

---

# DriveGPT: Scaling Autoregressive Behavior Models for Driving

---

Xin Huang<sup>1</sup> Eric M. Wolff<sup>1</sup> Paul Vernaza<sup>1</sup> Tung Phan-Minh<sup>1</sup> Hongge Chen<sup>1</sup> David S. Hayden<sup>1</sup>  
Mark Edmonds<sup>1</sup> Brian Pierce<sup>1</sup> Xinxin Chen<sup>1</sup> Pratik Elias Jacob<sup>1</sup> Xiaobai Chen<sup>1</sup> Chingiz Tairbekov<sup>1</sup>  
Pratik Agarwal<sup>1</sup> Tianshi Gao<sup>1</sup> Yuning Chai<sup>2</sup> Siddhartha S. Srinivasa<sup>1</sup>

## Abstract

We present DriveGPT, a scalable behavior model for autonomous driving. We model driving as a sequential decision-making task, and learn a transformer model to predict future agent states as tokens in an autoregressive fashion. We scale up our model parameters and training data by multiple orders of magnitude, enabling us to explore the scaling properties in terms of dataset size, model parameters, and compute. We evaluate DriveGPT across different scales in a planning task, through both quantitative metrics and qualitative examples, including closed-loop driving in complex real-world scenarios. In a separate prediction task, DriveGPT outperforms state-of-the-art baselines and exhibits improved performance by pretraining on a large-scale dataset, further validating the benefits of data scaling.

## 1. Introduction

Transformer-based foundation models have become increasingly prevalent in sequential modeling tasks across various machine learning domains. These models are highly effective in handling sequential data by capturing long-range dependencies and temporal relationships. Their success has been evident in natural language processing (Mann et al., 2020; Kaplan et al., 2020; Hoffmann et al., 2022), time-series forecasting (Zhou et al., 2021), and speech recognition (Kim et al., 2022), where sequential patterns play a crucial role. One of the key strengths of transformer-based models is their capacity to learn from large datasets including millions of training examples, enabling them to address complex tasks with increased model sizes, up to billions of model parameters.

---

<sup>1</sup>Cruise LLC, San Francisco, CA <sup>2</sup>Meta, Menlo Park, CA. Correspondence to: Xin Huang, Eric M. Wolff <{cyrus.huang, eric.wolff}@getcruise.com>.

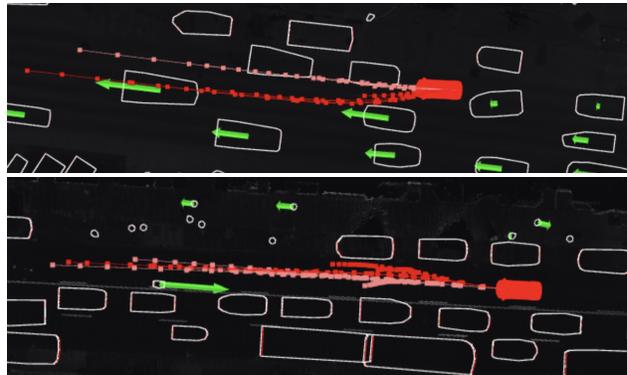


Figure 1. Through data and model scaling, DriveGPT (red) handles complex real-world driving scenarios, such as lane changing in heavy traffic and yielding to a cyclist in the opposite lane, compared to a smaller baseline trained on less data (pink).

Model	Parameters	Training Segments
VectorNet	72K	211K
PlanTF	2.1M	1M
QCNet-Argo2	7.3M	200K
QCNet-WOMD	7.5M	2.2M
Wayformer	20M	2.2M
MotionLM	27M	1.1M
MTR	66M	2.2M
GUMP	523M	2.6M
<b>DriveGPT (Ours)</b>	<b>1.4B</b>	<b>120M</b>

Table 1. DriveGPT is  $\sim 3x$  larger and is trained on  $\sim 50x$  more data sequences than existing published behavior models.

While scaling up model and dataset sizes has been critical for recent advances in sequential modeling for text prediction (Kaplan et al., 2020; Hoffmann et al., 2022), it remains unclear whether these scaling trends can be directly extended to behavior modeling, particularly in driving tasks, due to several unique challenges. First, driving tasks involve a wider range of input modalities, including agent trajectories and map information, unlike language tasks that rely solely on textual inputs. Second, behavior modeling demands spatial reasoning and an understanding of physical kinematics. Such capabilities are typically beyond the scope of language models. Finally, the collection of large-scale

driving datasets requires substantial effort and resources, making it far more challenging than gathering textual data. As a result, existing work is often constrained by the availability of training data or the scalability of the models, as summarized in Table 1.

In our work, we present a comprehensive study of scaling up data sizes and model parameters in the context of behavior modeling for autonomous driving, which predicts future actions of traffic agents to support critical tasks such as planning and motion prediction. Specifically, we train a transformer-based autoregressive behavior model on over 100 million high-quality human driving examples,  $\sim 50$  times more than existing open-source datasets, and scale the model over 1 billion parameters, outsize existing published behavior models.

As we scale up the volume of training data and the number of model parameters, we observe improvements in both quantitative metrics and qualitative behaviors. More importantly, large models trained on extensive, diverse datasets can better handle rare or edge-case scenarios, which often pose significant challenges for autonomous vehicles, as shown in Fig. 1<sup>1</sup>. As a result, we see great potential in scaling up behavior models through data and model parameters to improve the safety and robustness of autonomous driving systems.

Our main contributions are as follows:

1. We present DriveGPT, a large autoregressive behavior model for driving, by scaling up both model parameters and real-world training data samples.
2. We determine empirical driving scaling laws for an autoregressive behavior model in terms of data size, model parameters, and compute. We validate the value of scaling up training data and compute, and observe better model scalability as training data increases, consistent with the language scaling literature.
3. We quantitatively and qualitatively compare models from our scaling experiments to validate their effectiveness in real-world driving scenarios. We present real-world deployment of our model through closed-loop driving in challenging conditions.
4. We demonstrate the generalizability of our model on the Waymo Open Motion Dataset, which outperforms prior state-of-the-art on the motion prediction task and achieves improved performance through large-scale pretraining.

<sup>1</sup>The baseline model is trained on  $\sim 50$  times less data and uses  $\sim 50$  times fewer parameters.

## 2. Related Work

### 2.1. Behavior Modeling

Behavior modeling is a critical task in autonomous driving, which covers a broad spectrum of tasks including planning, prediction, and simulation. Taking multimodal inputs including agent history states and map information, behavior models predict the future states of these traffic agents by reasoning about agent dynamics (Cui et al., 2020; Song et al., 2022), interactions (Sun et al., 2022; Jiang et al., 2023), human intent (Shi et al., 2022; Huang et al., 2022; Sun et al., 2024), and driving environments (Liang et al., 2020; Kim et al., 2021).

Among learning-based models, transformers have gained popularity due to their ability to fuse multimodal inputs as encoders (Nayakanti et al., 2023; Zhang et al., 2023; Zhou et al., 2023; Jia et al., 2023; Gan et al., 2024) and model long-range temporal relationships as decoders (Seff et al., 2023; Shi et al., 2022; 2024). Despite the success of transformers in behavior modeling, existing literature is often restricted by the size of model parameters due to limited training data, which fails to capture the full scaling potential of transformer-based models. In our work, we scale up our transformer models to include billions of parameters, by training on a large-scale dataset including more than 100M driving demonstrations, and validate the scalability of transformer-based models in the context of autonomous driving.

### 2.2. Large Transformer Models

Large transformers have demonstrated great success in sequential modeling tasks, by scaling up model parameters and data sizes (Kaplan et al., 2020; Hoffmann et al., 2022; Zhai et al., 2022; Muennighoff et al., 2023). These scaling laws have pushed the boundary of many sequential modeling tasks including natural language processing (Mann et al., 2020), time-series forecasting (Zhou et al., 2021), and speech recognition (Kim et al., 2022).

Recent work has studied the scalability of behavior models in the context of motion prediction (Seff et al., 2023; Ettinger et al., 2024), planning (Sun et al., 2023), and simulation (Hu et al., 2024b), yet these studies are either constrained by limited data size (up to a couple of million training examples), or focused on a few orders of magnitude in terms of data and model scaling, limiting the potential to draw statistically significant conclusions over a large scaling range.

In this paper, we study the scaling properties (in terms of data samples, model parameters, and compute) across a much larger range compared to prior work. More specifically, we target an autoregressive decoder architecture that has been proven to be both scalable (as in the LLM liter-

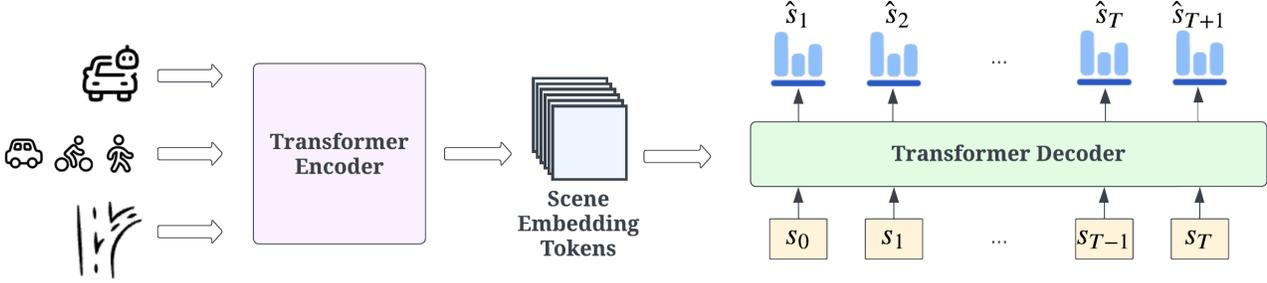


Figure 2. DriveGPT architecture, including a transformer encoder and a transformer decoder. The transformer encoder summarizes relevant scene context, such as target agent history, nearby agent history, and map information, into a set of scene embedding tokens. The transformer decoder follows an LLM-style architecture that takes a sequence of agent states as input and predicts a discrete distribution of actions at the next tick, conditioned on previous states.

ature) and effective in generating accurate trajectories for traffic agents (Seff et al., 2023; Chen et al., 2024; Hu et al., 2024a).

Beyond autonomous driving, there is limited relevant literature on building large transformer models for robotic tasks (O’Neill et al., 2024; Octo Model Team et al., 2024), which share a transformer architecture similar to our work. Robotics models often share different input features and dynamic models, and operate in different environments, making them difficult to apply directly to driving tasks.

### 3. Behavior Model

We use a standard encoder-decoder architecture as our behavior model, as shown in Fig. 2. We use transformer-based models as our encoder and decoder backbones due to their scalability in related sequential modeling tasks.

#### 3.1. Problem Formulation

We model the problem as a sequential prediction task over the future positions of the target agent up to horizon  $T$ , by applying the chain rule at each step, conditioning on driving context information  $\mathbf{c}$  and historical agent positions  $\mathbf{s}$ :

$$P(s_{1:T}|\mathbf{c}) = \prod_{t=1}^T P(s_t | s_{0:t-1}, \mathbf{c}). \quad (1)$$

The context information includes target agent history states  $c_{\text{target}}$ , nearby agent history states  $c_{\text{nearby}}$ , and map states  $c_{\text{map}}$ . The historical agent information includes agent positions from previous historical steps, i.e.  $s_{0:t-1}$  if we want to predict agent positions at step  $t$ .

We define “state” as a complete kinematic state including position, orientation, velocity, and acceleration, which is commonly available in agent history observations, and “position” as 2-D  $(x, y)$  coordinates to simplify the output space.

#### 3.2. Scene Encoder

The encoder follows a standard transformer encoder architecture (Shi et al., 2022) that fuses all input modalities into a set of scene embedding tokens. It consumes raw input features, including target agent history states, nearby agent history states, and map states as a set of vectors, and normalizes all inputs to an agent-centric view. Each vector is mapped to a token embedding through a PointNet-like encoder as in (Gao et al., 2020). At the end of the encoder, we apply a self-attention transformer (Vaswani et al., 2017) to fuse all input context into a set of encoder embeddings,  $\mathbf{c} \in \mathbb{R}^{n \times d}$ , where  $n$  is the number of vectors and  $d$  is the token dimensions, that summarize the driving scene.

#### 3.3. LLM-Style Trajectory Decoder

Inspired by the LLM literature (Radford et al., 2019), we follow (Seff et al., 2023) to use a transformer decoder architecture to predict the distribution of agent positions at each step in the future.

The decoder first tokenizes agent positions at all steps into embeddings with dimension  $d$  through a linear layer, followed by a LayerNorm layer and a ReLU layer. At each step  $t$ , the decoder takes agent embeddings up to  $t$ , and cross attends them with the encoder embeddings  $\mathbf{c}$  to predict the distribution of agent positions at the next step  $t + 1$ .

The output is a set of discrete actions  $a$  represented as the Verlet action (Rhinehart et al., 2018; Seff et al., 2023), as the second derivative of positions. We can apply the following equation to map Verlet actions to positions:

$$s_{t+1} = s_t + (s_t - s_{t-1}) + a_t, \quad (2)$$

where  $a_t$  is the predicted Verlet action, and  $(s_t - s_{t-1})$  assumes a constant velocity step. This representation helps predict smooth trajectories using a small set of actions.

### 3.4. Training

To train a DriveGPT model, we follow teacher forcing by applying ground truth future positions as input to the trajectory decoder. This allows us to predict all future steps in parallel.

We use a single cross-entropy classification loss over the action space, where the target action is selected as the one that is closest to the ground truth future trajectories. We refer the reader to Appendix A for more training details.

### 3.5. Inference

At inference time, we follow a standard LLM setup and roll out a trajectory over horizon  $T$  autoregressively, by repeating the process of predicting the action distribution at the next step, sampling an action, and adding it back to the input sequence.

We sample multiple trajectories in batch to approximate the distribution and then subsample to the desired number of modes using  $K$ -Means, as in (Seff et al., 2023).

## 4. Scaling Experiments

The goal of our scaling experiments is to determine the effect of data and model size on behavior prediction performance. Quantifying scaling laws similar to those seen in LLMs (Kaplan et al., 2020) can help prioritize the value of data and compute for future research directions in behavior modeling. We focus our effort on exploring the next frontier of data and model size – over an order of magnitude beyond previously published work.

**Large-scale driving dataset** From millions of miles of high-quality real-world human driving demonstrations, we curate a small subset of 120M segments for an internal research dataset. The dataset is carefully curated and balanced to represent diverse geographic regions across multiple cities and countries, including the United States, Japan, and the UAE. Data collection is evenly distributed between daytime and nighttime and is conducted primarily in urban environments. The dataset captures a wide range of challenging driving scenarios, such as lane changes, intersections, double-parked vehicles, construction zones, and close interactions with pedestrians and cyclists.

We extracted map information, target agent states, and nearby agent states into vectorized representation, as customary in behavior modeling literature (Gao et al., 2020).

**Scaling overview** We scale the model size across three orders of magnitude, from 1.5 million to 1.4 billion parameters, by increasing the embedding dimension in both encoder and decoder transformers. For each model size, we explore

Model Size	Hidden Dimension ( $d_{\text{model}}$ )	Max LR
1.5M	64	0.0050
4M	128	0.0020
8M	192	0.0014
12M	256	0.0010
26M	384	0.0010
94M	768	0.0007
163M	1024	0.0004
358M	1536	0.0002
629M	2048	0.0001
1.4B	3072	0.0001

Table 2. We vary the model size over three orders of magnitude through hidden dimensions. For each model size, we report the optimal learning rate, which decreases as the model size increases, matching the observations in the LLM scaling literature (Kaplan et al., 2020; Hoffmann et al., 2022).

multiple learning rate schedules using a cosine decay over the full training steps and select the maximum learning rate that yields the best performance. Table. 2 summarizes the optimal learning rate for each model size. Consistent with practices in large language model scaling (Kaplan et al., 2020), each model is trained for a single epoch.

We evaluate model performance using validation loss, computed on a comprehensive validation set of 10 million samples drawn from the same distribution as the training data, with no overlap. This set remains fixed across all scaling experiments to ensure consistency. We use validation loss as a proxy to measure model performance, following standard practices in scaling studies (Kaplan et al., 2020; Hoffmann et al., 2022; Muennighoff et al., 2023). This loss, calculated as cross-entropy on next-action prediction, serves as our primary performance metric. Additional driving-specific metrics are reported in Sec. 5.1.

### 4.1. Data Scaling

Data scaling results are summarized in Fig. 3. The smallest dataset of 2.2M samples mimics the size of Waymo Open Motion Dataset (WOMD) (Ettinger et al., 2021), a large open-source dataset for behavior modeling ( $\sim 44k$  scenarios with multiple target agents per scenario). We select a few subsets of our internal research dataset to study data scaling across different orders of dataset sizes. Our experiments use  $\sim 50x$  more data than WOMD, exploring a new region of the design space.

The results indicate that as the model is trained on more unique data samples, the performance improves, regardless of model size. Extrapolating from the scaling law in Fig. 3, to improve the best loss by another 10%, we would need to include 350M more training examples. A 20% improvement

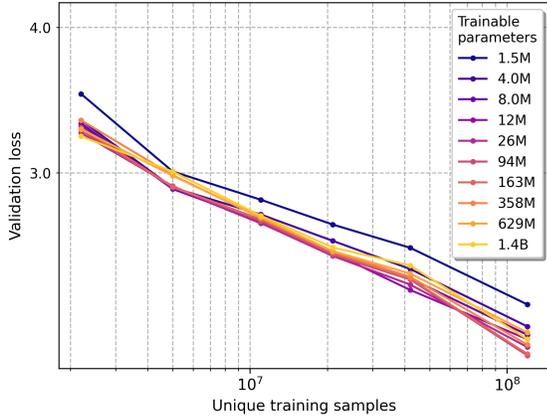


Figure 3. Performance increases with dataset size across a range of model parameters, indicating that data is a limiting factor. Both axes are on a log scale. An exponential fit was applied to all data points except for the 1.5M curve, resulting in the following relationship:  $\log(L) = -0.102 \log(D) + 2.663$  with an  $R^2$  value of 0.986, where  $L$  is the validation loss and  $D$  is the number of unique training samples.

would require about 1.4B more examples. As a result, we find that data remains the bottleneck for further improving driving performance.

Lastly, the scaling results remain relatively consistent across model sizes. This consistency indicates that data scaling comparisons can be done on reasonably small model sizes beyond 10M parameters.

### 4.2. Model Scaling

We now study model sizes across three orders of magnitude (1.5M to 1.4B parameters), as listed in Table 2. We increase model size by increasing the hidden dimensions of transformers for simplicity. We notice that modifying other parameters such as number of attention heads and hidden dimensions per head does not lead to noticeable changes in the results, as studied in Appendix B.

Training larger models is sensitive to learning rates, as observed in other scaling studies (Kaplan et al., 2020; Hoffmann et al., 2022). For each model size, we run multiple experiments at different learning rates to select the one with the optimal performance, as summarized in Table. 2.

Results in Fig. 4 demonstrate that increasing the amount of training data enhances the effectiveness of model scaling. Specifically, when the dataset size is up to 21M, the impact on the validation loss is barely noticeable across a large range of model sizes. Beyond 21M samples, validation losses improve with larger models – up to 12M parameters when trained on the 42M dataset and up to 94M parameters when trained on the 120M dataset – before reaching

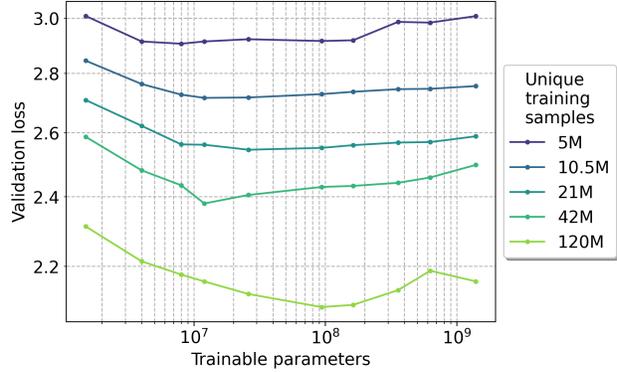


Figure 4. Model scaling is more effective as training data increases. The validation loss improves up to  $\sim 100M$  parameters when trained on the full dataset.

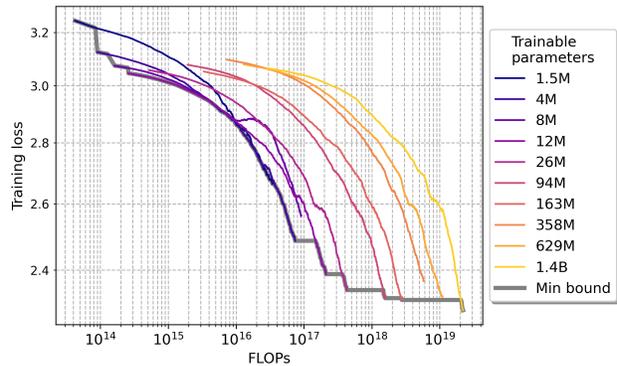


Figure 5. Relationship between (smoothed) training loss and FLOPs. Each curve represents an experiment corresponding to a specific model size, and the “min bound” indicates the best performance possible for a given FLOP budget.

a plateau and eventually overfitting. These findings further reinforce that data is the primary bottleneck for scaling models, aligning with observations in the LLM scaling literature (Kaplan et al., 2020). We leave further exploration of scaling behavior with larger datasets to future work.

### 4.3. Compute Scaling

In Fig. 5, we examine how compute affects training loss, where compute is measured by Floating Point Operations (FLOPs). We identify a monotonically decreasing “min-bound” boundary, which shows the lowest training loss observed up to the current compute value. As we increase compute, training loss generally decreases. Initially, this decrease is quite steep, but it gradually slows down at higher FLOPs values. This trend is consistent with observations in the LLM scaling literature, such as those reported by (Hoff-

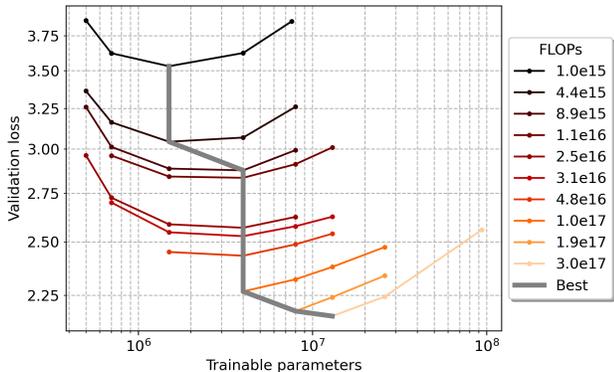


Figure 6. Performance as a function of model size for fixed compute budgets. The solid gray line connects the best result from each compute budget.

mann et al., 2022), covering a subset of the full FLOP range explored in these studies.

Next, we investigate whether there is an optimal combination of data size and model parameters for a fixed compute budget. Given the large computational expense for training models at the scales we are exploring, it is important to make the best use of our data. In this study, we fixed the compute budget in different FLOP groups. For a fixed compute budget, we can allocate resources either to model parameters or data samples, keeping their product constant.

Fig. 6 plots the performance with different compute budgets. The trend clearly shows that a larger compute budget leads to better performance, with optimal model size increasing accordingly, as indicated by the “best” gray line. The results further reveal that data is the main bottleneck, as the smallest model outperforms others in the three largest FLOP groups.

#### 4.4. Ablation Study on Decoder Architecture

We scale up two different model architectures by two orders of magnitude: our autoregressive decoder and a one-shot decoder. For the one-shot decoder, we follow (Nayakanti et al., 2023) to use a transformer decoder that takes a set of learned queries and cross attends them with scene embeddings to produce trajectory samples. This decoder is referred to as “one-shot” because it generates the full trajectory rollout at once, where an autoregressive decoder follows an LLM-style to produce trajectories one step at a time.

The results are summarized in Fig. 7, where we use minFDE at 6 seconds as a proxy to measure the model performance because of different loss definitions between two decoder architectures. Despite worse performance at small-scale parameters, our autoregressive decoder achieves better scalability and outperforms the one-shot baseline beyond 8M

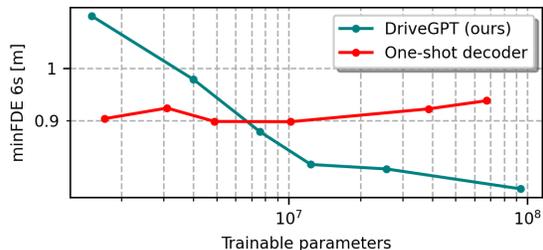


Figure 7. DriveGPT achieves better scalability in terms of minFDE, by adopting an autoregressive architecture compared to a one-shot architecture.

parameters. While we find it harder to scale the one-shot decoder, we confirm that our autoregressive decoder scales up to 100M parameters in terms of prediction accuracy, and defer further scalability study on the one-shot decoder as future work.

## 5. Planning and Prediction Experiments

In this section, we show detailed results of DriveGPT in a planning task using our internal research dataset and a motion prediction task using an external dataset. The results here further explore the impact of scaling from Section 4 and help ground those results in driving tasks and metrics.

### 5.1. Internal Evaluation: AV Planning

For the planning task, we train our model using our internal research dataset, composed of millions of high-quality human driving demonstrations, and generate AV trajectories by autoregressively predicting the next action, as described in Sec. 3.5.

We approximate the distribution by oversampling trajectories in batch and subsampling to 6 trajectories as in (Seff et al., 2023). While the AV must ultimately select a single trajectory for planning, returning multiple samples helps better understand multimodal behavior and aligns with motion prediction metrics.

We measure the planning performance on a comprehensive test set through a set of standard geometric metrics including minADE (mADE), minFDE (mFDE), and miss rate (MR). Additionally, we use semantic-based metrics including off-road rate (Offroad) that measures the ratio of trajectories that leave the road and collision rate (Collision) that measures the ratio of trajectories overlapping with traffic agents. We normalize these metrics across experiments to highlight relative performance changes.

#### 5.1.1. DATA SCALING RESULTS

We compare a DriveGPT model at 26M parameters trained on datasets of different sizes. The baseline dataset (2.2M) is selected to mimic the size of a typical behavior modeling

## DriveGPT: Scaling Autoregressive Behavior Models for Driving

Data	mADE ↓	mFDE ↓	MR ↓	Offroad ↓	Collision ↓
2.2M	1.000	1.000	1.000	1.000	1.000
21M	0.561	0.496	0.420	0.326	0.269
85M	0.496	0.441	0.332	0.238	0.217
120M	<b>0.489</b>	<b>0.433</b>	<b>0.317</b>	<b>0.198</b>	<b>0.196</b>

Table 3. As we scale up more training data, DriveGPT produces better AV trajectories as measured across all metrics. Metrics are normalized to highlight relative performance.

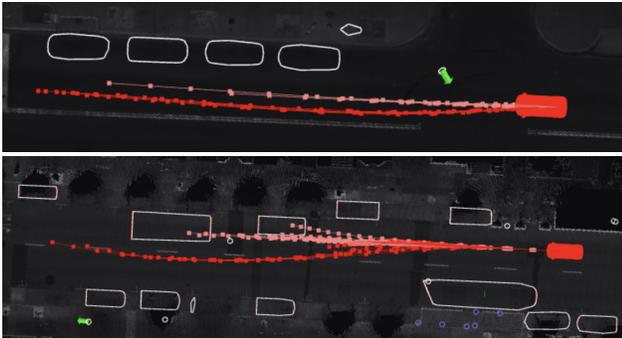


Figure 8. By training on 50x more data, DriveGPT (red) is able to produce high-quality trajectories in red that keep a safe lateral distance from a jaywalker (top) and go around two double-parked vehicles (bottom).

dataset such as WOMD.

The results are presented in Table 3, where we see that training on more data samples significantly improves the quality of the predicted AV trajectories, in terms of critical semantics metrics in driving including offroad rate and collision rate, as well as geometric metrics. These improvements are consistent with Sec. 4.1.

We further present two qualitative examples in Fig. 8 to illustrate the value of training on more data. In these examples, red trajectories represent DriveGPT trained on 120M samples, and pink trajectories are from the same model trained on 2.2M samples. The examples show that our method produces map-compliant and collision-free trajectories when trained on more data, successfully handling complicated interactions involving a jaywalking pedestrian and two double-parked vehicles.

### 5.1.2. MODEL SCALING RESULTS

We train four models using our 120M internal research dataset, and select the 8M model as the baseline. The baseline represents a reasonable size at which our model starts to outperform one-shot decoders, as shown in Sec. 4.4.

The results are presented in Table 4, where all metrics improve as the model size increases up to 94M parameters,

Model	mADE ↓	mFDE ↓	MR ↓	Offroad ↓	Collision ↓
8M	1.000	1.000	1.000	1.000	1.000
26M	0.954	0.950	0.902	0.858	0.915
94M	<b>0.937</b>	<b>0.925</b>	<b>0.866</b>	<b>0.815</b>	0.890
163M	0.943	<b>0.925</b>	0.875	<b>0.815</b>	<b>0.817</b>

Table 4. As we scale up more model parameters, all metrics improve up to 94M. Collision rate continues to improve at 163M. Metrics are normalized to highlight relative performance.

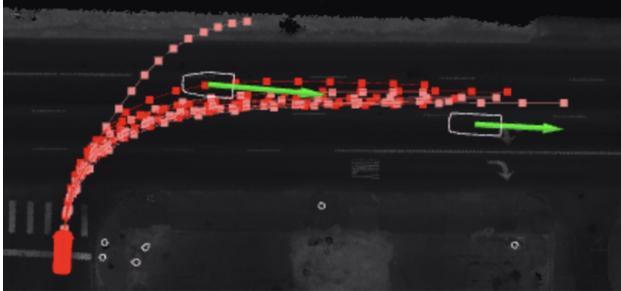


Figure 9. By training with 12x more parameters, DriveGPT (red) produces realistic trajectories that obey lane boundaries when performing a right turn into traffic.

with further gains in the collision metric at 163M parameters (see Sec. D.1 for qualitative examples). Although validation loss shows diminishing returns beyond 94M in Fig. 4, driving metrics continue to improve with increased model capacities, highlighting the potential benefits of introducing more parameters for enhanced driving performance.

We present a qualitative example in Fig. 9, where a larger DriveGPT including 94M parameters produces better trajectory samples in red that stay within the road boundary, compared to a smaller version including 8M parameters, in a right turn scenario.

### 5.1.3. CLOSED-LOOP DRIVING

We demonstrate the effectiveness of DriveGPT as a real-time motion planner deployed in a closed-loop setting. The model takes input features from an industry-level perception system that outputs agent states and map information. We used the 8M DriveGPT model, trained on the full dataset, to drive the car, achieving a latency of under 50ms on a single onboard GPU.

In Fig. 10, we present a challenging example in dense urban traffic where there are two double-parked vehicles blocking the path forward along with other oncoming vehicles. DriveGPT generates smooth and safe trajectories, bypassing the blocking vehicles and moving back to the original lane afterward. More examples are pre-

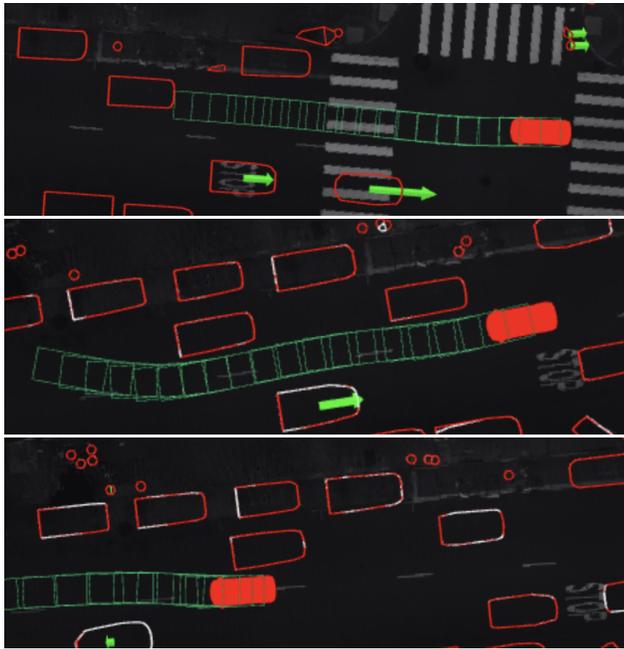


Figure 10. Example of DriveGPT running as a closed-loop planner in real time in dense urban traffic. It first produces a forward trajectory (top), updates it to bypass double-parked vehicles (middle), and drives back to the original lane at the end (bottom).

sented in the supplementary video at <https://youtu.be/-hLi44Pfy8g>, where DriveGPT alone is responsible for driving in real time.

## 5.2. External Evaluation: Motion Prediction

To directly compare with published results, we evaluate DriveGPT on the WOMD motion prediction task. Additionally, we explore the benefits of scale by pretraining on our internal research dataset and finetuning on the significantly smaller WOMD dataset.

### 5.2.1. OPEN-SOURCE ENCODER

For our external evaluation, we use the open-source MTR (Shi et al., 2022) encoder. This encoder is similar to the one described in Sec. 3.2. We make this change to improve the reproducibility of our results and take advantage of MTR’s open-source dataloading code for WOMD. We use the same autoregressive decoder as described in Sec. 3.3.

### 5.2.2. PRETRAINING SETUP

We made a couple of minor modifications to DriveGPT to be compatible with the WOMD dataset. First, we modify our map data to include the same semantics as in WOMD. Second, we modify our agent data to include the same kinematic

features for traffic agents as in WOMD.

We pretrain DriveGPT by training on our internal research dataset for one epoch (as in Sec. 4). We load the pretrained checkpoint and finetune the model using the same training setup as in the MTR codebase, where we train the model for 30 epochs using a weighted decay learning rate scheduler.

### 5.2.3. RESULTS

We measure model performance via a set of standard WOMD metrics, including minADE, minFDE, miss rate, and soft mAP. Each metric is measured on the test set and computed over three different time horizons.

We present two variants of our method to validate its effectiveness on the motion prediction task, including **DriveGPT-WOMD** that is trained on WOMD, and **DriveGPT-Finetune** that is pretrained on our 120M internal research dataset and finetuned on WOMD. For baselines, we use a set of representative state-of-the-art models.

We report results on the WOMD test set<sup>2</sup> in Table 5. The results demonstrate that our method outperforms existing state-of-the-art non-ensemble models in terms of geometric metrics. Compared to Wayformer (Nayakanti et al., 2023) and MotionLM (Seff et al., 2023) that use ensembles of up to 8 replicas, our model achieves the best minADE and minFDE metrics and the second-best miss rate metric without any ensembling.

While we prioritize geometric metrics that emphasize the recall of predicted trajectory samples (i.e. not missing the critical trajectory), our model shows lower soft mAP scores due to suboptimal probability estimates. These estimates suffer from accumulated noises over time, as they are computed by compounding the action probabilities over a long sequence (as described in Eq. (1)), resulting in less accurate sample probabilities and reduced soft mAP, which relies on accurate probability assignments across predicted samples. Consequently, we notice that the gap in soft mAP grows as the prediction horizon increases. This reveals a limitation of using an autoregressive decoder for accurate probability estimation, as also noted in the LLM literature (Jiang et al., 2021; Geng et al., 2024). We defer improving the probability estimates of autoregressive models for behavior modeling as future work. One potential direction is to train an additional probability prediction head for each sample, which could enhance probability estimates and lead to improved soft mAP scores.

We observe up to 3% additional gains by pretraining on our internal dataset, despite a large distribution shift between our internal dataset and the public WOMD dataset, in terms

<sup>2</sup>Metrics are sourced from (Zhang et al., 2023) and WOMD leaderboard.

	minADE↓	minFDE↓	Miss Rate↓	Soft mAP↑
MTR (Shi et al., 2022)	0.6050	1.2207	0.1351	0.4216
HDGT (Jia et al., 2023)	0.7676	1.1077	0.1325	0.3709
HPTR (Zhang et al., 2023)	0.5565	1.1393	0.1434	0.3968
ControlMTR (Sun et al., 2024)	0.5897	1.1916	0.1282	<b>0.4572</b>
MTR++ (Shi et al., 2024)	0.5906	1.1939	0.1298	0.4414
Wayformer <sup>†</sup> (Nayakanti et al., 2023)	0.5454	1.1280	0.1228	0.4335
MotionLM <sup>†</sup> (Seff et al., 2023)	0.5509	1.1199	<b>0.1058</b>	<u>0.4507</u>
DriveGPT-WOMD	<u>0.5279</u>	<u>1.0609</u>	0.1236	0.3795
DriveGPT-Finetune	<b>0.5240</b>	<b>1.0538</b>	<u>0.1202</u>	0.3857

Table 5. On the WOMD test set, DriveGPT achieves better results in geometric metrics without any ensembling, and improved performance after pretraining. Results are averaged over three agent types (vehicle, pedestrian, cyclist) and three prediction horizons (3s, 5s, 8s). The best metric is highlighted in bold and the second best is underlined. † denotes ensemble.

Vehicle	minADE↓	minFDE↓	Miss Rate↓
MTR	0.7642	1.5257	0.1514
DriveGPT-WOMD	0.6396	1.2823	0.1213
DriveGPT-Finetune	<b>0.6326</b>	<b>1.2636</b>	<b>0.1181</b>
Cyclist	minADE↓	minFDE↓	Miss Rate↓
MTR	0.7022	1.4093	0.1786
DriveGPT-WOMD	0.6317	<b>1.2617</b>	0.1756
DriveGPT-Finetune	<b>0.6294</b>	1.2641	<b>0.1691</b>
Pedestrian	minADE↓	minFDE↓	Miss Rate↓
MTR	0.3486	0.7270	0.0753
DriveGPT-WOMD	0.3122	0.6388	0.0738
DriveGPT-Finetune	<b>0.3100</b>	<b>0.6338</b>	<b>0.0733</b>

Table 6. DriveGPT achieves better geometric metrics across a diverse set of agent types. Results are averaged over 3 prediction horizons.

of trajectory distributions, feature noises, and differences in semantic definitions.

Diving deeper into agent-specific results in Table 6, we see consistent improvements across all agent types, compared to MTR that shares the same encoder as ours. This further validates the generalizability of our method, in addition to vehicle behavior modeling results described in Sec. 5.1.

#### 5.2.4. QUALITATIVE COMPARISON

We present two qualitative comparisons in Fig. 11, where DriveGPT produces better trajectories in terms of diversity (covering more distinct outcomes) and accuracy (matching with the ground truth future) compared to MTR. This improvement is evident in challenging scenarios with limited agent history information (top row) and multiple future modalities (bottom row).

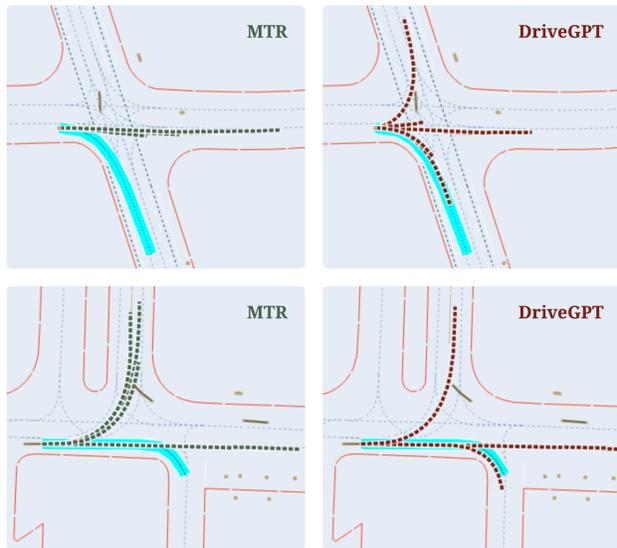


Figure 11. Compared to MTR (left), DriveGPT (right) produces more accurate and diverse trajectories in complex intersections. Blue represents ground truth future trajectories.

## 6. Conclusion

We introduced DriveGPT, an LLM-style autoregressive behavior model, to better understand the effects of model parameters and dataset size for autonomous driving. We systematically examined model performance as a function of both dataset size and model capacity, revealing LLM-like scaling laws for data and compute, as well as diminishing returns with increased model size. We showed the quantitative and qualitative benefits of scaling for planning in real-world driving scenarios. Additionally, we demonstrated our method on a public motion prediction benchmark, where DriveGPT outperformed state-of-the-art baselines and achieved improved performance through pretraining on a large-scale dataset.

## Acknowledgments

We thank, in alphabetical order, Abbas Shikari, Akshay Rangesh, Allan Lazarovici, Andy Zhuang, Ann Yan, Anton Spiridonov, Burkay Donderici, David Fang, Dennis Park, Florent Bekerman, Greg Meyer, Hans Zhang, Haytham Mohamed, Joe Liu, Jonathan Sorg, Junyuan Shi, Ling Li, Mao Ye, Matthew Creme, Peng Tang, Poonam Suryanarayan, Rizwan Chaudhry, Rosalie Zhu, Siva Karthik Mustikovela, Srinidhi Myadaboyina, Tanmay Agarwal, Xinlei Pan, Yi Hao, Yunlong He, Zaiwei Zhang, Zisu Dong, and many others from Cruise data and infra teams, for their contributions to the paper.

## Impact Statement

This paper aims to advance research in behavior modeling for autonomous vehicles. Since AVs interact with the public, our work has the potential to enhance their behavior around other road users, ultimately contributing to safer roads.

## References

- Chen, Y., Wang, Y., and Zhang, Z. DrivingGPT: Unifying driving world modeling and planning with multi-modal autoregressive transformers. *arXiv preprint arXiv:2412.18607*, 2024.
- Cui, H., Nguyen, T., Chou, F.-C., Lin, T.-H., Schneider, J., Bradley, D., and Djuric, N. Deep kinematic models for kinematically feasible vehicle trajectory predictions. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10563–10569. IEEE, 2020.
- Ettinger, S., Cheng, S., Caine, B., Liu, C., Zhao, H., Pradhan, S., Chai, Y., Sapp, B., Qi, C. R., Zhou, Y., et al. Large scale interactive motion forecasting for autonomous driving: The Waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9710–9719, 2021.
- Ettinger, S., Goel, K., Srivastava, A., and Al-Rfou, R. Scaling motion forecasting models with ensemble distillation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4812–4818. IEEE, 2024.
- Gan, Y., Xiao, H., Zhao, Y., Zhang, E., Huang, Z., Ye, X., and Ge, L. Multi-granular transformer for motion prediction with lidar. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 15092–15098. IEEE, 2024.
- Gao, J., Sun, C., Zhao, H., Shen, Y., Anguelov, D., Li, C., and Schmid, C. VectorNet: Encoding hd maps and agent dynamics from vectorized representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11525–11533, 2020.
- Geng, J., Cai, F., Wang, Y., Koepl, H., Nakov, P., and Gurevych, I. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 6577–6595, 2024.
- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- Hu, X., Yin, W., Jia, M., Deng, J., Guo, X., Zhang, Q., Long, X., and Tan, P. DrivingWorld: Constructing world model for autonomous driving via video GPT. *arXiv preprint arXiv:2412.19505*, 2024a.
- Hu, Y., Chai, S., Yang, Z., Qian, J., Li, K., Shao, W., Zhang, H., Xu, W., and Liu, Q. Solving motion planning tasks with a scalable generative model. In *European Conference on Computer Vision*, pp. 386–404. Springer, 2024b.
- Huang, X., Rosman, G., Gilitschenski, I., Jasour, A., McGill, S. G., Leonard, J. J., and Williams, B. C. HYPER: Learned hybrid trajectory prediction via factored inference and adaptive sampling. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2906–2912. IEEE, 2022.
- Jia, X., Wu, P., Chen, L., Liu, Y., Li, H., and Yan, J. HDGT: Heterogeneous driving graph transformer for multi-agent trajectory prediction via scene encoding. *IEEE transactions on pattern analysis and machine intelligence*, 2023.
- Jiang, C., Cornman, A., Park, C., Sapp, B., Zhou, Y., Anguelov, D., et al. MotionDiffuser: Controllable multi-agent motion prediction using diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9644–9653, 2023.
- Jiang, Z., Araki, J., Ding, H., and Neubig, G. How can we know when language models know? On the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- Kim, B., Park, S. H., Lee, S., Khoshimjonov, E., Kum, D., Kim, J., Kim, J. S., and Choi, J. W. LaPred: Lane-aware prediction of multi-modal future trajectories of dynamic agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14636–14645, 2021.

- Kim, S., Gholami, A., Shaw, A., Lee, N., Mangalam, K., Malik, J., Mahoney, M. W., and Keutzer, K. Squeezeformer: An efficient transformer for automatic speech recognition. *Advances in Neural Information Processing Systems*, 35:9361–9373, 2022.
- Liang, M., Yang, B., Hu, R., Chen, Y., Liao, R., Feng, S., and Urtasun, R. Learning lane graph representations for motion forecasting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 541–556. Springer, 2020.
- Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 1, 2020.
- Muennighoff, N., Rush, A., Barak, B., Le Scao, T., Tazi, N., Piktus, A., Pyysalo, S., Wolf, T., and Raffel, C. A. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376, 2023.
- Nayakanti, N., Al-Rfou, R., Zhou, A., Goel, K., Refaat, K. S., and Sapp, B. Wayformer: Motion forecasting via simple & efficient attention networks. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2980–2987. IEEE, 2023.
- Octo Model Team, Ghosh, D., Walke, H., Pertsch, K., Black, K., Mees, O., Dasari, S., Hejna, J., Xu, C., Luo, J., Kreiman, T., Tan, Y., Chen, L. Y., Sanketi, P., Vuong, Q., Xiao, T., Sadigh, D., Finn, C., and Levine, S. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.
- O’Neill, A., Rehman, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlekar, A., Jain, A., et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- Rhinehart, N., Kitani, K. M., and Vernaza, P. R2P2: A reparameterized pushforward policy for diverse, precise generative path forecasting. In *ECCV*, pp. 772–788, 2018.
- Seff, A., Cera, B., Chen, D., Ng, M., Zhou, A., Nayakanti, N., Refaat, K. S., Al-Rfou, R., and Sapp, B. MotionLM: Multi-agent motion forecasting as language modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8579–8590, 2023.
- Shi, S., Jiang, L., Dai, D., and Schiele, B. Motion transformer with global intention localization and local movement refinement. *Advances in Neural Information Processing Systems*, 35:6531–6543, 2022.
- Shi, S., Jiang, L., Dai, D., and Schiele, B. MTR++: Multi-agent motion prediction with symmetric scene modeling and guided intention querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3955–3971, 2024.
- Song, H., Luan, D., Ding, W., Wang, M. Y., and Chen, Q. Learning to predict vehicle trajectories with model-based planning. In *Conference on Robot Learning*, pp. 1035–1045. PMLR, 2022.
- Sun, J., Yuan, C., Sun, S., Wang, S., Han, Y., Ma, S., Huang, Z., Wong, A., Tee, K. P., and Ang, M. H. ControlMTR: Control-guided motion transformer with scene-compliant intention points for feasible motion prediction. In *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1507–1514. IEEE, 2024.
- Sun, Q., Huang, X., Gu, J., Williams, B. C., and Zhao, H. M2I: From factored marginal trajectory prediction to interactive prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6543–6552, 2022.
- Sun, Q., Zhang, S., Ma, D., Shi, J., Li, D., Luo, S., Wang, Y., Xu, N., Cao, G., and Zhao, H. Large trajectory models are scalable motion predictors and planners. *arXiv preprint arXiv:2310.19620*, 2023.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30(2017), 2017.
- Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12104–12113, 2022.
- Zhang, Z., Liniger, A., Sakaridis, C., Yu, F., and Van Gool, L. Real-time motion prediction via heterogeneous polyline transformer with relative pose encoding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11106–11115, 2021.
- Zhou, Z., Wang, J., Li, Y.-H., and Huang, Y.-K. Query-centric trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17863–17873, 2023.

## A. Training Detail

### A.1. Internal Evaluation: AV Planning

We train our models on the internal research dataset for 1 epoch, as customary in the LLM scaling literature (Kaplan et al., 2020; Hoffmann et al., 2022).

Our models follow the implementations described in (Nayakanti et al., 2023; Seff et al., 2023), and are trained using a batch size of 2048 and a standard Adam optimizer adopted in (Kaplan et al., 2020). We follow the optimal learning rate schedule discovered in (Hoffmann et al., 2022), which applies a cosine decay with a cycle length equivalent to the total number of training steps. All models were trained on 16 NVIDIA H100 GPUs.

### A.2. External Evaluation: Motion Prediction

We train our models on the WOMD data following the same setup in (Shi et al., 2022). More specifically, we use a batch size of 80 and an AdamW optimizer with a learning rate of 0.0001. The models are trained for 30 epochs, where the learning rate is decayed by a factor of 0.5 every 2 epochs, starting from epoch 20.

## B. Additional Ablation Studies

### B.1. Ablation Study on Attention Heads

In Fig. 12, we present an ablation study examining the impact of the number of heads in the encoder and decoder transformers, while keeping the hidden dimension the same, as specified in Table. 4. For each hidden dimension ( $d_{model}$ ), we conduct experiments across multiple configurations<sup>3</sup> of attention heads and present their validation losses using a distinct color. While we notice a pattern where more decoder attention heads lead to better performance for small models, the scaling trend is more influenced by the hidden dimension size than by the number of attention heads when the hidden dimension remains fixed.

### B.2. Ablation Study on Decoder Scaling

In Fig. 13, we present model scaling results (in dashed lines) where we only scale up the autoregressive decoder, while keeping the encoder at a fixed size. Compared to scaling both encoder and decoder (in solid lines), scaling only the decoder exhibits a similar trend but yields worse performance, especially for models exceeding 10M parameters in two larger datasets. Therefore, our main results focus on scaling both encoder and decoder.

<sup>3</sup>For the last group, we omit a few models due to memory constraints.

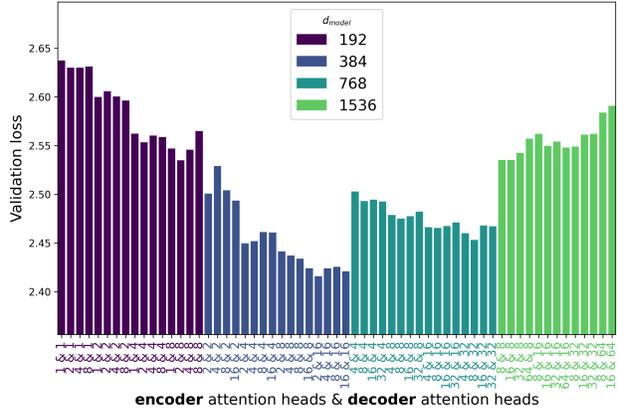


Figure 12. Adjusting the number of attention heads has a smaller effect on the trend of validation losses, compared to changing the hidden dimension.

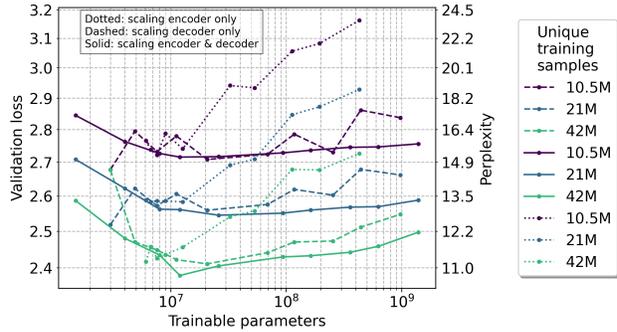


Figure 13. Scaling only the decoder exhibits a similar trend but worse results compared to scaling both encoder and decoder. Scaling only the encoder has little impact beyond 10M parameters.

### B.3. Ablation Study on Encoder Scaling

In Fig. 13, we present model scaling results (in dotted lines) where we only scale up the encoder, while keeping the decoder at a fixed size. The results indicate that encoder scaling is not as effective as decoder scaling.

## C. Scaling From a Perplexity View

Following the LLM scaling literature, we present our scaling results as a function of validation loss. Another key aspect to highlight is the perplexity scale. Perplexity, defined as the exponentiation of entropy, quantifies how well a probability model predicts a sample. It reflects the effective number of choices the model considers: a perplexity of  $k$  implies the model is  $k$ -way perplexed.

In Fig. 14, we re-plot the scaling results, adding a secondary  $y$ -axis to represent perplexity. In Fig. 15, we plot the per-

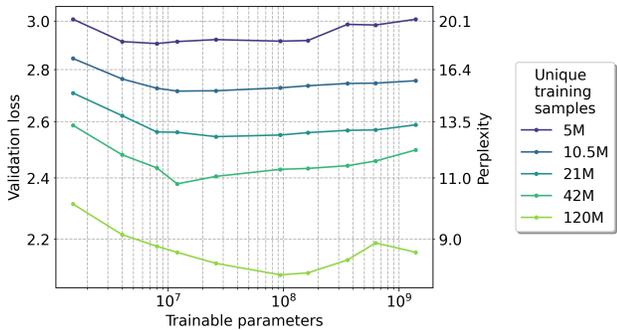


Figure 14. Fig. 4 replotted with a perplexity scale on the right  $y$ -axis for ease of reference.

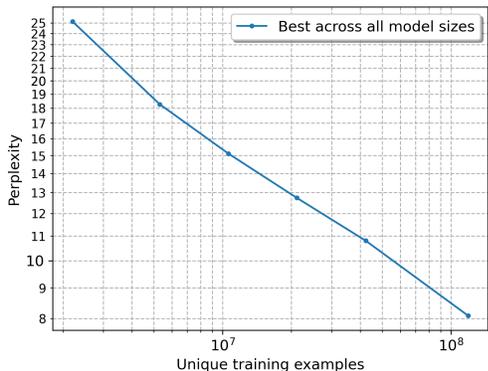


Figure 15. We fit the best perplexity values from all models in Fig. 3 as a function of dataset size. A simple linear regression yields  $\log(P) = -0.28 \log(D) + 7.22$  with an  $R^2$  value of 0.994, where  $P$  is the validation perplexity and  $D$  is the number of unique training examples.

plexity of our best models as a function of unique training examples. Extrapolating from the perplexity fit in the figure, a perplexity of 7 would require about 70M more training examples and a perplexity of 5 would require 500M more training examples.

## D. Additional Qualitative Examples

### D.1. Qualitative Comparison of Large Models

In Fig. 16, we present two qualitative examples demonstrating that a larger DriveGPT model (163M parameters) exhibits better collision avoidance capabilities, compared to its smaller 94M-parameter variant.

### D.2. Additional Closed-Loop Driving Examples

In the supplementary video at <https://youtu.be/-hLi44Pfy8g>, we present additional examples of deploying DriveGPT as a real-time motion planner in a closed-loop

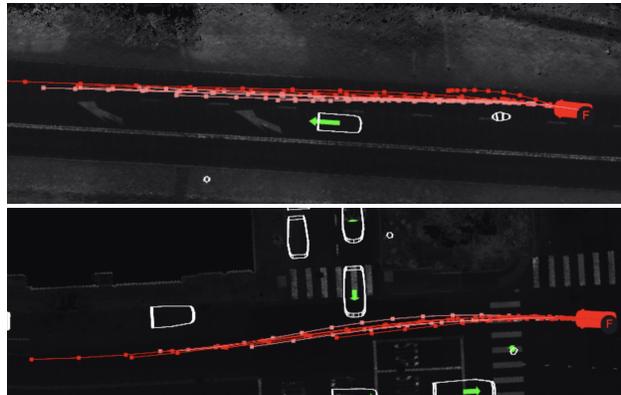


Figure 16. A larger DriveGPT model at 163M parameters generates trajectories (red) that maintain a safer distance from other traffic agents and road debris compared to a smaller model at 94M parameters (pink).

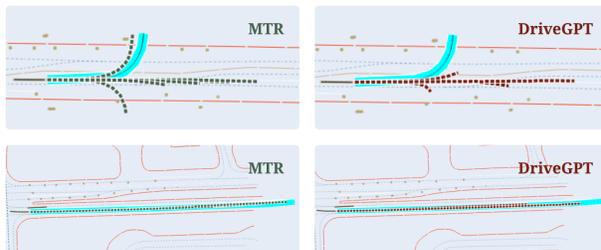


Figure 17. In examples where DriveGPT has the largest error regressions compared to MTR, it predicts less accurate trajectories due to the presence of lane boundary (top row) or missing map information (bottom row). Blue represents ground truth future.

setting. The video covers representative challenging scenarios in driving, including a) unprotected left turn, b) double parked vehicle, c) construction zone, d) blow-through cyclist, and e) lane change in heavy traffic.

### D.3. Failure Cases on WOMD

We show two representative failure examples of DriveGPT in Fig. 17, selected based on the largest minFDE regressions of our method compared to the MTR baseline on WOMD. In the first example (top row), our method fails to predict trajectories that make a left turn, likely due to the presence of a lane boundary. In the second example (bottom row), DriveGPT predicts an undershooting trajectory that results in a larger longitudinal error due to missing map information in the data.

Vehicle	Data	mADE↓	mFDE↓	MR↓
DriveGPT-WOMD	N/A	0.6381	1.2846	0.1233
DriveGPT-Finetune	40M	0.6368	1.2788	0.1217
DriveGPT-Finetune	80M	0.6341	1.2736	0.1207
DriveGPT-Finetune	120M	<b>0.6310</b>	<b>1.2655</b>	<b>0.1195</b>

Table 7. Using more training samples during pretraining gives better results (averaged over 3 horizons) in the finetuned model. This verifies the benefit of data scaling as shown in Sec. 4.1.

Model Parameters	mADE↓	mFDE↓	MR↓
2M	0.6953	1.4297	0.1489
4M	0.6534	1.3202	0.1298
9M	0.6423	1.2955	0.1249
14M	<b>0.6381</b>	<b>1.2846</b>	0.1233
21M	0.6396	1.2871	<b>0.1225</b>

Table 8. Scaling model parameters in DriveGPT-WOMD improves vehicle prediction metrics up to 14M model parameters. Results are averaged over 3 prediction horizons.

## E. Additional WOMD Validation Results

In this section, we present additional ablation studies on the WOMD validation set, due to the submission limit of the WOMD test server.

### E.1. WOMD Data Scaling in Pretraining

We validate the effectiveness of data scaling, by pretraining DriveGPT on various sizes of our internal dataset. The results in Table 7 indicate that pretraining on more unique samples leads to better results in the finetuned model, aligning with our findings in Sec. 4.1 that data scaling improves model performance.

### E.2. WOMD Model Scaling

We present an ablation study on model scaling using DriveGPT-WOMD in Table 8. The results show that the model performance continues to improve up to 14M parameters, with no further gains beyond this point due to the limited sample size in WOMD. Therefore, we choose to report the results from the 14M model in Sec. 5.2.