

# ETA: Efficiency through Thinking Ahead, A Dual Approach to Self-Driving with Large Models

Shadi Hamdan<sup>1,2\*</sup> Chonghao Sima<sup>3</sup> Zetong Yang<sup>4</sup> Hongyang Li<sup>3</sup> Fatma Güney<sup>1,2</sup>  
<sup>1</sup>Koç University <sup>2</sup>KUIS AI Center <sup>3</sup>The University of Hong Kong <sup>4</sup>OpenDriveLab

{shamdan17, fguney}@ku.edu.tr, chonghaosima@connect.hku.hk, hongyang@hku.hk, tomztyang@gmail.com

## Abstract

How can we benefit from large models without sacrificing inference speed, a common dilemma in self-driving systems? A prevalent solution is a dual-system architecture, employing a small model for rapid, reactive decisions and a larger model for slower but more informative analyses. Existing dual-system designs often implement parallel architectures where inference is either directly conducted using the large model at each current frame or retrieved from previously stored inference results. However, these works still struggle to enable large models for a timely response to every online frame. Our key insight is to shift intensive computations of the current frame to previous time steps and perform a batch inference of multiple time steps to make large models respond promptly to each time step. To achieve the shifting, we introduce Efficiency through Thinking Ahead (ETA), an asynchronous system designed to: (1) propagate informative features from the past to the current frame using future predictions from the large model, (2) extract current frame features using a small model for real-time responsiveness, and (3) integrate these dual features via an action mask mechanism that emphasizes action-critical image regions. Evaluated on the Bench2Drive CARLA Leaderboard-v2 benchmark, ETA advances state-of-the-art performance by 8% with a driving score of 69.53 while maintaining a near-real-time inference speed at 50 ms. Code and checkpoints can be found [here](#).

## 1. Introduction

Large models, including multimodal foundation models [5, 26] and vision foundation models with huge amount of parameters [6, 18], have been increasingly used in self-driving because of their impressive perception and reasoning capabilities [4, 9, 24, 36]. Due to the real-time demand of self-

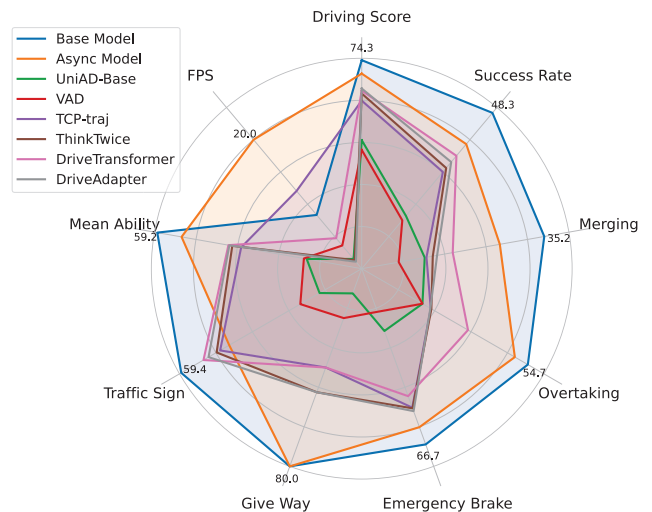


Figure 1. **Comparison on Bench2Drive [13].** Our **Base model** achieves the best performance across all metrics with a high latency. By improving latency to 20 FPS, our **Async model** achieves the second best in all metrics except Traffic Sign handling.

driving, inference-time speed becomes a critical concern when using large models. For instance, the best-performing models [11, 14] on the Bench2Drive closed-loop benchmark [13] currently have latencies in the order of hundreds of milliseconds due to their large transformer structures, which do not satisfy the real-time requirement. In this work, we address the challenge of utilizing large models for self-driving *without* compromising inference-time speed.

The efficiency bottleneck in large models has driven researchers to develop dual approaches for embodied systems [3, 8, 20, 30, 37, 39], including self-driving [25, 32, 33, 41]. Inspired by the System-1 and System-2 framework from cognitive science [16], these dual frameworks utilize a small model for fast, reactive decisions and a large model for slower, more thoughtful decisions. While the dual approach facilitates the use of large models in online systems, the inference results from these models in existing paradigms [33, 41] are often infrequently available.

\*Work partially done during internship at OpenDriveLab.

Primary contact: shamdan17@ku.edu.tr

Most outcomes are derived from immediate small model inference coupled with delayed large model inference (See Fig. 2). One potential solution for integrating large models into every frame is to store the outputs of the large model for complex cases in memory and retrieve relevant instances as needed using the small model, as explored in prior research [25]. However, managing an expanding memory bank in dynamic environments poses a challenge and can be difficult to generalize. In this paper, we manage to make the inference of large models timely available for each frame.

Our key insight is to enable large models for online inferencing every frame by an Efficiency through Thinking Ahead (ETA) pipeline, shifting the time-consuming computations of large models on the current frame to a previous time step, and performing a batch inference for multiple frames on one time to trade space complexity for time complexity, as shown in Fig. 2 (right). Consequently, each online frame is processed using large models from prior time steps, yielding outcomes that integrate both timely large model inference and small model inference.

However, shifting the computation of the current time step to a prior time step is non-trivial. As the information of the current time step is not available in previous frames. Our second novelty lies in updating the information from the prior time step for invisible online time step through i) propagating accurate information from the large model via future prediction; ii) processing the current time step with a small model to capture changes that are difficult to predict. Furthermore, we encourage observations and predicted actions to align more closely in the feature space, enabling the dual framework to function effectively. We design these modules to be lightweight, with real-time consideration in mind. Future prediction is supervised using the features of the large model in the current time step, but only during training, so the inference-time speed remains unaffected.

We evaluate the proposed ETA framework on the challenging Bench2Drive benchmark in the Leaderboard-v2 setting. Our extensive experiments demonstrate that using the large model asynchronously with the smaller model effectively balances performance and latency (Fig. 1). The proposed model demonstrates significant improvements over baselines, achieving a Driving Score (DS) of 69.53, which represents an improvement of 8.2% over the previous best DS, particularly in complex scenarios such as merging, overtaking, emergency braking, and yielding. Notably, the dual model achieves these results with a latency of 50 ms, the second lowest on the benchmark, following an MLP-based model [38], which was used as a sanity check.

The contributions are summarized as follows:

1. We introduce a new ETA paradigm to benefit from large models with real-time consideration;
2. We propose a dual framework with batched predictive large model inference and timely small model adjust-

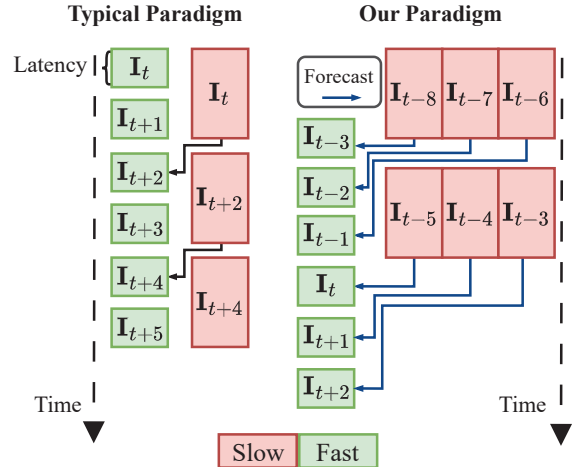


Figure 2. **Comparison of Dual Approaches.** Typically, dual-system approaches accommodate the larger latency of the slow, low-frequency model (red) by using the slow model predictions a fraction of the time. The fast, high frequency model (green) outputs predictions at every time step, incorporating the slow model’s outputs when available. In our approach, we batch and forecast in order to benefit from the larger model at every time step.

ment to implement the ETA paradigm;

3. We achieve competitive results with (close to) real-time performance in the closed-loop setting on the challenging Bench2Drive benchmark.

## 2. Related Work

### 2.1. Dual Approach to Embodied Systems

**Autonomous Driving:** Due to the large computational cost of LLMs and VLMs, a dual approach similar to ours has recently been explored for driving. DriveVLM-Dual [33] uses a larger VLM running asynchronously with a smaller traditional modular pipeline for 3D perception and motion planning by refining coarse waypoints predicted by the VLM. Despite being heavily optimized, 2B-parameter VLM takes 300ms to run, rendering closed-loop evaluation infeasible. Given observations and high-level textual instructions, AD-H [41] uses an LLM to output a higher level plan in the form of mid-level driving commands, such as turn left, etc., and conditioned on that, a small LLM predicts low-level waypoints. LeapAD [25] proposes a dual approach with a heuristic system for fast decisions based on samples retrieved from a memory and a slow analytical system that uses GPT4 for reasoning in the case of failures of the fast part and then stores the corrected action in the memory.

In a dual paradigm, the frequency of querying large model vs. small model is an important design decision, resulting in a trade-off between accuracy and speed. Our work distinguishes from prior works in that we enable utilization

of relevant large model features at every time step. We propose an efficient dual framework with batched predictive large model inference and timely small model adjustment.

**Robotics:** LLMs and VLMs are significantly welcomed in the robotics domain to provide open-world reasoning [7, 43]. Very recent works resort to a hierarchical structure to combine the large foundation models and low-level policy networks as a dual framework to realize both instruction decomposition and real-time control [29, 39]. Li *et al.* [20] propose to encode observations to latents with a VLM in low frequency and employ an action decoder conditioned on the latents with high prediction frequency. Instead, RoboDual [3] and Hi Robot [30] integrate different models for the dual system, adopting vision-language-action models for System-2 and System-1, respectively. This methodology has also been successfully applied in industrial companies such as Figure AI [8], where generalized performance and efficient deployment are achieved for whole upper-body control of humanoid robots.

## 2.2. Recent Progress on CARLA

**Closed-Loop Large Models on Leaderboard-v1:** Earlier autonomous driving frameworks utilizing large models evaluate on Town05 [25, 34] or LAV benchmark [40] in Leaderboard-v1 setting for closed-loop evaluation. These approaches first process the input using a large encoder, such as Qformer [34] or CLIP ViT [40], to tokenize the scene or a VLM to create a textual description of the scene [25] and then feed its output to an LLM/MLLM to predict action. LangAuto, another benchmark used by LLM/VLM-based agents [28, 41], provides a semi-closed-loop evaluation but with predefined routes with explicit instructions as input instead of route waypoints.

**Leaderboard-v2:** Due to the challenging long scenarios in CARLA Leaderboard-v2, there are only a few submissions, each scoring extremely low driving scores. The modular [41] or hybrid [17] approaches fall behind end-to-end approaches. End-to-end imitation learning approaches require high-quality driving demonstrations from an expert. There are two experts, i.e., with access to privileged data, commonly used for data collection in the Leaderboard-v2 setting: Think2Drive [19], which is a model-based RL agent, and PDMLite [2, 31], which is a rule-based planner. TransFuser++ [10], the best imitation learning agent in Leaderboard-v1, performs the second best in Leaderboard-v2 using data collected by the PDMLite expert.

LLM4AD/CarLLaVA [27], the winner of the CARLA Autonomous Driving Challenge 2.0, encodes the input with LLaVA-Next and then feeds its output to an LLM to predict action. Our base model is similar to LLM4AD/CarLLaVA but is still far behind real-time constraints at 102 ms, which is improved to 50 ms with our dual approach. As of Decem-

ber 2024, the CARLA leaderboard test server is temporarily closed and does not accept submissions.

**Bench2Drive Benchmark:** Due to the difficulty of comparing methods without significant differences in driving scores on challenging, long test routes of Leaderboard-v2, Bench2Drive [13] creates a benchmark with short routes, each focused on evaluating a distinct skill set. Several previous end-to-end driving methods are benchmarked by training on official training data collected by the Think2Drive expert [19]. DriveTransformer [14] proposes a transformer framework with different types of attention, achieving SOTA performance on Bench2Drive; however, it still has a high latency. Our dual framework achieves the best results with significantly lower latency (Fig. 1). Recent work [1, 42] reports results by collecting data from other experts, creating an unfair comparison due to changes to training data in terms of resolution, diversity, etc.

## 3. Methodology

### 3.1. Overall

In this work, we propose a dual system that leverages the concept of shifting intensive computations of large models from the current frame to previous frames. Additionally, we introduce the asynchronous batch inference technique, which performs feature extraction using large models across multiple frames simultaneously, thereby trading space complexity for time complexity. The integration of these two approaches enables timely inference from large models for every frame, effectively unleashing their potential in online systems.

**Predictive Large Model:** The predictive large model is specifically designed to transfer the computation of the current frame to a previous time step. Given two RGB images,  $\mathbf{I}_t$  and  $\mathbf{I}_{t-\Delta}$ , captured  $\Delta$  seconds apart, along with the current speed  $\mathbf{v}_t$  and the target waypoints  $\mathbf{w}_t$ , our objective is to predict the corresponding driving action  $\mathbf{a}_t$ . This action comprises a path made up of equidistant points and varying distance waypoints, as in [27].

To accomplish this, we first introduce a base model that employs a large encoder (Section 3.2), which, while accurate, is inefficient in terms of processing time. Subsequently, we develop our dual framework (Section 3.3), which processes the two time steps asynchronously. This method not only reduces latency but also retains the advantages of utilizing a large model.

**Asynchronous Batch Inference:** The concept of asynchronous batch inference is illustrated in Fig. 2. As demonstrated, to facilitate the use of large models for every frame, we parallelize the inference of previous frames to enable large models for per frame. By simultaneously processing multiple frames, we can significantly reduce the latency

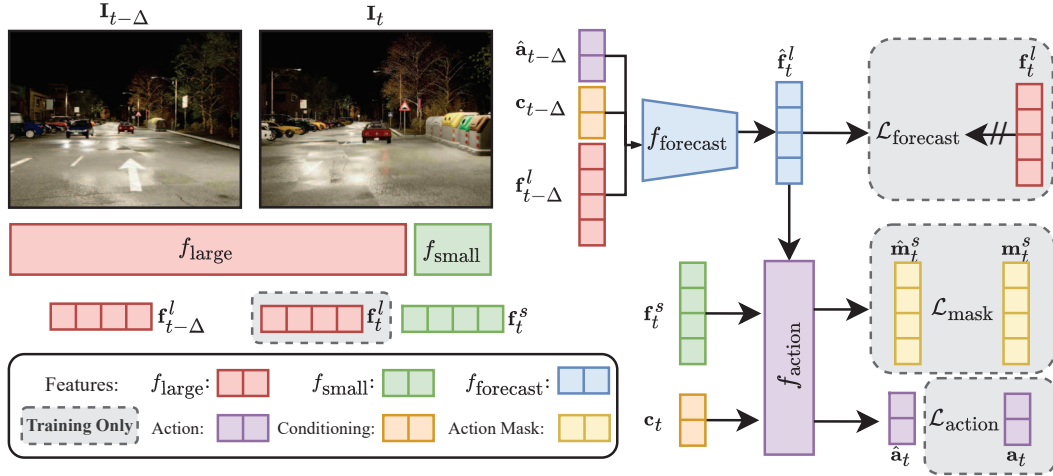


Figure 3. **Overview of the Asynchronous (Async) Model.** Our model processes two frames,  $\Delta$  apart in time, using the large model  $f_{\text{large}}$  for the previous frame  $\mathbf{I}_{t-\Delta}$  and the small model  $f_{\text{small}}$  for the current frame  $\mathbf{I}_t$ . Based on the previous frame’s features from the large model,  $\mathbf{f}_{t-\Delta}^l$ , along with conditioning inputs  $\mathbf{c}_t$  and  $\hat{\mathbf{a}}_t$ , we predict the current time-step features,  $\hat{\mathbf{f}}_t^l$ . The action  $\hat{\mathbf{a}}_t$  is then predicted using the action model  $f_{\text{action}}$ , which takes as input the forecasted features  $\hat{\mathbf{f}}_t^l$ , the current frame’s features from the small model  $\mathbf{f}_t^s$ , and the conditioning input  $\mathbf{c}_t$ . The forecasted features are supervised using the large model’s features at the current time step,  $\mathbf{f}_t^l$ . Since this supervision occurs during training alone, inference-time speed remains unaffected. Training-time computations are highlighted with gray boxes. In addition to forecasting and action losses, we apply a mask loss to better align observations with actions.

burden typically associated with large models, as multiple frames are processed at the same time, enabling more responsive and efficient online systems. This approach not only optimizes resource utilization but also enhances the overall performance of the predictive large model, ensuring timely and accurate driving actions.

### 3.2. Base Large Model

By using a large model in all time steps, we can train a base model to perform well but with high latency. This allows us to explore the upper bounds of driving performance with various large models if latency were not a concern.

For example, following the approach proposed by the winner [27] of the CARLA Autonomous Driving Challenge 2.0, we can use a large vision encoder  $f_{\text{large}}$ , e.g. a VLM or MLLM, to extract features from sensory input  $\mathbf{I}_t$  at time  $t$ , resulting in  $\mathbf{f}_t^l$ . We then feed it to an action model  $f_{\text{action}}$  to predict action  $\hat{\mathbf{a}}_t$  conditioned on  $\mathbf{c}_t$ , which is the concatenation of speed  $v_t$  and target waypoints  $\mathbf{w}_t$  at time  $t$ :

$$\mathbf{f}_t^l = f_{\text{large}}(\mathbf{I}_t), \quad (1)$$

$$\hat{\mathbf{a}}_t = f_{\text{action}}(\mathbf{I}_t, \mathbf{c}_t). \quad (2)$$

Following LLaVA’s anyres [22, 23] approach, we divide the image into two patches and then process them with the vision encoder. To reduce the number of output tokens, we apply  $2 \times 2$  spatial pooling, decreasing the token count by a factor of 4. Despite this reduction, the best-performing base models with a large encoder still fail to achieve a latency close to real-time constraints.

**Action Mask:** When we visualized the attention maps, we observed that the model struggles to focus on image patches that are relevant to action. To better align the predicted actions with the observations, we compute attention between patch features and the encoded queries corresponding to the action. This process generates a mask,  $\hat{\mathbf{m}}_t$ , showing the regions of the input image toward which the ego vehicle is likely to move (Fig. 4). We supervise this action mask with the ground truth mask, which is obtained by projecting expert actions onto the image (Section 3.4).

### 3.3. Asynchronous (Async) Model

As illustrated in Fig. 3, we propose a dual framework with a slow, large model  $f_{\text{large}}$  and a smaller, faster model  $f_{\text{small}}$ . Our small model is designed with real-time performance in mind. However, large models with strong performance typically exceed real-time constraints. To utilize the large model without increasing latency, we process the input at a previous time step  $t - \Delta$ , rather than the current time step  $t$ :

$$\mathbf{f}_{t-\Delta}^l = f_{\text{large}}(\mathbf{I}_{t-\Delta}). \quad (3)$$

$\mathbf{I}_{t-\Delta}$  denotes the input image at time  $t - \Delta$  with the corresponding features  $\mathbf{f}_{t-\Delta}^l$  obtained from the large model.

**Forecasting:** Given features  $\mathbf{f}_{t-\Delta}^l$  from the large model at a previous time step, we train a model  $f_{\text{forecast}}$  to predict features at the current time step. Similar to a world model [21], we condition this model on additional inputs, including the action predicted,  $\hat{\mathbf{a}}_{t-\Delta}$ , and  $\mathbf{c}_{t-\Delta}$ , which is the concatenation



Figure 4. **Action Mask.** In the top row, we show two examples of the RGB image with the path (purple) and waypoints (blue) projected onto it. Patches containing a path or waypoint are marked as 1 (yellow), and all other patches are marked as 0 (purple), creating the binary mask, overlaid on the image below for visualization.

tion of speed  $v_{t-\Delta}$  and target waypoints  $\mathbf{w}_{t-\Delta}$ :

$$\hat{\mathbf{f}}_t^l = f_{\text{forecast}}(\mathbf{f}_{t-\Delta}^l, \hat{\mathbf{a}}_{t-\Delta}, \mathbf{c}_{t-\Delta}). \quad (4)$$

The result of forecasting is the predicted features  $\hat{\mathbf{f}}_t^l$  for the current time step  $t$ .

**Small Model:** In addition to forecasting the features for the current time step, we use the small model,  $f_{\text{small}}$ , to process the current input  $\mathbf{I}_t$ , resulting in features  $\mathbf{f}_t^s$ :

$$\mathbf{f}_t^s = f_{\text{small}}(\mathbf{I}_t). \quad (5)$$

The small model is designed to capture changes between the previous and current time steps, particularly those that are difficult to predict, such as a traffic light’s state or the sudden appearance of a pedestrian.

**Action Prediction:** We concatenate the predicted features  $\hat{\mathbf{f}}_t$  from forecasting with the features  $\mathbf{f}_t^s$  from the small model and provide them as input to the action model,  $f_{\text{action}}$ , along with the conditioning input  $\mathbf{c}_t$ , resulting in the predicted action  $\hat{\mathbf{a}}_t$ :

$$\hat{\mathbf{a}}_t = f_{\text{action}}(\hat{\mathbf{f}}_t, \mathbf{f}_t^s, \mathbf{c}_t). \quad (6)$$

It is observed that the key to performance lies in forecasting and the small model working together to efficiently update past information from the large model to the current state.

### 3.4. Training

**Action Loss:** For the prediction of action, we apply an  $L_1$  loss between the predicted actions and the expert action, including the path and waypoints. Instead of directly predicting absolute positions, we learn to predict residuals. At test time, we sum these residuals to reconstruct the original waypoints. A similar approach is applied to the path tokens.

We find that this formulation stabilizes training compared to directly regressing path and waypoint coordinates.

**Action Mask Loss:** To generate supervision for the action mask, we project the expert action onto the input RGB image using the camera parameters. Patches containing a waypoint or path point are assigned a value of 1, while all others are set to 0 (Fig. 4). We apply a binary cross-entropy loss between the ground truth mask  $\mathbf{m}_t$  created this way and the predicted mask  $\hat{\mathbf{m}}_t$ .

The final loss of the base model is a weighted sum of the two loss terms where  $\lambda_1 = 1/16$ :

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{action}} + \lambda_1 \mathcal{L}_{\text{mask}}. \quad (7)$$

**Async Model with Forecasting Loss:** For the async model, we optimize the predicted action loss and mask loss as in the case of the base model (7), but the mask loss is applied to the features of the small model in the async case. Additionally, we introduce a forecasting loss,  $\mathcal{L}_{\text{forecast}}$ , to supervise the predicted features based on the large model’s features from the previous time step. With the forecasting loss, we aim to bring the predicted features closer to the ground truth features extracted from the current time step using the same large model. Specifically, we compute the absolute difference between ground truth features  $\mathbf{f}_t^l$  and the predicted features  $\hat{\mathbf{f}}_t^l$  from (4):

$$\mathcal{L}_{\text{forecast}}(\mathbf{f}_t^l, \hat{\mathbf{f}}_t^l) = |\mathbf{f}_t^l - \hat{\mathbf{f}}_t^l|. \quad (8)$$

We stop gradients from flowing through the ground truth features to prevent collapse. The weight of the mask loss remains the same as in the base model, and we set the weight of the forecasting loss to 0.5.

## 4. Experiments

### 4.1. Training Setting

We train all models for 40 epochs on  $8 \times$  A100 GPUs. We use CLIP-ViT-L-336px as the base large model encoder and the first 8 layers of CLIP-ViT-L-336px as the small model encoder, which we ablate in Supplementary. For the action decoder, we use a 12-layer transformer decoder. We use the base Adam optimizer with an initial learning rate of  $3 \cdot 10^{-5}$  and a cosine annealing schedule. The learning rate (LR) resets 4 times before reaching the minimum LR near the end of training. We set the global batch size to 160. For the Async Model, we set  $\Delta$  to 0.5s.

### 4.2. Evaluation Details

**Data:** The Bench2Drive [13] Benchmark comes with two data splits collected using the Think2Drive RL agent [19]. The base split, which consists of 1000 routes, is used to train and evaluate multiple models in the original paper

Table 1. **Comparison to SOTA on Bench2Drive.** In addition to the existent baselines, we compare our method to the recent DriveTransformer [14], selecting the best-performing variant for each model. Both our base and async models outperform *all* prior arts, with the async model achieving significantly lower latency with the proposed dual framework. Latency is reported in milliseconds.

Method	DS $\uparrow$	SR $\uparrow$ (%)	Efficiency $\uparrow$	Comfort $\uparrow$	Latency $\downarrow$
AD-MLP [38]	18.05	0.00	48.45	22.63	3
UniAD-Base [9]	45.81	16.36	129.21	43.58	663.4
VAD [15]	42.35	15.00	157.94	46.01	278.3
TCP-traj [35]	59.90	30.00	76.54	18.08	83
ThinkTwice [12]	62.44	31.23	69.33	16.22	762
DriveTransformer-Large [14]	63.46	35.01	100.64	20.78	211.7
DriveAdapter [11]	64.22	33.08	70.22	16.01	931
Base Model (Ours)	74.33 $\pm$ 0.99	48.33 $\pm$ 3.56	186.04 $\pm$ 6.47	25.77 $\pm$ 1.13	102
Async Model (Ours)	69.53 $\pm$ 1.62	38.64 $\pm$ 0.98	184.51 $\pm$ 1.54	28.43 $\pm$ 1.50	50

and following work. To maintain direct comparability to the evaluated models, we only train on the base split and do not use any additional data. We divide the data into a 94%-3%-3% training-validation-test split during our experiments. Similarly to CarLLaVA [27], we create different buckets containing different interesting scenarios, such as turning, near-collision, or intersections. During training, we use a weighted sampling approach to sample from the buckets. Please see Supplementary for details.

**Metrics:** We report the Driving Score (DS) and Success Rate (SR), along with Comfortness and Efficiency scores. Additionally, we provide per-ability scores to assess the model’s performance across different scenario categories. Finally, we include the latency of each model in milliseconds (ms). For our main models, we report results averaged over three runs with different seeds to account for variance in initialization and closed-loop evaluation. We report results from a single seed for ablation studies.

### 4.3. Comparison to State-of-the-Art

In Table 1, we compare our method to other approaches, including the originally reported AD-MLP [38], UniAD [9], VAD [15], TCP [35], ThinkTwice [12], and DriveAdapter [11]. Note that we are unable to quantitatively compare to other dual approaches [25, 33, 41] due to the lack of comparable closed-loop results.

Our base model, without asynchronous future prediction, achieves the highest driving score among all counterparts, with a latency of 102 ms. However, despite the strong performance of our base model, its 102ms latency imposes an upper bound of approximately 10Hz on the control frequency, which is relatively low for real-time control.

Using our asynchronous model, which performs the computation of the full vision encoder asynchronously until the time it is needed, we reduce latency to 50ms, enabling a control frequency of 20Hz. This marks a significant improvement over the base model, allowing for much more

responsive control. Despite not having access to the large model at the current time step, our async model still performs very well in terms of driving score and success rate, ranking second after our base model.

### 4.4. Ablation Studies

We perform an ablation study of our design choices in Table 2 to address the following research questions:

- Does supervising the large model for forecasting improve performance, or is the improvement solely due to scaling up the model parameters?
- Is the small model necessary, or can we rely solely on the forecasted features?
- Is additional supervision with the action mask essential?
- How does the small model perform without forecasting?
- Would better forecasting lead to better performance?
- In which scenarios do we observe improvements, and what are the underlying reasons?

**A: Without supervising forecasting,** the driving score drops to 54.92, highlighting the importance of training the large model to learn forecasting. The performance improvement cannot be solely attributed to the increased parameter count, as the large model’s features do not enhance performance without being explicitly trained for future prediction. By supervising the model to forecast with ground truth features, the large vision encoder effectively improves the driving score by approximately 15 points. Interestingly, the overall performance remains lower than when using only the small model (**D**), possibly due to confusion caused by suboptimal features from the large model.

**B: Without small model,** the performance drops to 42.49, indicating that forecasting with the large model alone is insufficient without up-to-date information from the small model. This may be due to scenarios where the future state is difficult to predict, such as changes in traffic light status. As shown in Table 3, removing the small model (**B**) results in a significant decline in emergency braking

Table 2. **Ablation Study.** We conduct an ablation study by removing each component of the model (A–C). Additionally, we report the performance of the small model alone (D) and further analyze forecasting by using ground truth features during both training and testing (E) or only during testing (F). ‘GT’ indicates ground truth. See the text for a detailed analysis.

ID	Method	DS↑	SR↑(%)	Efficiency↑	Comfort↑	Latency↓
	Base Model (Ours)	74.33	48.33	186.04	25.77	102
	Async Model (Ours)	69.53	38.64	184.51	28.43	50
A	Without Forecasting	54.92	30.45	177.52	18.71	50
B	Without Small Model	42.49	17.27	170.94	14.10	31
C	Without Action Mask	42.67	17.27	167.32	27.99	50
D	Small Model Only	61.30	35.00	183.63	43.91	50
E	GT Forecasting	74.12	48.18	192.72	24.76	124
F	GT Forecasting (Test Time)	60.72	32.27	151.53	17.25	124

Table 3. **Ablation Study according to Distinct Abilities.** To identify which components contribute to improvements in different scenarios, we report model performance in the ablation study based on distinct capabilities. For instance, removing the small model (C) leads to significant performance drops in the emergency brake (E. **Brake**) and traffic sign (T. **Sign**). See the text for a detailed analysis.

ID	Method	Merging↑	Overtaking↑	E. Brake↑	Give Way↑	T. Sign↑	Mean↑	Latency↓
	Base Model (Ours)	35.24	54.70	66.67	80.00	59.43	59.21	102
	Async Model (Ours)	26.66	50.42	60.13	80.00	43.64	52.17	50
A	Without Forecasting	20.00	35.90	41.18	80.00	36.18	42.65	50
B	Without Small Model	14.86	17.50	16.36	40.00	30.56	23.85	31
C	Without Action Mask	20.00	17.95	9.80	80.00	30.92	31.73	50
D	Small Model Only	27.14	41.03	47.06	60.00	44.74	43.99	50
E	GT Forecasting	29.33	57.50	56.36	60.00	53.89	51.41	124
F	GT Forecasting (Test)	28.57	38.46	43.14	60.00	43.42	42.72	124

ability (60.13→16.36) and proper handling of traffic signs (43.64→30.56). These results highlight the critical role of the small model in providing the model with up-to-date information by allowing it to glance at the current state.

**C: Without action mask,** the performance drops to 42.67, highlighting the importance of aligning the predicted actions with the observations. It might be surprising that such a small adjustment has such a significant impact on the async model, but this was one of the first insights we gained while attempting to make the base model work by examining the visualization of attention maps. This additional supervision helps the model better relate the predicted action to the observations. We found that removing it had a less severe effect on the base model (see Supplementary).

**D: Using only the small model,** performance drops to 61.30, demonstrating that the predicted features from the large model enhance overall performance. With our dual formulation, we can leverage these features without impacting latency. Comparing the async model with forecasting to using only the small model without forecasting (D) in terms of abilities in Table 3, we observe a notable decline in overtaking (50.42→41.03) and emergency braking (60.13→47.06) behaviors. A possible explanation is that

the large model provides strong features for predictable scenarios, enabling the small model to focus on detecting deviations, such as emergencies. Without the large model, the small model must handle all situations, leaving fewer resources for effectively managing emergencies.

**E & F: Using ground truth features for forecasting,** we further examine the impact of forecasting on performance. When ground truth features replace forecasted features during both training and testing (E), the driving score improves to 74.12, surpassing the score achieved with our async model’s forecasted features (69.53). This gap suggests that forecasting can still be improved, and with perfect forecasting, the model can perform at the same level as our base model (74.33), ruling out potential design limitations. However, when ground truth features are provided only at test time (F), the driving score drops significantly to 60.72. This suggests a distribution mismatch between the predicted and ground truth features, even though the model is trained with supervision from ground truth features.

#### 4.5. Qualitative Analysis

We visually investigate the contributions of the small model and the large model with forecasting through two example

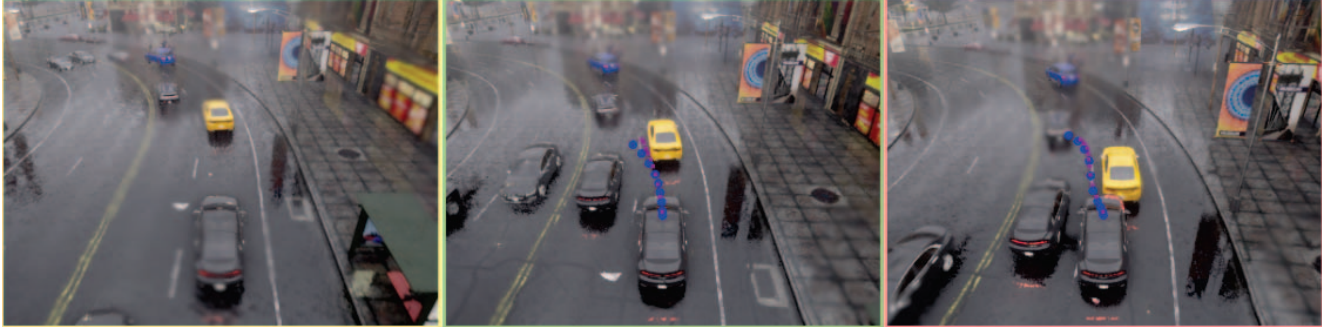


Figure 5. **Hard Brake without Small Model.** **Left:** The initial state of the scene where the yellow car ahead suddenly brakes, requiring the ego vehicle to perform a hard brake. **Middle:** The Async Model detects the hazard using the small model and stops in time. **Right:** The version without the small model (**B**) relies solely on forecasting and fails to detect the sudden stop in time, resulting in a crash. The predicted action for each model is overlaid on the image, with waypoints shown in blue and the path in red.



Figure 6. **Lane Change without Forecasting.** **Left:** The initial state of the scene, where the ego vehicle is tasked with switching to the left lane. **Middle:** The Async Model successfully executes the lane change while avoiding collisions by leveraging forecasting. **Right:** The version without forecasting (**A**) fails to anticipate the future position of the vehicle behind, resulting in a rear-end collision. The predicted action for each model is overlaid on the image, with waypoints shown in blue and the path in red.

scenarios: a hard brake without the small model (Fig. 5) and a lane change without forecasting supervision (Fig. 6). These correspond to rows **B** and **A** in the ablation study (Table 2), respectively. In Fig. 5, the version in **B** fails to perform a hard brake, as it cannot detect the sudden stop of the yellow vehicle ahead without the small model, highlighting its role in detecting emergencies that appear in the current time step. Similarly, in Fig. 6, the version in **C** results in a rear-end crash due to its failure to anticipate the future position of the vehicle behind without forecasting.

## 5. Conclusion

We introduce a novel dual paradigm that shifts computation to the previous time step, enabling more efficient processing with large models in potential applications in embodied systems. To implement this paradigm in self-driving, we proposed ETA, a dual framework that combines forecasting to propagate valuable information from a large model at the prior time step to the current time step and a small model to capture unpredictable changes, such as sudden brakes

due to emergencies. ETA advances state-of-the-art driving performance on Bench2Drive by 8%, also reducing latency from hundreds of milliseconds in previous methods to just 50 ms. Through detailed ablation studies, we demonstrate the importance of small model and forecasting based on the large model’s output, highlighting the effectiveness of ETA to combine the two in a dual framework.

**Limitations and Future Work:** While our async model significantly narrows the gap to the base model in critical driving scenarios such as overtaking, emergency braking, and yielding, some challenges remain. The largest discrepancies between the base and async models occur in the traffic sign handling and merging scenarios.

Future work will explore improving forecasting, as our analysis shows that using ground truth features instead of forecasting brings the async model to the level of the base model. Furthermore, we see potential for pushing performance even further by leveraging larger and more advanced vision encoders, which could enhance both perception and decision-making capabilities.

## Acknowledgements

This project is funded by the European Union (ERC, ENSURE, 101116486) with additional compute support from Leonardo Booster (EuroHPC Joint Undertaking, EHPC-AI-2024A01-060). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This study is also supported by National Natural Science Foundation of China (62206172) and Shanghai Committee of Science and Technology (23YF1462000). We extend our gratitude to Li Chen and the rest of the members from OpenDriveLab for their profound discussions and supportive help in writing and related work. We also extend our gratitude to Görkay Aydemir and the rest of the members from the Autonomous Vision Group at Koc University for their constructive feedback and discussions.

## References

- [1] Fazel Arasteh, Mohammed Elmahgiubi, Behzad Khamidehi, Hamidreza Mirkhani, Weize Zhang, Cao Tongtong, and Kasra Rezaee. Validity learning on failures: Mitigating the distribution shift in autonomous vehicle planning. *arXiv preprint arXiv:2406.01544*, 2024. 3
- [2] Jens Beißwenger. PDM-Lite: A rule-based planner for carla leaderboard 2.0. 2024. 3
- [3] Qingwen Bu, Hongyang Li, Li Chen, Jisong Cai, Jia Zeng, Heming Cui, Maoqing Yao, and Yu Qiao. Towards synergistic, generalized, and efficient dual-system for robotic manipulation. *arXiv preprint arXiv:2410.08001*, 2024. 1, 3
- [4] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE TPAMI*, 2024. 1
- [5] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024. 1
- [6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1
- [7] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model. In *ICML*, 2023. 3
- [8] Figure. Helix: A vision-language-action model for generalist humanoid control. 2025. 1, 3
- [9] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *CVPR*, 2023. 1, 6
- [10] Bernhard Jaeger, Kashyap Chitta, and Andreas Geiger. Hidden biases of end-to-end driving models. In *ICCV*, 2023. 3
- [11] Xiaosong Jia, Yulu Gao, Li Chen, Junchi Yan, Patrick Langechuan Liu, and Hongyang Li. DriveAdapter: breaking the coupling barrier of perception and planning in end-to-end autonomous driving. In *ICCV*, 2023. 1, 6
- [12] Xiaosong Jia, Penghao Wu, Li Chen, Jiangwei Xie, Conghui He, Junchi Yan, and Hongyang Li. Think twice before driving: Towards scalable decoders for end-to-end autonomous driving. In *CVPR*, 2023. 6
- [13] Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2Drive: towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. In *NeurIPS 2024 Datasets and Benchmarks Track*, 2024. 1, 3, 5
- [14] Xiaosong Jia, Junqi You, Zhiyuan Zhang, and Junchi Yan. DriveTransformer: unified transformer for scalable end-to-end autonomous driving. In *ICLR*, 2025. 1, 3, 6
- [15] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. VAD: vectorized scene representation for efficient autonomous driving. In *ICCV*, 2023. 6
- [16] Daniel Kahneman. *Thinking, fast and slow*. Farrar, Straus and Giroux, 2011. 1
- [17] Gemb Kaljavesi, Xiyan Su, and Frank Diermeyer. Integrating end-to-end and modular driving approaches for online corner case detection in autonomous driving. *arXiv preprint arXiv:2409.01178*, 2024. 3
- [18] Youngwan Lee, Joong-won Hwang, Sangrok Lee, Yuseok Bae, and Jongyoul Park. An energy and gpu-computation efficient backbone network for real-time object detection. In *CVPR Workshop*, 2019. 1
- [19] Qifeng Li, Xiaosong Jia, Shaobo Wang, and Junchi Yan. Think2Drive: efficient reinforcement learning by thinking in latent world model for quasi-realistic autonomous driving (in carla-v2). In *ECCV*, 2024. 3, 5
- [20] Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, Xiaofan Wang, Bei Liu, Jianlong Fu, Jianmin Bao, Dong Chen, Yuanchun Shi, Jiaolong Yang, and Baining Guo. CogACT: a foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024. 1, 3
- [21] Yingyan Li, Lue Fan, Jiawei He, Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang, and Tieniu Tan. Enhancing end-to-end autonomous driving with latent world model. In *ICLR*, 2025. 4
- [22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 4
- [23] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 4

- [24] Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023. 1
- [25] Jianbiao Mei, Yukai Ma, Xuemeng Yang, Licheng Wen, Xinyu Cai, Xin Li, Daocheng Fu, Bo Zhang, Pinlong Cai, Min Dou, Botian Shi, Liang He, Yong Liu, and Yu Qiao. Continuously learning, adapting, and improving: A dual-process approach to autonomous driving. In *NeurIPS*, 2024. 1, 2, 3, 6
- [26] OpenAI. OpenAI: Introducing ChatGPT. 2022. 1
- [27] Katrin Renz, Long Chen, Ana-Maria Marcu, Jan Hünermann, Benoit Hanotte, Alice Karnsund, Jamie Shotton, Elahe Arani, and Oleg Sinavski. CarLLaVA: Vision language models for camera-only closed-loop driving. *arXiv preprint arXiv:2406.10165*, 2024. 3, 4, 6
- [28] Hao Shao, Yuxuan Hu, Letian Wang, Steven L. Waslander, Yu Liu, and Hongsheng Li. LMDrive: Closed-loop end-to-end driving with large language models. In *CVPR*, 2024. 3
- [29] Yide Shentu, Philipp Wu, Aravind Rajeswaran, and Pieter Abbeel. From LLMs to Actions: Latent codes as bridges in hierarchical robot control. *arXiv preprint arXiv:2405.04798*, 2024. 3
- [30] Lucy Xiaoyang Shi, Brian Ichter, Michael Equi, Liyiming Ke, Karl Pertsch, Quan Vuong, James Tanner, Anna Walling, Haohuan Wang, Niccolo Fusai, Adrian Li-Bell, Danny Driess, Lachy Groom, Sergey Levine, and Chelsea Finn. Hi Robot: Open-ended instruction following with hierarchical vision-language-action models. *arXiv preprint arXiv:2502.19417*, 2025. 1, 3
- [31] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. DriveLM: Driving with graph visual question answering. In *ECCV*, 2024. 3
- [32] Chonghao Sima, Kashyap Chitta, Zhiding Yu, Shiyi Lan, Ping Luo, Andreas Geiger, Hongyang Li, and Jose M Alvarez. Centaur: Robust end-to-end autonomous driving with test-time training. *arXiv preprint arXiv:2503.11650*, 2025. 1
- [33] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Zhiyong Zhao, Yang Wang, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. DriveVLM: The convergence of autonomous driving and large vision-language models. In *CoRL*, 2024. 1, 2, 6
- [34] Wenhai Wang, Jiangwei Xie, ChuanYang Hu, Haoming Zou, Jianan Fan, Wenwen Tong, Yang Wen, Silei Wu, Hanming Deng, Zhiqi Li, Hao Tian, Lewei Lu, Xizhou Zhu, Xiaogang Wang, Yu Qiao, and Jifeng Dai. DriveMLM: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv preprint arXiv:2312.09245*, 2023. 3
- [35] Penghao Wu, Xiaosong Jia, Li Chen, Junchi Yan, Hongyang Li, and Yu Qiao. Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. In *NeurIPS*, 2022. 6
- [36] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. DriveGPT4: Interpretable end-to-end autonomous driving via large language model. *RA-L*, 2024. 1
- [37] Ping Yu, Jing Xu, Jason Weston, and Ilia Kulikov. Distilling system 2 into system 1. *arXiv preprint arXiv:2407.06023*, 2024. 1
- [38] Jiang-Tian Zhai, Ze Feng, Jinhao Du, Yongqiang Mao, Jiang-Jiang Liu, Zichang Tan, Yifu Zhang, Xiaoqing Ye, and Jingdong Wang. Rethinking the open-loop evaluation of end-to-end autonomous driving in nuscenes. *arXiv preprint arXiv:2305.10430*, 2023. 2, 6
- [39] Jianke Zhang, Yanjiang Guo, Xiaoyu Chen, Yen-Jen Wang, Yucheng Hu, Chengming Shi, and Jianyu Chen. HiRT: Enhancing robotic control with hierarchical robot transformers. In *CoRL*, 2024. 1, 3
- [40] Jimuyang Zhang, Zanming Huang, Arijit Ray, and Eshed Ohn-Bar. Feedback-guided autonomous driving. In *CVPR*, 2024. 3
- [41] Zaibin Zhang, Shiyu Tang, Yuanhang Zhang, Talas Fu, Yifan Wang, Yang Liu, Dong Wang, Jing Shao, Lijun Wang, and Huchuan Lu. AD-H: autonomous driving with hierarchical agents. *arXiv preprint arXiv:2406.03474*, 2024. 1, 2, 3, 6
- [42] Julian Zimmerlin, Jens Beißwenger, Bernhard Jaeger, Andreas Geiger, and Kashyap Chitta. Hidden biases of end-to-end driving datasets. *arXiv preprint arXiv:2412.09602*, 2024. 3
- [43] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *CoRL*, 2023. 3