. Jiang, S. Lan, M. Shi, N. Chang, J. Kautz, Y. Li, and J. M. Alvarez, "Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

[32] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, 2022.

[33] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," *Advances in neural information processing systems*, vol. 30, 2017.

[34] T. Tian and K. Goel, "Direct post-training preference alignment for multi-agent motion generation model using implicit feedback from pre-training demonstrations," in *The Thirteenth International Conference on Learning Representations*, 2025.

[35] C. Zhang, P. Hao, X. Cao, X. Hao, S. Cui, and S. Wang, "Vtla: Vision-tactile-language-action model with preference learning for insertion manipulation," *arXiv preprint arXiv:2505.09577*, 2025.

[36] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen *et al.*, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint arXiv:2507.06261*, 2025.