# DIRECT MULTI-AGENT MOTION GENERATION PREFERENCE ALIGNMENT WITH IMPLICIT FEEDBACK FROM DEMONSTRATIONS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Recent advancements in Large Language Models (LLMs) have transformed motion generation models in embodied applications such as autonomous driving and robotic manipulation. While LLM-type motion models benefit from scalability and efficient formulation, there remains a discrepancy between their token-prediction imitation objectives and human preferences. This often results in behaviors that deviate from human-preferred demonstrations, making post-training behavior alignment crucial for generating human-preferred motions. Post-training alignment requires a large number of preference rankings over model generations, which are costly and time-consuming to annotate in multi-agent motion generation settings. Recently, there has been growing interest in using expert demonstrations used in the pre-training phase to scalably build preference data for alignment. However, these methods often adopt a worst-case scenario assumption, treating all generated samples from the reference model as unpreferred and relying on expert demonstrations to directly or indirectly construct preferred generations. This approach overlooks the rich signal provided by preference rankings among the model's own generations. In this work, instead of treating all generated samples as equally unpreferred, we propose a principled approach that leverages the implicit preferences encoded in expert demonstrations to construct preference rankings among the generations produced by the reference model, offering more nuanced guidance at a lower-cost. We present the first investigation of direct preference alignment for multi-agent motion token-prediction models using implicit preference feedback from demonstrations. We apply our approach to large-scale traffic simulation and demonstrate its effectiveness in improving the realism of generated behaviors involving up to 128 agents, making a 1M token-prediction model comparable to state-of-the-art large models by relying solely on implicit feedback from demonstrations, without requiring additional human annotations or incurring high computational costs. Furthermore, we provide an in-depth analysis of preference data scaling laws and their effects on over-optimization, offering valuable insights for future investigations.

## 1 INTRODUCTION

The recent advances in Large Language Models (LLMs) (Achiam et al., 2023) have significantly impacted the design of motion generation for embodied tasks such as autonomous driving (Seff et al., 2023; Philion et al., 2024), robot manipulation (Brohan et al., 2023), and humanoid locomotion (Radosavovic et al., 2024). Formulating motion generation as a next-token prediction task not only provides a unified framework for modeling sequential decision-making tasks but also provides opportunities for leveraging pre-trained LLMs for more cost-effective training and leveraging the common-sense reasoning capabilities inherent in these models (Tian et al., 2024). However, despite the remarkable progress, robots relying on these large token-prediction models do not automatically become better at doing what *humans* prefer due to the misalignment between the training objective and the underlying reward function that incentivizes the expert demonstrations. This discrepancy underscores the challenge of ensuring that motion models trained with next-token prediction are effectively aligned with expert-preferred behaviors.

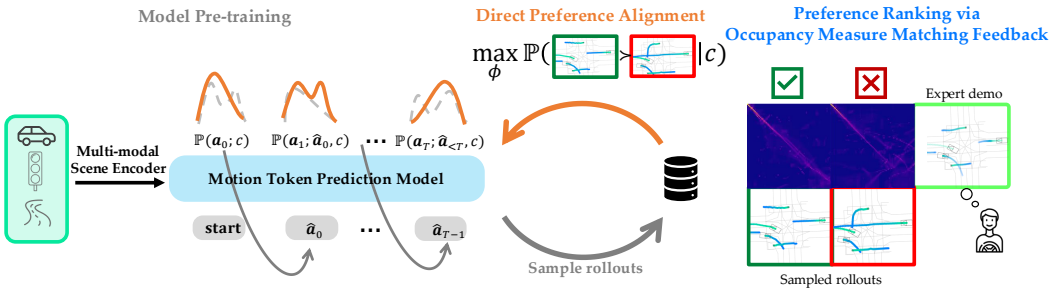


Figure 1: **Direct preference alignment from occupancy measure matching feedback for realistic traffic simulation.** DPA-OMF is a simple yet effective alignment-from-demonstration approach that aligns a pre-trained traffic simulation model with human preferences. It leverages the implicit preferences encoded in expert demonstrations to construct preference rankings among generations sampled from the reference model, and updates the model. The gray dotted lines above the motion token prediction model indicate the reference model's motion token distributions at each prediction step, and the orange lines represent the probabilities after the alignment process. $\hat{\mathbf{a}}_t$ denotes agents' action tokens sampled from the predicted distribution $\mathbb{P}(\mathbf{a}_t; c, \hat{\mathbf{a}}_{t-1})$ during inference time, $c$ denotes the scene context representation (more details in Section 3.1).

Preference-based alignment has emerged as a crucial component in LLM post-training stage, aiming to reconcile the disparity between the next-token prediction objective and human preferences. Among various frameworks, direct alignment algorithms (e.g., direct preference optimization (Rafailov et al., 2024b)) are particularly appealing due to their simplicity of training and computational efficiency. Specifically, direct alignment algorithms collect preference rankings from humans over model generations and directly update the model to maximize the likelihood of preferred behaviors over unpreferred ones. However, in complex embodied settings, such as joint motion generation across hundreds of agents, obtaining such preference data can be very challenging. Human annotators must analyze intricate and nuanced motions, which is a time-consuming process, making the scalability of direct alignment methods difficult in these scenarios.

While soliciting rankings from experts provides explicit preference information, we argue that expert demonstrations used in the pre-training phase inherently encode implicit human preferences, which can be reused to align a pre-trained motion generation model in a cost-effective and efficient way, beyond their role in supervised pre-training. Recently, Alignment from Demonstrations (AFD) (Li et al., 2024; Sun & van der Schaar, 2024; Chen et al., 2024b) has emerged as a valuable technique for automatically generating preference data using existing expert demonstrations, allowing preference alignment to scale at a low cost. However, previous methods typically assume a worst-case scenario: treating all generations from the reference model as unpreferred, and relying on expert demonstrations to directly (Chen et al., 2024b) or indirectly (Sun & van der Schaar, 2024) construct preferred generations. This assumption overlooks the rich signals provided by the preference rankings among the generated samples from the reference model, which can offer more nuanced guidance than simply treating all generated samples as unpreferred.

*Instead of treating all generated samples as equally bad, we propose leveraging the implicit preferences encoded in expert demonstrations to construct preference rankings* ***among*** *the generations produced by the reference model, offering more nuanced guidance at a lower-cost.*

Our approach draws inspiration from inverse reinforcement learning (Abbeel & Ng, 2004; Ho & Ermon, 2016), where alignment between a generated behavior and the expert demonstration is measured through occupancy measure matching. We propose Direct Preference Alignment from Occupancy Measure Matching Feedback, DPA-OMF, a principled approach using optimal transport to define an implicit preference distance function. This function measures the alignment between a generated sample and the expert demonstration through occupancy measure matching in a semantically meaningful feature space, and is then used to rank the generated samples according to their alignment with expert demonstrations, producing more nuanced preference data at scale to align the motion generation model (Figure 1).

We present the first investigation of direct preference alignment for multi-agent motion token-prediction models using implicit preference feedback from demonstrations. We apply our approach

to large-scale realistic traffic simulations and demonstrate its effectiveness in improving the realism of generated behaviors involving up to 128 agents, making a 1M token-prediction model comparable to state-of-the-art large models by relying solely on implicit feedback from demonstrations, without requiring additional human annotations or incurring high computational costs.

**Contribution**. In this work, we consider the problem of efficient post-training alignment of a token-prediction model for multi-agent motion generation. We propose Direct Preference Alignment from Occupancy Measure Matching Feedback (DPA-OMF), a simple yet principled approach that leverages the expert demonstrations to generate implicit preference feedback and significantly improves the pre-trained model's generation quality without additional human preference annotation, reward learning, or complex reinforcement learning. To the best of our knowledge, our work is the first work that demonstrates the benefits of preference alignment on large-scale multi-agent motion generation models using implicit preference feedback from demonstrations, and provides detailed analysis on preference data scaling laws and their effects on preference over-optimization.

## 2 Related Works

**Scaling alignment using AI feedback.** Previous works have proposed leveraging synthetic AI feedback to scale preference data in LLM applications, either from a single model (Bai et al., 2022; Lee et al.; Mu et al.), an ensemble of teacher models (Tunstall et al., 2023), or through targeted context prompting, where the model is instructed to generate both good and bad outputs for comparison (Yang et al., 2023; Liu et al., 2024). Unfortunately, these frameworks are not directly applicable to embodied contexts, such as autonomous driving, due to the absence of high-quality, open-source, and input-unified foundational models. What makes input-unified foundational models especially challenging is that different embodied models often use incompatible input modalities or features, making it difficult to transfer feedback across models effectively.

**Alignment from expert demonstrations.** Alignment from Demonstrations (AFD) has emerged as a valuable technique for automatically generating preference data using expert demonstrations, enabling preference alignment to scale effectively. For example, Chen et al. (2024b) fine-tunes a pre-trained reference model in a self-play manner, where the optimized model maximizes the log-likelihood ratio between expert demonstrations and self-generated samples. However, this method treats all model-generated samples as unpreferred, overlooking the valuable information embedded in the preference rankings among those samples. Additionally, it suffers from bias introduced by the *heterogeneity* of the preference data: since the preference data are drawn from two different sources (human vs. optimized model), the discrimination objective may emphasize the differences between models rather than focusing on the key aspects that truly evaluate the quality of the generated behaviors. To address the heterogeneity issue, Sun & van der Schaar (2024) proposes using expert demonstrations to first fine-tune the reference model through supervised learning, creating an expert model. Then, samples generated by the fine-tuned expert model are treated as positive examples, while samples from the initial model are treated as negative examples to construct a preference dataset. While this approach helps mitigate the heterogeneity problem, it requires training an additional expert model, and there is no guarantee that all generations from the expert model will consistently be superior to those from the initial model. *Unlike previous works, our approach constructs preference rankings among the generations produced by the reference model, effectively addressing the heterogeneity issue while providing more nuanced guidance at a lower cost.*

**Motion alignment via divergence minimization.** Adversarial imitation learning (IL) aims to align the behavior of the learning agent with expert demonstrations by minimizing the Jensen-Shannon divergence between the agent's action occupancy measure and that of the expert (Ho & Ermon, 2016; Song et al., 2018). These methods jointly train a policy model and a discriminator: the policy model is trained to imitate expert demonstrations, while the discriminator is trained to separate the generated samples from the expert demonstrations and informs the policy model. Additionally, adversarial training is known for its instability and high computational cost (Kodali et al., 2017; Yang et al., 2022), making it particularly problematic in post-training alignment, where stability and computational efficiency are crucial. *Similar to adversarial IL, we employ occupancy measure matching to assess the alignment between generated samples and expert demonstrations. However, our ap-*

*proach differs by using preference rankings over the generated samples to inform the alignment process, rather than relying on a signal learned from separating all generated samples from expert demonstrations. Our approach provides more informative guidance while also being significantly more computationally efficient and more suitable for post-training alignment.*

**Preference alignment for realistic traffic simulation.** Preference-based alignment has recently been used to enhance the realism of traffic generation (Cao et al., 2024; Wang et al., 2024b). However, these previous works rely on Reinforcement Learning with Human Feedback (RLHF) as the alignment approach and learn rewards from low-fidelity simulator data. In contrast, our work seeks to improve traffic generation models without the need for reinforcement learning or explicit reward learning. Instead, we utilize an implicit preference distance derived from real human demonstrations to guide the alignment, and demonstrate improvements in a much larger scale.

# 3 Direct Multi-agent Motion Generation Preference Alignment with Implicit feedback from demonstrations

## 3.1 Multi-agent motion generation as a next-token prediction task

We assume access to a set of expert demonstrations, $\mathcal{D}_{\text{e}}$, where each example consists of a joint action sequence $\xi = \{\mathbf{a}_0, \ldots, \mathbf{a}_T\}$ and a scene context representation $c$. The joint action sequence represents the behaviors of $N$ agents over a generation horizon $T$, conditioned on the given scene context $c$. Each joint action at time step $t$ is a collection of action tokens for all $N$ agents, denoted as $\mathbf{a}_t = (a_t^1, \ldots, a_t^N)$. Following Seff et al. (2023), our learning objective is to train a generative model parameterized by $\theta$, which maximizes the likelihood of the ground truth joint action at time step $t$, conditioned on all previous joint actions and the initial scene context:

$$\max_{\theta} \mathbb{E}_{(\xi,c)\sim\mathcal{D}_{\text{e}}} \left[\Pi_{t=0}^{T} \pi_\theta(\mathbf{a}_t|\mathbf{a}_{<t}; c)\right], \tag{1}$$

where $\pi_\theta(\mathbf{a}_t|\mathbf{a}_{<t}; c) := \mathbb{P}(\mathbf{a}_t|\mathbf{a}_{t-1}, \ldots, \mathbf{a}_0; c)$. We denote the model trained with the token prediction objective as the reference model $\pi_{ref}$. Note that the optimal solution of (1) is the same as seeking a policy $\pi$ that minimizes the forward Kullback-Leibl (KL) divergence to the expert policy: $\min_\theta \text{KL}(\pi_\theta||\pi_{\text{e}})$. Consequently, the resulting policy tends to exhibit mass-covering behavior (Wang et al., 2024a), making it prone to deviating from the expert's true policy, which necessitates post-training alignment.

## 3.2 Implicit preference feedback from demonstrations

In this section, we introduce our approach to leveraging expert demonstrations for scalable construction of preference feedback. The key idea is to define an implicit preference distance function that measures the alignment between a generated sample and the expert demonstration using occupancy measure matching. This distance is then used to rank the samples generated by the reference model, forming the basis for constructing preference feedback.

With the reference model pre-trained using the next-token prediction objective (1), we can construct a rollout (a generated motion sampled from the model) set, $\mathcal{D}_{\pi_{\text{ref}}}$, in which each example contains a set of $K$ rollouts, $\{\xi^1, \ldots, \xi^K\}$, sampled from the reference model, and the associated scene context $c$. We aim to find a preference distance function $d$ that approximately measures the similarity over a triplet $(\xi^i, \xi^j, \xi^k)$: We treat $\xi^i$ as an anchor, then if the preference distance between $\xi^i$ and $\xi^j$ is smaller than that between $\xi^i$ and $\xi^k$, $\xi^j$ is more similar to the anchor than $\xi^k$: $d(\xi^i, \xi^j) > d(\xi^i, \xi^k) \implies \xi^i \succ \xi^j \succ \xi^k$.

Fundamental work in IRL (Abbeel & Ng, 2004) advocates for using occupancy measure matching to assess the alignment between the policy induced by a recovered reward function and human demonstrations. If the policy's occupancy measure closely matches that of the human, the recovered reward function is considered better aligned with the human's true preferences. Building on this insight, we leverage the Optimal Transport (OT) method (Villani et al., 2009) as a principled approach to define an implicit preference distance function that measures the alignment between a rollout and the expert demonstration. We then use this preference distance to rank rollouts from the reference model to construct the preference dataset. OT has been successfully used to measure alignment between behaviors in prior single-agent reinforcement learning works (Xiao et al., 2019) with the key

difference in our work being used in multi-agent settings to fine-tune a generative model without the need for reinforcement learning (see Q4 in Appendix A for more details).

**Rollout occupancy measure**. Let $\mathbf{o} = \{o_t\}_{t=0}^{T}$ represent a sequence of scene observations obtained by rolling out a joint action sequence $\xi$ in the initial scene $c$ over $T$ time steps. The empirical occupancy measure associated with $(\xi, c)$ is a discrete probability measure defined as $\rho_{(\xi,c)} = \frac{1}{T}\sum_{t=0}^{T}\delta_{\Phi(o_t)}$, where $\delta_{\Phi(o_t)}$ is a Dirac distribution centered on $\Phi(o_t) = [\phi(o_t^1), \ldots, \phi(o_t^N)]$. Here, $\phi$ is a feature encoder that maps each agent's information into a feature vector. Intuitively, the rollout occupancy measure represents the distribution of features visited by the generation policy throughout the generation horizon.

**Implicit preference distance.** Optimal transport measures the distance between two discrete probability measures by solving the optimal coupling $\mu^* \in \mathbb{R}^{T\times T}$ that transports a rollout occupancy measure, $\rho_{(\xi_s,c)}$, to the expert occupancy measure, $\rho_{(\xi_e,c)}$, with minimal cost. Instead of computing the scene-level optimal coupling, we compute the agent-level optimal coupling and then aggregate them for computing the scene-level alignment, assuming the generative model only generates the behaviors of pre-defined agents but not insert nor remove agents from the scene. Specifically, for each agent $i$, we compute optimal coupling between agent $i$'s empirical occupancy measure induced by the sampled rollout $\xi^i$ and agent $i$'s empirical occupancy measure from the demonstration $\xi_e^i$ by minimizing the Wasserstein distance the two:

$$\mu^{i,*} = \underset{\mu\in\mathcal{M}(\rho_{(\xi_e^i,c)},\rho_{(\xi_s^i,c)})}{\arg\min} \sum_{t=1}^{T}\sum_{t'=1}^{T} c\big(\phi(o_{t,s}^i), \phi(o_{t',e}^i)\big)\mu_{t,t'}. \tag{2}$$

where $\mathcal{M}(\rho_{(\xi_s^i,c)}, \rho_{(\xi_e^i,c)}) = \{\mu \in \mathbb{R}^{T\times T} : \mu\mathbf{1} = \rho_{(\xi_s^i,c)}, \mu^T\mathbf{1} = \rho_{(\xi_e^i,c)}\}$ is the set of coupling matrices and $c : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a cost function defined on the support of the measure (e.g., cosine distance) with $n$ being the dimension of the feature. This gives rise to the following distance that measures the alignment between a rollout and expert demonstration:

$$d(\xi_s, \xi_e; c) = \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T} c\big(\phi(o_{t,e}^i), \phi(o_{t',e}^i)\big)\mu_{t,t'}^{i,*}. \tag{3}$$

We use this implicit preference distance function to build pair-wise preference rankings using the rollouts from $\mathcal{D}_{\pi_{\text{ref}}}$ and construct the preference dataset $\mathcal{D}_{\pi_{\text{pref}}}$. Each example in $\mathcal{D}_{\pi_{\text{pref}}}$ contains $N_{\text{pref}}$ pairwise comparisons, $(\xi^+ \succ \xi^-)^{1,\ldots,N_{\text{pref}}}$, along with the associated scene context $c$.

### 3.3 Direct preference alignment via contrastive preference learning

Using the automatically constructed preference dataset $\mathcal{D}_{\pi_{\text{pref}}}$, we apply a multi-agent extension of the contrastive preference learning algorithm (Hejna et al., 2023), combined with reference policy regularization, to directly fine-tune the reference model:

$$\max_{\theta} \mathbb{E}_{(\xi^+,\xi^-,c)\sim\mathcal{D}_{\text{pref}}}\left[-\log\frac{\exp\sum_t \gamma^t\alpha\log\frac{\pi_\theta(\mathbf{a}_t^+|\mathbf{a}_{<t}^+,c)}{\pi_{\text{ref}}(\mathbf{a}_t^+|\mathbf{a}_{<t}^+,c)}}{\exp\sum_t \gamma^t\alpha\log\frac{\pi_\theta(\mathbf{a}_t^+|\mathbf{a}_{<t}^+,c)}{\pi_{\text{ref}}(\mathbf{a}_t^+|\mathbf{a}_{<t}^+,c)} + \exp\sum_t \gamma^t\alpha\log\frac{\pi_\theta(\mathbf{a}_t^-|\mathbf{a}_{<t}^-,c)}{\pi_{\text{ref}}(\mathbf{a}_t^-|\mathbf{a}_{<t}^-,c)}}\right]. \tag{4}$$

## 4 Experiment Design

**Realistic traffic scene generation.** We validate our approach in the large-scale realistic traffic scene generation challenge (WOSAC), where the model is tasked with generating eight seconds of realistic interactions among multiple heterogeneous agents (up to 128) at 10 Hz, based on one second of past observations (Montali et al., 2024). The WOSAC challenge uses a realism metric to evaluate the quality of the traffic generation model. The realism of a generative model is defined as the empirical negative log likelihood of the ground truth scene evolution under the distribution induced by 32 scene rollouts sampled from that model (binned into discrete histograms).

**Token prediction model and preference dataset construction.** We use the MotionLM model (Seff et al., 2023) as our generation model (1M trainable parameters) (more details in Appendix C)

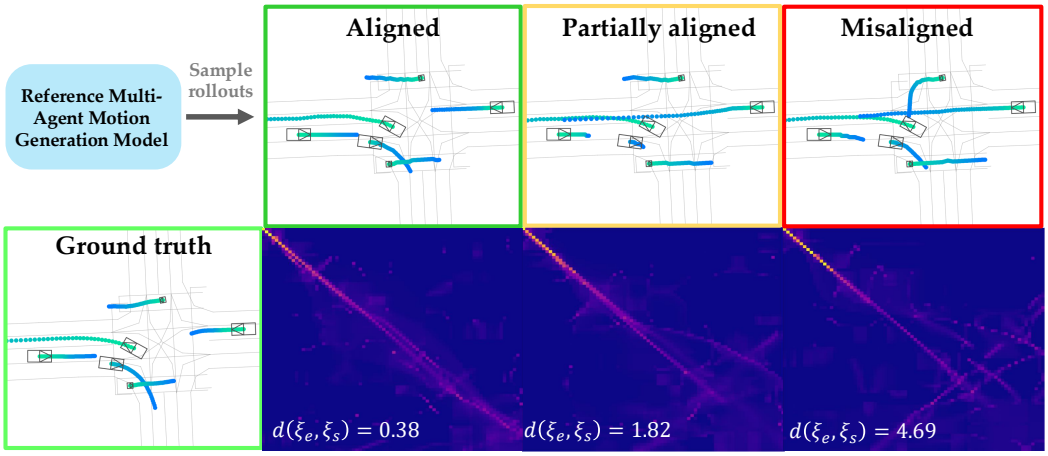


Figure 2: **Alignment visualization.** The heat map visualizes the optimal coupling between a generated traffic simulation and the ground truth scene evolution. More peaks along the diagonal indicate better alignment between the behaviors (i.e., a smaller preference distance). The traffic simulation with a small preference distance (left) shows behavior that is well-aligned with the ground truth, while the simulation with a larger distance (right) exhibits inconsistencies, such as the pedestrian's generated motion colliding with an oncoming vehicle.

and first train it using the next-token prediction objective on the Waymo Open Motion Dataset for 1.2M steps to obtain a reference model. For each training example, we sample 64 rollouts from the reference model and rank them using the preference distance function. The 16 closest rollouts are treated as preferred samples, while the 16 farthest are considered unpreferred, constructing 16 comparisons per example. Following previous work on learning rewards for autonomous driving (Sun et al., 2018; Chen et al., 2024a), we use features that capture the modeled agent's safety, comfort, and progress to encode the agent's state information at each time step when building the rollout occupancy measure, including: [collision status, distance to road boundary, minimum clearance to other road users, control effort, speed]. When solving the optimal transport plan (2), we use the L2 cost between features with the following weights: [10, 5, 2, 1, 1]. These features are also used to encode the agent's state in the realism metric. It is important to note that the preference distance is not the same as the realism metric. The preference distance measures the alignment between a rollout and the expert demonstration, whereas the realism metric assesses the likelihood of the expert demonstration given all the rollouts.

## 5 Results

### 5.1 On the validity of the implicit preference distance function

In this section, we qualitatively and quantitatively investigate if the implicit preference distance reflects the behavior alignment (i.e., smaller distance means more aligned behavior).

**Qualitative example.** In Figure 2, we present a qualitative example illustrating how the coupling matrix reflects the behavior alignment between generated traffic simulations and expert demonstrations. Traffic simulations with smaller preference distances exhibit behavior more closely aligned with the expert demonstrations. Additional qualitative examples can be found in the Appendix D.

**Baseline**. To quantitatively evaluate our preference distance function in a controlled experiment, we conduct a post-selection analysis, where we select sampled traffic simulations from the reference model and analyze the relationship between the realism of these samples and their

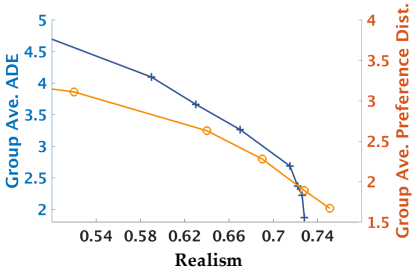


Figure 3: Relationship between the WOSAC realism of selected traffic simulations and their group-averaged distance from the expert demonstration.

group-averaged distance to the expert demonstration. We use the Average Displacement Error (ADE) (L2 distance between a sampled traffic simulation and the expert demonstration) as a baseline, which is a commonly used distance for evaluating trajectory generation performance in autonomous driving.

**Experiment setup**. We control the distance to the expert demonstration, and measure the realism of the selected sampled traffic simulations. Specifically, we first sample 128 traffic simulations from the pre-trained reference model, rank them using the candidate distance (ADE or preference distance), and then select traffic simulations as the final model output using a sliding window of size 32. For example, model variant 1 outputs the top 32 ranked traffic simulations, model variant 2 outputs simulations ranked from 2 to 33, and so on. We then measure the corresponding WOSAC realism of each model variant to study its relationship to each candidate distance metric.

**Results.** In Figure 3, we illustrate the relationship between the WOSAC realism of selected traffic simulations and their group-averaged distance from the expert demonstration. We observe that ADE is informative in reflecting behavior alignment only up to a certain point: initially, as ADE decreases, the realism improves. However, once the traffic simulations reach a reasonable level of realism, further reductions in ADE do not result in significant improvements in the realism. In contrast, the preference distance function correlates more strongly with the realism and demonstrates a better effective range when using the same reference model. This highlights its effectiveness in measuring the alignment between a generated traffic simulation and the expert demonstration.

### 5.2 On the value of preference alignment

In this section, we compare our approach with various post-training alignment methods to demonstrate its effectiveness. Our experiments focus on evaluating the impact of two factors: 1) different feedback signals, and 2) the alignment approach.

**Baseline.** We compare **DPA-OMF** against (1) **DPA-ADEF**, which constructs preference rankings using ADE as the distance function to rank traffic simulations sampled from the reference model, then fine-tunes the reference model using (4); (2) **SFT-bestOA**, which selects the top 32 ranked traffic simulations, measured by the preference distance, from the reference model and uses them as new labels for supervised fine-tuning, aiming to directly distill the preference into the reference model; (3) **SFT-bestADE**, similar to **SFT-bestOA** but use the ADE to pick the top ranked samples;(4) We also list the performance of SOTA models that are typically much larger with sophisticated designs. We use the same reference model for post-training alignment and fine-tune for 200k steps in all experiments.

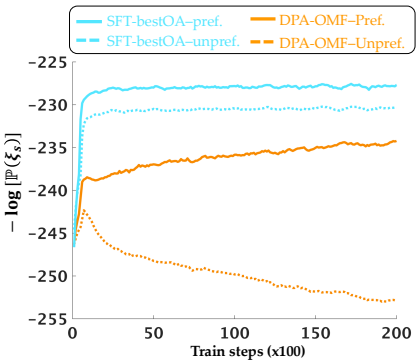


Figure 4: The log-likelihood of preferred/unpreferred rollouts from the reference model when using **DPA-OMF** versus **SFT-bestOA** for alignment.

**Results.** We present the quantitative evaluation of each method's realism and trajectory L2 error in Table 1. **DPA-OMF** significantly outperforms all baselines in terms of alignment between the generated traffic simulations and the expert demonstrations, while the baselines struggle to improve—and in some cases, even degrade (**DPA-ADEF** and **SFT-bestADE**)—the realism. Interestingly, **SFT-bestOA** does not seem to improve the realistic metric too much although it is using the same preferred traffic simulations as learning signals just like **DPA-OMF**. Figure 4 sheds more light on this. In Figure 4, we show the negative-loglikelihood of both preferred traffic simulations and unpreferred traffic simulations from the reference model when using **DPA-OMF** and **SFT-bestOA** to align the model. We can see that the likelihood of preferred traffic simulations is increasing and the likelihood of unpreferred traffic simulations is decreasing when using **DPA-OMF** to align the model, while the likelihood of unpreferred traffic simulations is increasing when using **SFT-bestOA** to align the model. This may be due to the fact that both the preferred and the unpreferred samples are from the same distribution. This demonstrates the importance of explicitly considering the negative signals when aligning the model. In Figure 5, we present a qualitative visualization of the generated traffic simulations from each approach. We sample 64 rollouts from each model and display the most likely

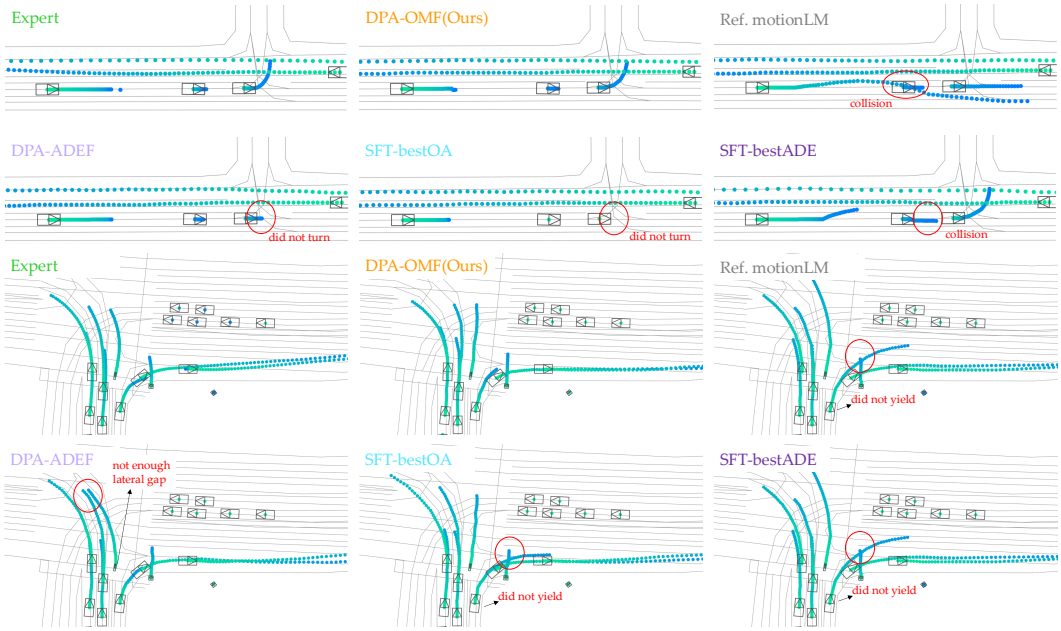


Figure 5: **Traffic simulation generation visualization.** Our approach produces traffic simulations that are more closely aligned with expert demonstrations, while baseline models generate simulations that are only partially aligned or misaligned (highlighted in red texts).

generation. Our approach produces traffic simulations that are more closely aligned with expert demonstrations, whereas the baseline models generate simulations that are only partially aligned or misaligned. While our approach improves the reference model through cost-effective preference alignment, it still falls short a bit compared to some SOTA methods. We provide more discussions in $Q3$ of Appendix A.

| Method | Kinematic ↑ | Interactive↑ | Map compliance↑ | Composite realism↑ | minADE↓ |
|---|---|---|---|---|---|
| **MotionLM** (1M reference model) | **0.417** | 0.778 | 0.815 | 0.721 | 1.398 |
| **DPA-ADEF** | 0.393 | 0.780 | 0.812 | 0.714 | **1.379** |
| **SFT-bestADE** | 0.406 | 0.773 | 0.816 | 0.715 | 1.392 |
| **SFT-bestOA** | 0.410 | 0.781 | 0.826 | 0.723 | 1.428 |
| **DPA-OMF** (ours) | 0.415 | **0.786** | **0.867** | **0.739** | 1.413 |
| BehaviorGPT(Zhou et al., 2024) (3M) | 0.433 | 0.799 | 0.859 | 0.747 | 1.415 |
| Trajeglish(Philion et al., 2024) (35M) | 0.415 | 0.786 | 0.867 | 0.721 | 1.544 |
| SMART(Wu et al., 2024) (102M) | 0.479 | 0.806 | 0.864 | 0.761 | 1.372 |

Table 1: Realism of difference methods. Our approach improves the realism of the reference model (shaded in grey) without requiring additional reward learning or reinforcement learning.

**On importance of the features.** Despite the OT-based preference score's effectiveness in measuring the alignment between a generated traffic simulation and the expert demonstration, the relationship between the OT-based preference score and human preference is not strictly monotonic. This relationship heavily depends on the features used to compute the feature occupancy measure. We consider a set of features commonly used in IRL research to compute the preference distance (described in Section 4). In Table 2, we present an ablation study analyzing the effect of each feature on the effectiveness of the approach. Although the best performance is achieved when all features are active, it is interesting to note that using only the collision status when computing the preference distance for model alignment still leads to improvements. We hypothesize that this is because the reference model already performs reasonably well in generating behaviors but lacks awareness of collisions. However, when using only the progress or comfort feature to compute the preference distance, both the realism (due to an increased collision rate) and ADE regress. This highlights the importance of using a comprehensive set of features to accurately characterize driving behaviors. Further discussions on the limitations and potential solutions can be found in $Q1$ of Appendix A.

| Features | Kinematic | Interactive | Map compliance | Composite realism | minADE |
|---|---|---|---|---|---|
| **Collision only** | 0.389 | **0.788** | 0.833 | 0.724 | 1.527 |
| **Progress only** | **0.421** | 0.760 | 0.812 | 0.710 | 1.483 |
| **Comfort only** | 0.411 | 0.744 | 0.820 | 0.705 | 1.581 |
| **Full** | 0.415 | **0.786** | **0.867** | **0.739** | 1.413 |

Table 2: The effect of each feature on the effectiveness of the approach.

### 5.3 Preference Scaling

One of the key advantages of **DPA-OMF** is its ability to leverage expert demonstrations to automatically construct preference rankings without requiring additional human annotations, making it highly scalable. In this section, we evaluate the performance of **DPA-OMF** as we scale the number of preference rankings. In Figure 6(left), we show the relationship between the aligned model's performance and the average number of preference rankings per training example (e.g., a value of 4 indicates that the dataset used for alignment is four times larger than the original training set). Interestingly, we observe that a small amount of preference feedback not only fails to improve the model but actually degrades it. However, as the number of preference rankings increases, the alignment objective begins to demonstrate its effectiveness. We hypothesize that this degradation is due to preference over-optimization, which we explore further in the next section.

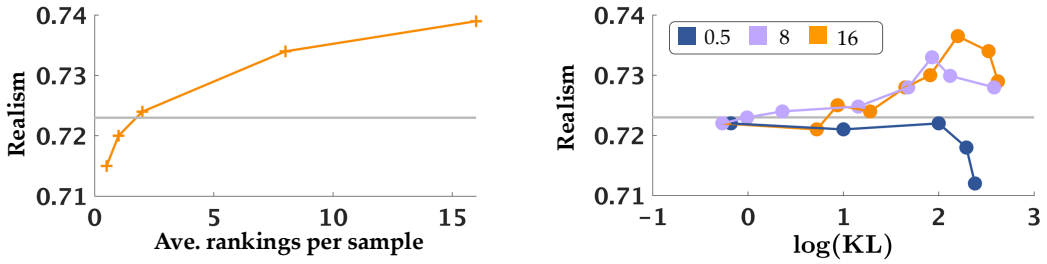


Figure 6: **[left - preference scaling]**: Performance of the alignment under different preference training data sizes. The gray line represents the performance of the reference model; **[right - preference over-optimization]**: The trade-off between the policy drift and the fine-tuning performance gain under various preference data sizes

.

### 5.4 Preference over Exploitation

Preference over-optimization has been studied in the context of LLMs for both online methods (e.g., RLHF) and direct preference alignment methods (Tang et al., 2024; Rafailov et al., 2024a). In this section, we investigate this phenomenon in the context of multi-agent motion generation. While previous research has focused on the impact of preference over-optimization across different model sizes, our study examines its effects with varying data sizes.

Goodhart's law states that "when a measure becomes a target, it ceases to be a good measure" (Goodhart & Goodhart, 1984). In our case, this effect manifests when there is insufficient feedback, causing the model to over-optimize an incomplete preference signal. Following (Tang et al., 2024), we examine this effect by measuring the KL divergence between the reference model and the fine-tuned model at various preference data sizes during the alignment process. KL divergence quantifies how much the optimized policy deviates from the reference model's policy during preference learning, and can be interpreted as the optimization cost incurred by the alignment. In Figure 6 (right), with little preference data (e.g., 0.5 rankings per training example), the optimized policy drifts away from the reference but degrades performance. As we scale the preference data, the same policy drift budget results in better performance, though further increases in the KL budget eventually reduce performance. Nevertheless, this investigation shows that scaling preference data can mitigate the effects of preference over-optimization and highlights the importance of doing so in a cost-effective manner.

## 6 CONCLUSION

In this work, we consider the problem of efficient post-training alignment of a token-prediction model for multi-agent motion generation. We propose Direct Preference Alignment from Occupancy Measure Matching Feedback, a simple yet principled approach that leverages expert demonstrations to generate implicit preference feedback and improves the pre-trained model's generation quality without additional human preference annotations, reward learning, or complex reinforcement learning. We presented the first investigation of direct preference alignment for multi-agent motion token-prediction models using implicit preference feedback from demonstrations. We applied our approach to large-scale traffic simulation and demonstrated its effectiveness in improving the realism of generated behaviors involving up to 128 agents, making a 1M token-prediction model comparable to state-of-the-art large models by relying solely on implicit feedback from demonstrations, without requiring additional human annotations or incurring high computational costs. Additionally, we provided an in-depth analysis of preference data scaling laws and their effects on over-optimization, offering valuable insights for future investigations.

**Reproducibility Statement**. To enhance the reproducibility of our research, we have provided a detailed explanation of our motion generation model and its source in Appendix C. Additionally, we have thoroughly described the calculation of preference distance in Section 3.2 and the process of constructing our preference dataset in Section 4.

## REFERENCES


Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 1, 2004.

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.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Yulong Cao, Boris Ivanovic, Chaowei Xiao, and Marco Pavone. Reinforcement learning with human feedback for realistic traffic simulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14428–14434. IEEE, 2024.

Zhaorun Chen, Siyue Wang, Zhuokai Zhao, Chaoli Mao, Yiyang Zhou, Jiayu He, and Albert Sibo Hu. Escirl: Evolving self-contrastive irl for trajectory prediction in autonomous driving. In *8th Annual Conference on Robot Learning*, 2024a.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024b.

Charles AE Goodhart and CAE Goodhart. *Problems of monetary management: the UK experience*. Springer, 1984.

Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, Scott Niekum, W Bradley Knox, and Dorsa Sadigh. Contrastive prefence learning: Learning from human feedback without rl. *arXiv preprint arXiv:2310.13639*, 2023.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

Siddharth Karamcheti, Suraj Nair, Annie S Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. *arXiv preprint arXiv:2302.12766*, 2023.

Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. On convergence and stability of gans. *arXiv preprint arXiv:1705.07215*, 2017.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. In *Forty-first International Conference on Machine Learning*.

Jiaxiang Li, Siliang Zeng, Hoi-To Wai, Chenliang Li, Alfredo Garcia, and Mingyi Hong. Getting more juice out of the sft data: Reward learning from human demonstration improves sft for llm alignment. *arXiv preprint arXiv:2405.17888*, 2024.

Aiwei Liu, Haoping Bai, Zhiyun Lu, Xiang Kong, Simon Wang, Jiulong Shan, Meng Cao, and Lijie Wen. Direct large language model alignment through self-rewarding contrastive prompt distillation. *arXiv preprint arXiv:2402.11907*, 2024.

Yicheng Luo, zhengyao jiang, Samuel Cohen, Edward Grefenstette, and Marc Peter Deisenroth. Optimal transport for offline imitation learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MhuFzFsrfvH.

Nico Montali, John Lambert, Paul Mougin, Alex Kuefler, Nicholas Rhinehart, Michelle Li, Cole Gulino, Tristan Emrich, Zoey Yang, Shimon Whiteson, et al. The waymo open sim agents challenge. *Advances in Neural Information Processing Systems*, 36, 2024.

Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. Rule based rewards for language model safety.

Nigamaa Nayakanti, Rami Al-Rfou, Aurick Zhou, Kratarth Goel, Khaled S Refaat, and Benjamin Sapp. Wayformer: Motion forecasting via simple & efficient attention networks. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2980–2987. IEEE, 2023.

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.

Jonah Philion, Xue Bin Peng, and Sanja Fidler. Trajeglish: Traffic modeling as next-token prediction. In *The Twelfth International Conference on Learning Representations*, 2024.

Ilija Radosavovic, Bike Zhang, Baifeng Shi, Jathushan Rajasegaran, Sarthak Kamat, Trevor Darrell, Koushil Sreenath, and Jitendra Malik. Humanoid locomotion as next token prediction. *arXiv preprint arXiv:2402.19469*, 2024.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. *arXiv preprint arXiv:2406.02900*, 2024a.

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, 2024b.

Ari Seff, Brian Cera, Dian Chen, Mason Ng, Aurick Zhou, Nigamaa Nayakanti, Khaled S Refaat, Rami Al-Rfou, and Benjamin Sapp. Motionlm: Multi-agent motion forecasting as language modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8579–8590, 2023.

Jiaming Song, Hongyu Ren, Dorsa Sadigh, and Stefano Ermon. Multi-agent generative adversarial imitation learning. *Advances in neural information processing systems*, 31, 2018.

Hao Sun and Mihaela van der Schaar. Inverse-rlignment: Inverse reinforcement learning from demonstrations for llm alignment. *arXiv preprint arXiv:2405.15624*, 2024.

Liting Sun, Wei Zhan, and Masayoshi Tomizuka. Probabilistic prediction of interactive driving behavior via hierarchical inverse reinforcement learning. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 2111–2117. IEEE, 2018.

Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, Rémi Munos, Bernardo Ávila Pires, Michal Valko, Yong Cheng, et al. Understanding the performance gap between online and offline alignment algorithms. *arXiv preprint arXiv:2405.08448*, 2024.

Ran Tian, Chenfeng Xu, Masayoshi Tomizuka, Jitendra Malik, and Andrea Bajcsy. What matters to you? towards visual representation alignment for robot learning. *arXiv preprint arXiv:2310.07932*, 2023.

Thomas Tian, Boyi Li, Xinshuo Weng, Yuxiao Chen, Edward Schmerling, Yue Wang, Boris Ivanovic, and Marco Pavone. Tokenize the world into object-level knowledge to address long-tail events in autonomous driving. In *8th Annual Conference on Robot Learning*, 2024.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints. In *The International Conference on Learning Representations*, 2024a.

Yuting Wang, Lu Liu, Maonan Wang, and Xi Xiong. Reinforcement learning from human feedback for lane changing of autonomous vehicles in mixed traffic. *arXiv preprint arXiv:2408.04447*, 2024b.

Wei Wu, Xiaoxin Feng, Ziyan Gao, and Yuheng Kan. Smart: Scalable multi-agent real-time simulation via next-token prediction. *arXiv preprint arXiv:2405.15677*, 2024.

Huang Xiao, Michael Herman, Joerg Wagner, Sebastian Ziesche, Jalal Etesami, and Thai Hong Linh. Wasserstein adversarial imitation learning. *arXiv preprint arXiv:1906.08113*, 2019.

Ceyuan Yang, Yujun Shen, Yinghao Xu, Deli Zhao, Bo Dai, and Bolei Zhou. Improving gans with a dynamic discriminator. *Advances in Neural Information Processing Systems*, 35:15093–15104, 2022.

Kevin Yang, Dan Klein, Asli Celikyilmaz, Nanyun Peng, and Yuandong Tian. Rlcd: Reinforcement learning from contrast distillation for language model alignment. *arXiv preprint arXiv:2307.12950*, 2023.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. *International Conference on Learning Representations*, 2020.

Zikang Zhou, Haibo Hu, Xinhong Chen, Jianping Wang, Nan Guan, Kui Wu, Yung-Hui Li, Yu-Kai Huang, and Chun Jason Xue. Behaviorgpt: Smart agent simulation for autonomous driving with next-patch prediction. *arXiv preprint arXiv:2405.17372*, 2024.

# A MOTIVATING QUESTIONS

Inspired by the appendix of (Karamcheti et al., 2023), in this section, we list some motivating questions that may arise from reading the main paper.

***Q1.*** *The features used when computing the preference distance are manually designed, what are the benefits and limitations?*

The alignment between generated agents' behavior and expert demonstrations requires measuring the distance between their occupancy measures in a semantically meaningful feature space. In this work, we use manually designed features that are well-validated and widely used in the autonomous driving industry. This allows us to conduct controlled experiments and evaluate the performance of our approach using proven, effective features. Additionally, in industry settings, it is often necessary to align a pre-trained motion model to specific criteria during post-training, such as progress. Using semantically meaningful features enables efficient and effective alignment for such cases. While encoding the generated traffic simulation into a learnable feature space could potentially provide more information and improve the informativeness of the preference distance, it also risks introducing spurious correlations (Zhang et al., 2020) and requires additional tasks to train the encoder to extract valuable features. Recent work (Tian et al., 2023) proposes to use human input to explicitly calibrate the feature space learning process to reduce the risk of learning spurious correlations, but the proposed method is only validated in simple simulation settings. We are excited to explore this direction in future work to further enhance the performance of our approach.

***Q2.*** *The preference distance can better reflect the alignment between a generated motion and the expert demonstrations, can it be used to directly train the motion model end-to-end and provide better results?*

In this work, we focus specifically on LLM-type token-prediction models, as they are becoming the backbone of motion models in various embodied tasks. These auto-regressive models typically use a teacher forcing training scheme for efficiency, which is not naturally compatible with the preference distance. While it is possible to use preference distance as a loss signal to train the auto-regressive model, this approach introduces additional complexity in the training pipeline and significantly increases computational cost, as it requires solving the optimal transport problem at each step.

***Q3.*** *Why the performance is still worse than SOTA methods even with preference alignment?*

In Table 1, while the aligned model underperforms compared to some state-of-the-art (SOTA) models, we believe this is primarily due to the architecture and capacity limitations of the reference model. We anticipate that applying our approach to larger SOTA models would result in significant performance improvements, as our method provides more nuanced alignment and could fully leverage the capabilities of more advanced architectures.

***Q4.*** *What makes the paper different from previous works that use optimal transport based reward for robot behavior learning via RL?*

The optimal transport method has been used to generate reward signals by measuring the distance between a rollout and expert demonstrations in single-agent RL settings (Xiao et al., 2019; Luo et al., 2023; Tian et al., 2023). In contrast, our work focuses on post-training alignment of multi-agent motion token-prediction models using expert demonstrations. To overcome the overly conservative assumption in previous works, which treat all model-generated samples as unpreferred, we leverage optimal transport to construct preference rankings among sampled rollouts. While it is feasible to convert the preference distance into per-step rewards and align the model using RLHF, this approach would require significantly more computational resources, as optimal transport must be solved in a multi-agent setting at every RL step. We are excited to explore the compute-performance trade-off between using preference distance in direct alignment versus RLHF in future work.

***Q5.*** *Why is using expert demonstrations as the preferred samples in preference learning less informative compared to constructing preference data using the generated samples?*

The expert demonstrations are first used to fine-tune the model with (1), thus the likelihood of expert demonstrations are already much higher than the sampled generations from the model. If using expert demonstrations as preferred samples and all the sampled generations as unpreferred

samples in preference alignment, the contrastive loss can be improved but this will further suppress the likelihood of the samples overall.

## B Extended Related Works

**Motion generation as next-token prediction.** Motion generation has traditionally been approached using one-shot prediction techniques, where the entire motion sequence is generated in a single forward pass conditioned on scene information (Nayakanti et al., 2023). However, these approaches often struggle with modeling long-term dependencies and maintaining temporal coherence between actions. Next-token prediction framework — where each subsequent token is predicted based on the previously generated ones — has proven highly successful in language modeling (Achiam et al., 2023). Similar to language, human or robotic motion unfolds as a series of continuous actions over time, with each movement serving as a "token" that depends on prior movements. Next-token prediction provides a natural autoregressive framework for generating these sequential actions, making it a powerful tool for motion generation in domains such as autonomous driving (Seff et al., 2023; Philion et al., 2024), robot manipulation (Brohan et al., 2023), and humanoid locomotion (Radosavovic et al., 2024). In our work, we explore the use of next-token prediction for multi-agent motion generation. However, unlike previous approaches, we focus on aligning a pre-trained motion generation model with human preference.

**Alignment of Large Language Models using preference feedback.** Next-token prediction optimizes for local coherence between individual tokens but often lacks long-term consistency between the generated sequence and the ground truth. This misalignment between the training objective and the human internal reward function, which governs their behavior, can lead to suboptimal outcomes and even safety-critical motions in embodied settings. To address this misalignment, both online and offline methods have been developed to better align large language models (LLMs) with human values using preference feedback. One prominent online approach is Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022), which fine-tunes the model by first learning a reward model and then using reinforcement learning to optimize the generative model's behavior to maximize the learned reward. However, this two-phase approach introduces significant complexity to the alignment process and incurs substantial computational costs. As an alternative, offline methods—often referred to as direct alignment methods (Rafailov et al., 2024b)—bypass the reward learning step by directly optimizing the model to maximize the margin between the log-likelihood ratio of preferred samples and that of unpreferred ones. While offline methods have demonstrated empirical efficiency and are often more favorable compared to online methods (Tunstall et al., 2023), collecting preference feedback remains time-consuming and difficult to scale. This challenge is especially pronounced in embodied settings, where human annotators must analyze intricate and nuanced multi-agent motions, making the process even more labor-intensive.

## C Motion generation model

We follow the MotionLM architecture to implement our multi-agent motion generation model (Seff et al., 2023).

Our model utilizes an early fusion network as the scene encoder to encode multi-modal scene inputs. The scene encoder integrates multiple input modalities, including the road graph, traffic light states, and the trajectory history of surrounding agents. These inputs are first projected into a common latent space through modality-specific encoders. The resulting latent embeddings for each modality are then augmented with learnable positional encodings to preserve spatial and temporal relationships. The augmented embeddings are concatenated and passed through a self-attention encoder, which generates a scene embedding for each modeled agent. These scene embeddings are subsequently used by the autoregressive model, via cross-attention, to predict the actions of each agent.. An agent's action token is obtained via discritizing acceleration control into a finite number of bins (169) and a joint action token denotes the collection of all agents' action token in the scene. Please refer to Section App.A of (Seff et al., 2023) about the bins used for tokenization.

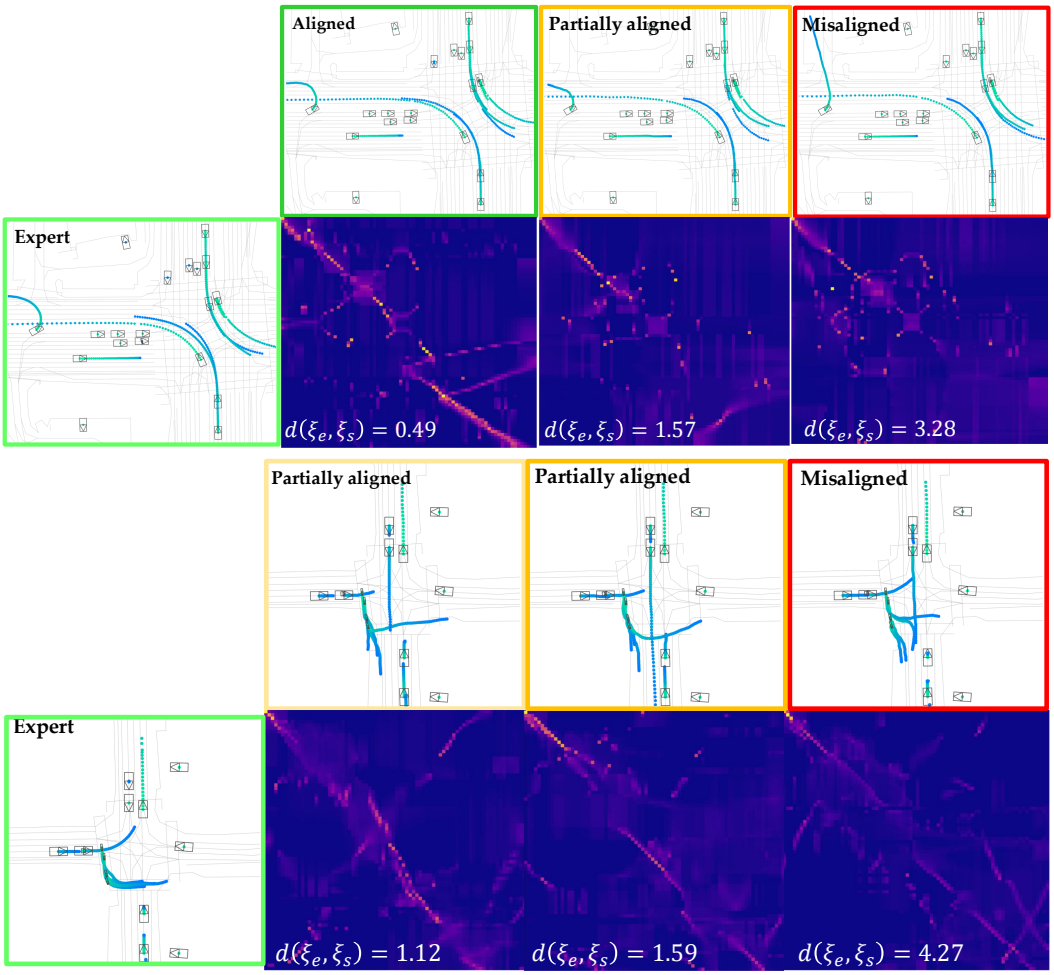


Figure 7: **Alignment visualization.** The heat map visualizes the optimal coupling between a generated traffic simulation and the ground truth scene evolution. More peaks along the diagonal indicate better alignment between the behaviors (i.e., a smaller preference distance).

## D QUALITATIVE EXAMPLES OF PREFERENCE DISTANCE

In Figure 7, we show qualitative examples that illustrates how the coupling matrix reflects the behavior alignment between generated traffic simulations and the expert demonstrations. We see that traffic simulations with smaller preference distance demonstrate more aligned behavior compared to the expert demonstration.

## E QUALITATIVE EXAMPLES DEMONSTRATING THE EFFECTIVENESS OF OUR APPROACH

In Figure 8, we include more visualizations to demonstrate the performance of our approach. For each model, we sample 64 rollouts from the model and we show the most-likely one.

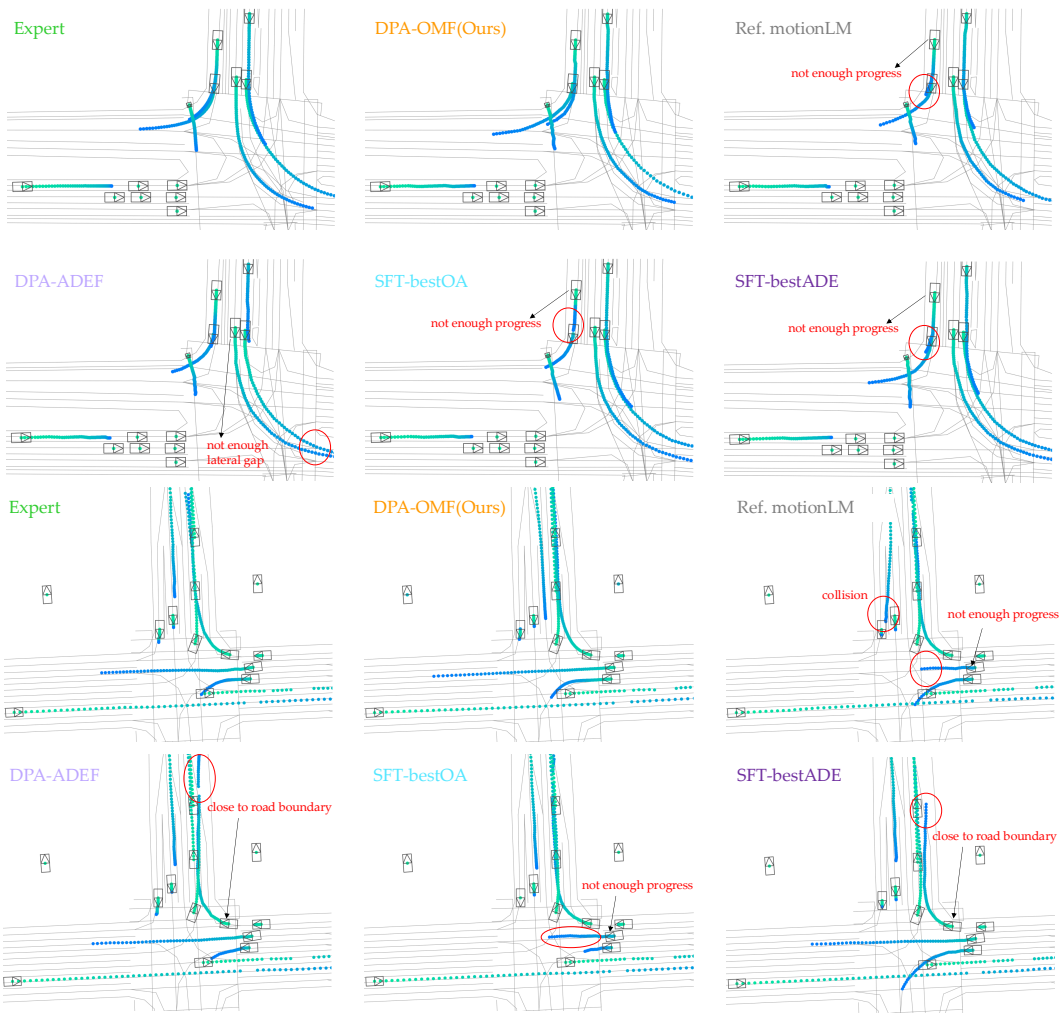


Figure 8: **Traffic simulation generation visualization.** Our approach produces traffic simulations that are more closely aligned with expert demonstrations, while baseline models generate simulations that are only partially aligned or misaligned.

## F Additional result - Comparison between DPA-OMF and adversarial preference alignment from demonstrations

In this section, we compare our method with the AFD approach, which treats all samples from the reference model as negative samples. For each training sample, we construct 16 rankings by sampling 16 generated traffic simulations from the reference model (i.e., both our method and AFD utilize the same amount of preference data). We measure the WOSAC realism of the fine-tuned model, the model's ability to assign higher likelihood to preferred traffic simulations ranked by our preference distance (measured as classification accuracy), and the minADE. As shown in Table 3, our approach significantly outperforms the adversarial AFD in all metrics, demonstrating the effectiveness of our method.

| Features | Classification accuracy ↑ | Composite realism ↑ | minADE ↓ |
|---|---|---|---|
| **Ours** | **0.84** | **0.739** | **1.413** |
| **Adversial AFD** | 0.52 | 0.720 | 1.539 |

Table 3: The comparison between DPA-OMF with adversarial AFD. Our approach significantly outperforms the adversarial AFD in all metrics.

To further analyze why adversarial preference alignment is less effective, we plot the negative log-likelihood of expert demonstrations, preferred traffic simulations, and unpreferred traffic simulations in Figure 9.

The results reveal that the likelihood of expert demonstrations is consistently much higher than that of both preferred and unpreferred samples throughout the alignment process. This stems from the pre-training phase, where expert demonstrations are used to train the reference model. Moreover, during the preference alignment phase, the model primarily increases the likelihood of expert demonstrations while leaving the likelihood of preferred and unpreferred samples relatively unchanged. This indicates that the model is unable to capture nuanced differences between preferred and unpreferred samples, leading to suboptimal alignment performance.

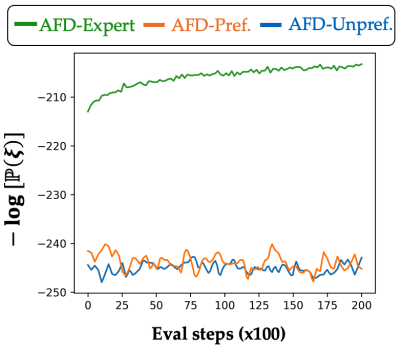


Figure 9: The log-likelihood of expert demos and preferred/unpreferred rollouts from the reference model when AFD for alignment.

## G Additional result - Bias introduced by the heterogeneity of the preference data

In the previous section, we show that using expert demonstrations as preferred samples and model generations as unpreferred samples results in increasing the likelihood of expert demonstrations without significantly affecting the likelihood of either preferred or unpreferred generated samples. This suggests that the model struggles to associate the features that make expert demonstrations preferred with the generated preferred samples. To further explore this, we conducted a separate experiment demonstrating how a discriminative objective using expert demonstrations as positive samples and model generations as negative samples can lead to spurious correlations.

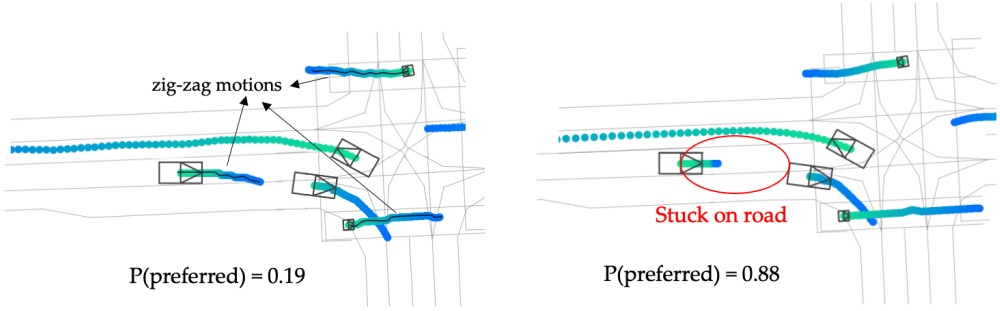


Figure 10: **Spurious correlation introduced by the heterogeneity of the preference data.** The model relies heavily on trajectory smoothness to differentiate between expert demonstrations and model generations, which can lead to incorrect predictions about preferred behaviors. For example, in the traffic simulation on the left, the trajectories exhibit zig-zag patterns but demonstrate more human-like behaviors compared to the simulation on the right. However, the model incorrectly predicts the likelihood of the left humam-like simulation being preferred by humans as only 0.19 and predicts the likelihood of the right unhumam-like simulation being preferred by humans as 0.88, highlighting its inability to fully capture nuanced human preferences.

In this experiment, we trained a discriminator using a contrastive objective to distinguish between expert demonstrations and model generations. The discriminator achieved a classification accuracy of 0.83 on the evaluation dataset, indicating it can reasonably classify motions as either expert demonstrations or model generations. When the discriminator was used to rank pairs of model-generated motions, we observed a pattern: motions with zig-zag trajectories are often classified as unpreferred, while relatively smooth motions are classified as preferred, even when there is unhuman-like behaviors (e.g., stuck on roads as shown in Figure 10).

This behavior arises because of the heterogeneity of the two data sources: most human demonstrations exhibit smooth motions, while model generations are not constrained by vehicle dynamics. Consequently, the contrastive objective may incentivize the model to pick up this spurious correlation, prioritizing smoothness over other critical attributes such as staying on the road.

## H ADDITIONAL RESULT - THE COST OF QUERYING HUMANS FOR PREFERENCES IN MULTI-AGENT TRAFFIC GENERATIONS

To quantify the human cost associated with providing preference rankings for multi-agent traffic simulations, we conducted an Institutional Review Board (IRB)-approved human subject study to measure the effort required. In this study, we presented paired traffic simulations to participants and asked them to rank the pairs based on how realistic the simulations were compared to their personal driving experience. We varied the number of traffic agents in the simulations and recorded the time needed to provide rankings. Five participants ranked 500 pairs of traffic simulations, and Table 4 summarizes the time required to complete this task. The results show a clear trend: as the number of traffic agents increases, the time required for human annotators to rank simulations grows significantly.

| **Num. of agents in the scene** | 1 | 10 | 20 | 40 | 80 |
|---|---|---|---|---|---|
| **Average time used for ranking [s]** | 0.7 | 4.9 | 9.8 | 29.4 | 42.1 |

Table 4: Average time required for a human to rank traffic sumulations.

Although this study was conducted under time constraints and is not exhaustive, it provides an useful estimate of the human cost for constructing preference rankings at scale. Specifically, for the preference data used in our experiments, the estimated average time required for one human annotator is approximately **633** days. This result underscores the practical challenges of scaling preference ranking annotations in multi-agent scenarios, motivating our approach to leverage existing demonstrations to construct preference rankings efficiently.