ECoSim: Data Efficient Fine-Tuning for Controllable Traffic Simulation

Yu-Hsiang Chen[1,2*], Wei-Jer Chang[2*], Yi-Ting Chen[2], and Masayoshi Tomizuka[2]

[1] National Yang Ming Chiao Tung University
[2] University of California, Berkeley

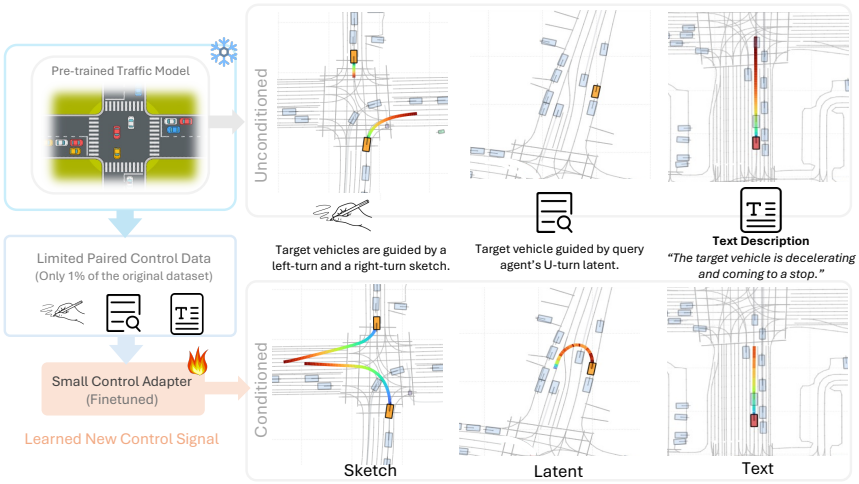

Fig. 1: Data-Efficient Multi-Modal Control of Pretrained Traffic Models. We introduce a data-efficient adaptation framework that steers frozen traffic simulators via sketch, latent, or text signals. Utilizing only 1% of training data, it transforms unconditioned pretrained traffic model into precise, user-defined maneuvers without compromising the base generative prior.

Abstract. Controllable traffic simulation is critical for testing autonomous driving systems, yet existing approaches often require retraining large generative models with extensive annotated data. We introduce a lightweight control adaptation framework that endows frozen state-of-the-art traffic models—including both diffusion and autoregressive backbones—with multi-modal controllability (sketch, latent behavior codes, and text). By modulating intermediate features through identity-initialized FiLM layers, our method enables precise control while preserving the base model's generative prior. Evaluated on Waymo Open Sim Agents Challenge, our approach demonstrates strong controllability with less than **1% of the training data**. Furthermore, through context-aware condition transfer, our framework enables counterfactual scenario generation and long-tail synthesis while maintaining stable closed-loop driving realism and safety.

* Equal contribution.

Our framework unlocks new possibilities for controllable traffic simulation, enabling targeted scenario generation through lightweight adaptation of pretrained generative models.

Keywords: Controllable Traffic Simulation · Data Efficient Adaptation

1 Introduction

Traffic simulation is essential to autonomous vehicle (AV) development, enabling scalable and repeatable evaluation in controlled settings. It supports systematic validation under diverse driving conditions, shortens development cycles, and provides a safe platform for training and assessment. A key requirement is *controllability*: the ability to deliberately induce targeted behavior patterns that probe specific system responses. Many traffic scenes are non-interactive and trivial, offering limited insight into failure modes. With controllability, simulation enables the construction of counterfactual "what-if" scenarios that expose potential weaknesses and improve diagnostic effectiveness.

Despite its importance, enabling controllable traffic simulation remains challenging. Prior approaches typically fall into two paradigms: inference-time guidance and direct conditioning through retraining. Inference-time methods, such as diffusion via optimization or gradient-based control during sampling [2,12,14], offer flexible guidance. However, such approaches often rely on carefully designed objectives and introduce additional computational overhead at inference. Alternatively, direct conditioning enables models [3, 20, 21, 26] to internalize control signals during training, often yielding stable and faithful behavior. However, this comes at the cost of large-scale paired data and expensive retraining for each new control modality. Also, most existing evaluations focus on alignment with ground-truth trajectories, leaving the ability to generate controllable counterfactual scenarios across diverse contexts less explored.

In this work, we introduce a data-efficient fine-tuning framework that enables controllable traffic simulation using pretrained models trained on large-scale data. We freeze the base model and learn lightweight control adapters that inject conditioning signals into intermediate features. The adapters are trained under closed-loop simulation to enable controllability while preserving long-horizon interaction dynamics. Our framework supports multiple control modalities. Trajectory sketches provide precise spatial guidance, latent behavior embeddings extracted by our BehaviorVAE capture driving patterns learned from data, and natural language offers an intuitive user interface for specifying high-level behaviors. The same adaptation mechanism generalizes across different model paradigms, including diffusion-based and autoregressive traffic models. Finally, we introduce a context-aware retrieval pipeline that transfers learned behaviors into compatible scenes, enabling counterfactual and long-tail scenario generation.

We evaluate on the Waymo Open Sim Agents Challenge (WOSAC) [15] under closed-loop simulation. Across both diffusion-based and autoregressive models, our lightweight adapters reduce control error by up to 83% while improving

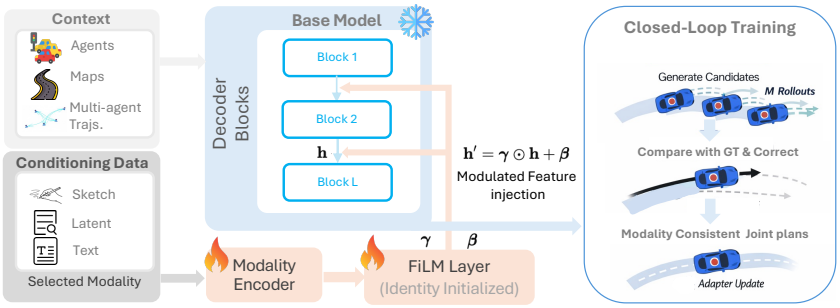

Fig. 2: Model-Agnostic Control Adaptation Architecture. A lightweight adapter injects multi-modal control signals into a frozen pretrained traffic model by predicting FiLM parameters $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ that modulate intermediate features $\mathbf{h}$. The design is compatible with both autoregressive and diffusion backbones, and identity initialization preserves the base model's generative prior.

closed-loop realism by 0.02–0.03 absolute points in the WOSAC Meta score (see Sec. 4.2). Remarkably, these gains are achieved using as little as 0.01%–1% paired fine-tuning data. Despite this data efficiency, our method matches or surpasses fully fine-tuned baselines trained on substantially larger datasets [20].

Our main contributions can be summarized as follows:

- We propose a data-efficient control adaptation framework that enables multi-modal controllability (*sketch*, *latent*, and *text*) for both diffusion and state-of-the-art autoregressive traffic models.
- We demonstrate strong **sample efficiency**. With identity-initialized FiLM adapters, our method achieves competitive controllability using minimal paired supervision (as few as 500 scenarios, 0.1% data) while keeping the pretrained backbone frozen.
- The proposed context-aware condition transfer mechanism enables counterfactual and long-tail scenario generation while preserving closed-loop realism.

2 Related Work

Controllable Traffic Scenario Generation. Recent work has made significant progress in building data-driven, realistic traffic simulators that support multi-agent interactions in closed-loop environments [1, 10, 13]. Recent autoregressive (AR) and reactive simulators achieve strong closed-loop realism by modeling traffic generation as sequential next-token prediction [22, 24, 28]. However, these models are primarily designed for unconditional simulation and provide limited mechanisms for controllable scenario synthesis, which is crucial for structured, target-case testing [9, 17].

Several works explore the generation of controllable traffic scenarios to synthesize targeted or safety-critical situations [2, 4, 9]. Existing approaches can be broadly categorized into three mechanisms: (1) *Inference-time guidance* steers unconditional samples during generation via gradient-based or constraint-based

objectives [14, 27]. While flexible, it often introduces significant inference overhead. (2) *Direct conditioning* trains generative models from scratch to internalize explicit control signals (e.g., goals specifications, scene constraints or language descriptions) [3, 20, 21, 26]. However, introducing new modalities requires expensive retraining. (3) *Retrieval-based control* guides generation using similar scenarios retrieved from large datasets, such as RealGen [7]. In contrast, we adapt pretrained traffic models to support new control modalities without retraining the backbone. Our framework further extends retrieval-based guidance by introducing multiple control interfaces, including behavior latents learned via BehaviorVAE, enabling context-aware control within closed-loop simulation.

Control Adaptation in Generative Models. Recent works in image and video synthesis explore adapting pretrained generative models to enable controllable generation. Rather than retraining large conditional models from scratch, these approaches augment pretrained backbones with mechanisms that incorporate new conditioning signals. For example, ControlNet [23] introduces dedicated control branches connected via zero-initialized convolutions that progressively learn to modulate a pretrained diffusion model in response to structural inputs such as edges or poses. Similarly, LoRA [11] adapts pretrained models by injecting low-rank updates into existing weight matrices, enabling new capabilities while preserving the original model parameters. These works demonstrate that powerful generative models can be extended to new control modalities by leveraging pretrained generative priors.

Despite their success in visual generation, such control adaptation strategies remain largely unexplored in traffic modeling, where preventing compounding errors requires closed-loop training. We introduce a control adaptation framework optimized under closed-loop simulation to enable controllability while preserving the behavior of pretrained traffic models.

3 Method

3.1 Problem Formulation

Let π_{ref} denote a model pre-trained on large-scale, unconditional driving data that generates future trajectories in a closed-loop manner given a scene context c. Our goal is to extend this model to conditional generation, $\pi_\theta(\mathbf{S}_{t:t+T} \mid c, r)$, where r denotes a control signal (e.g., trajectory sketches, latent behavior codes, or language descriptions) and to enable controllable multi-agent generation in a data-efficient manner. The notation $\mathbf{S}_{t:t+T}$ denotes the joint multi-agent future states over a prediction horizon. The generator should control designated agents according to r while preserving realistic interactions of the pre-trained model π_{ref}. For each control modality m, we assume access to a small control-annotated dataset $\mathcal{D}_m \subset \mathcal{D}$ and aim to adapt the pretrained simulator with limited supervision.

3.2 Control Adaptation via Feature Modulation

Given a raw control input r_m from modality m, we first map it to a unified d-dimensional conditioning embedding $z \in \mathbb{R}^d$ via a modality-specific encoder (where $d = 256$ in our implementation). For each agent i, only targeted agents receive their corresponding z_i, while non-targeted agents are assigned $z_i = \mathbf{0}$.

To enable controllable generation while preserving the pretrained traffic model prior, we inject the conditioning embedding z_i into intermediate representations of the frozen backbone through lightweight feature modulation. We adopt Feature-wise Linear Modulation (FiLM) [16], which conditions intermediate activations via an affine transformation. For an intermediate feature map $\mathbf{h}^{(l)} \in \mathbb{R}^D$ at the l-th decoder layer (where D is the hidden dimension of the backbone), FiLM applies:

$$\mathrm{FiLM}(\mathbf{h}^{(l)}; z_i) = \boldsymbol{\gamma}^{(l)}(z_i) \odot \mathbf{h}^{(l)} + \boldsymbol{\beta}^{(l)}(z_i), \tag{1}$$

where $\boldsymbol{\gamma}^{(l)}, \boldsymbol{\beta}^{(l)} \in \mathbb{R}^D$ are layer-specific modulation functions (implemented as lightweight MLPs) predicted from z_i.

A key design principle of our framework is extensibility: pretrained traffic models should be incrementally augmented with new controllability without retraining or destabilizing the original model. To this end, we initialize the modulation layers near identity ($\boldsymbol{\gamma}^{(l)} \approx \mathbf{1}$, $\boldsymbol{\beta}^{(l)} \approx \mathbf{0}$), ensuring that the adapted simulator initially reproduces the base model exactly. This identity-preserving initialization is particularly important in closed-loop traffic simulation, where small perturbations can compound over long horizons and disrupt multi-agent interactions. Conceptually related to zero-initialized conditioning strategies such as ControlNet [23], this design enables controllability to be learned gradually through lightweight additional parameters while maintaining the backbone's learned dynamics.

3.3 Modality-Specific Control Representations

Sketch and Language Conditioning. Trajectory sketches provide coarse spatial guidance in the form of future waypoints expressed in the agent's local frame. We encode the waypoint sequence using a lightweight temporal encoder and aggregate it into a fixed-dimensional embedding z_i, capturing the intended motion trend while allowing reactive interactions.

Following [3], we utilize a DistilBERT [19] model adapted via Low-Rank Adaptation (LoRA) [11] for natural language commands (e.g., "turn left at the intersection"). While language instructions are inherently abstract, we establish spatial and kinematic grounding by jointly optimizing the LoRA parameters with the downstream closed-loop control objective. Specifically, given a text prompt p_i, we extract the language features $\mathbf{H}_i = \mathrm{BERT}(p_i)$. We pool the `[CLS]` token embedding and apply a learned linear projection to match the shared d-dimensional control space.

Context-Aware Behavior Latent. Beyond explicit sketches or language commands, we introduce a context-aware behavior latent that captures high-level

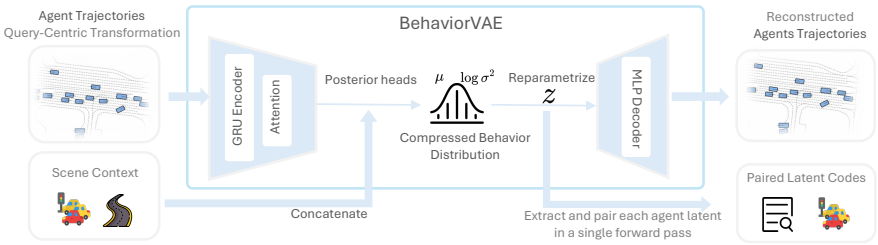

Fig. 3: BehaviorVAE overview. Agent trajectories and scene context are encoded into a Gaussian latent posterior; reparameterized per-agent latents are decoded for trajectory reconstruction and exported as paired latent codes in a single forward pass.

driving patterns directly from data, without manual annotations. This compact control space abstracts interaction styles such as yielding or merging, enabling scalable and data-efficient controllability (see Sec. 4.3).

To learn this representation, we employ a Conditional VAE (CVAE) [25], referred to as *BehaviorVAE*. We adopt a CVAE instead of deterministic autoencoder-based approaches [7, 18] because KL regularization encourages a smooth latent space where similar behaviors cluster together. This representation enables strong sample efficiency, as shown in Sec. 4.3. Given a traffic scene, the encoder jointly processes all agents' future motion sequences together with scene context in a single forward pass. By modeling relative geometry and multi-agent dependencies, it infers a latent variable z_i for each agent that captures structured interaction behaviors in a context-aware manner (See Fig. 3).

Formally, the encoder produces a Gaussian latent distribution $q_\phi(z_i \mid \boldsymbol{\xi}, c_{\text{env}})$, where $\boldsymbol{\xi}$ denotes the set of agent trajectories in the scene. Because latent inference is performed jointly at the scene level, the resulting z_i encodes interaction-aware behavior patterns rather than independent per-agent dynamics. The model is trained using a conditional ELBO objective:

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_q[\log p_\psi(\boldsymbol{\xi}_i \mid z_i)] - \beta D_{\text{KL}}(q_\phi(z_i \mid \boldsymbol{\xi}, c_{\text{env}}) \parallel p(z)), \tag{2}$$

with KL annealing for stable optimization.

At inference time, sampling z_i from the prior enables counterfactual behavior generation, and the latent is projected into the shared conditioning space for downstream control.

3.4 Backbone Integration and Closed-Loop Fine-Tuning

To enable controllable traffic simulation, the proposed adapters must be integrated into existing generative traffic models while preserving their learned generative priors. A key challenge is that the model operates in a closed-loop setting, where small prediction errors can accumulate over time and destabilize multi-agent interactions. We integrate control adapters into the generative cores of both diffusion [12] and autoregressive traffic models [22] while keeping backbone parameters frozen. In both cases, control is injected at the decoder level

to directly influence trajectory generation, systematically modulating the representations at every transformer decoder layer output across both the diffusion and autoregressive backbones. Only the control encoders and FiLM layers are updated during closed-loop fine-tuning [3, 24]. By leveraging FiLM-based feature modulation at the decoder level, the same control adaptation mechanism can be applied to different generative backbones without modifying their core architectures.

The key idea of our closed-loop fine-tuning is to bias sampled trajectories toward those aligned with the control target while maintaining stable multi-agent interactions. The exact procedure depends on the generative backbone.

Diffusion: Receding Horizon Control. Following a "plan–select–execute" paradigm [3], we generate M candidate rollouts at each replanning step and select the candidate closest to the ground-truth trajectory:

$$m^* = \underset{m}{\operatorname{argmin}} \mathcal{L}_{\text{match}}(\hat{\mathbf{S}}^{(m)}, \mathbf{S}^{gt}), \tag{3}$$

where $\hat{\mathbf{S}}^{(m)}$ denotes the m-th predicted joint trajectory and $\mathbf{S}^{gt}$ is the ground truth. The adapter parameters are updated using the selected rollout to compute the training loss during closed-loop fine-tuning.

Autoregressive: Closed-Loop Fine-Tuning. For the autoregressive (AR) model, trajectories are generated sequentially under control conditioning z_i. We optimize a cross-entropy objective under closed-loop unrolling:

$$\mathcal{L}_{\text{AR}} = -\frac{1}{NT} \sum_{i,t} \log \pi_\theta(s_{i,t}^{gt} \mid \hat{\mathbf{S}}_{\leq t}, z_i), \tag{4}$$

where $\hat{\mathbf{S}}_{\leq t}$ denotes the dynamically generated joint state history up to time t, where N is the number of agents and T is the prediction horizon. To reinforce controllability, we bias sampling toward trajectories [24] aligned with the conditioning signal, enabling the adapter to recover from its own prediction errors.

3.5 Context-Aware Scenario Transfer.

In traffic simulation, it is important to explore alternative futures or synthesize rare interactions that deviate from recorded data while remaining realistic. However, directly forcing arbitrary control signals often leads to implausible behaviors in the current traffic context. To address this, we introduce a **context-aware scenario transfer pipeline** that retrieves compatible agents from similar scenarios and transfers their control signals into the target traffic scene, enabling realistic variations within existing environments.

Specifically, we use *context embedding similarity* to retrieve compatible agents across scenarios. Given a target agent, we search the dataset for agents whose encoded scene context embeddings are similar (see Fig. 4). The retrieved agents serve as sources of plausible behaviors that can be transferred to the target scene.

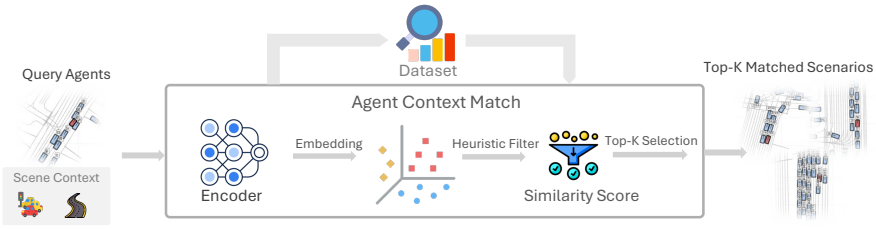

Fig. 4: Context-Match Retrieval Pipeline. Query agents are encoded into a shared embedding space. Following heuristic filtering for dynamic feasibility, candidates are ranked via similarity scoring to retrieve the Top-K environmentally compatible scenarios from the dataset.

Using this scenario transfer mechanism, we support two types of controllable scenario generation: **(i) Counterfactual Rollouts**, where we retrieve context-matched agents representing different driving intentions and transfer their behavior signals (e.g., sketch or latent trajectory) to the target scene to generate alternative futures, and **(ii) Long-Tail Synthesis**, where we identify rare behaviors using SMART likelihood and transfer their latent behavior codes to context-matched agents in new scenes to synthesize rare interactions.

4 Experiments

In this section, we demonstrate that our framework enables effective multi-modal controllability for both autoregressive and diffusion traffic models using only small amounts of paired data, and further show its practical utility for counterfactual traffic simulation.

4.1 Experimental Setup

Datasets and Benchmark. We evaluate our framework on the Waymo Open Motion Dataset (WOMD) [8], following the standard WOSAC [15] protocol for closed-loop traffic simulation.
Backbone Models. To demonstrate the architecture-agnostic nature of our framework, we evaluate two state-of-the-art generative traffic models as unconditional backbones: **VBD** [12], a diffusion-based model, and **SMART** [22], an autoregressive model.
Metrics. Following established practices in controllable generation [3, 20], we evaluate along two axes: **Realism**, measured by the WOSAC Meta score (↑), evaluates distributional realism using likelihood-based metrics across multiple aspects of driving behavior. **Controllability**, measured by the mean Average Displacement Error (mADE ↓) and relative mADE Gain (↑), which quantify alignment with the intended control signals. For counterfactual scenario synthesis (Sec. 4.4), we additionally report driving quality and safety metrics, including collision and off-road rates, as well as the closed-loop **PDMScore** [5].

Table 1: Quantitative Results by Modality on WOSAC. We evaluate our proposed framework against ProSim [20]. Grouping by control modality explicitly highlights our sample and parameter efficiency: despite utilizing **only 1%** of the data and freezing the backbone (updating merely 1.3∼2.1M parameters), our FiLM-based adapters (VBD and SMART) achieve significantly lower mADE and higher realism (Meta) compared to ProSim training on 100% of the dataset.

Backbone (Data Ratio)	Trainable Params	Meta ↑	Kinematic ↑	Interactive ↑	Map ↑	mADE ↓	mADE Gain ↑
Unconditional (Base Models)							
ProSim	11.5M	0.7035	0.4198	0.7249	0.8382	2.6787	–
VBD-CL	12.3M	0.7186	0.4181	0.7705	0.8237	2.7223	–
SMART-tiny-CLSFT	7.0M	**0.7728**	**0.4660**	**0.8057**	**0.9058**	**1.5044**	–
Sketch Control							
ProSim (100%)	–	0.7490	0.4527	0.7733	0.8872	1.0993	58.96%
VBD-CL (1%)	**1.3M**	0.7432	0.4720	0.7592	0.8776	0.9645	64.56%
SMART-tiny-CLSFT (1%)	2.1M	**0.7984**	**0.5305**	**0.8146**	**0.9307**	**0.2561**	**82.97%**
Latent Control							
VBD-CL (1%)	1.9M	0.7440	0.4625	0.7801	0.8584	0.9221	66.12%
SMART-tiny-CLSFT (1%)	**1.4M**	**0.7912**	**0.5227**	**0.8077**	**0.9234**	**0.3238**	**78.48%**
Text Control							
ProSim (100%)	–	0.7090	0.4251	0.7324	0.8412	2.3297	13.02%
VBD-CL (1%)	**1.9M**	0.7306	0.4270	0.7670	0.8572	2.0005	**26.51%**
SMART-tiny-CLSFT (1%)	2.0M	**0.7767**	**0.4751**	**0.8072**	**0.9098**	**1.3537**	10.02%

Control Signal Generation. To quantitatively evaluate controllability, we derive control signals from held-out ground truth (GT) trajectories: (i) **Sketch:** We downsample the GT future trajectories to obtain sparse waypoints as the sketch input; (ii) **Latent:** We process the GT trajectories through our pretrained BehaviorVAE encoder to extract latent codes $\mathbf{z}$; and (iii) **Text:** We utilize natural language descriptions of each agent's motion from the **ProSim-Instruct-520k** dataset [20]. We explicitly note that derived controls are used only for standardized quantitative evaluation (e.g., mADE), not as a deployment assumption. Practical non-oracle controllability is evaluated via retrieval-based sketch/latent/text transfer in Sec. 4.4.

4.2 Controllable Simulation Evaluation

Tab. 1 reports controllable simulation results across both diffusion (VBD) and autoregressive (SMART) traffic models. To contextualize our performance, we compare against ProSim [20], the closest prior method supporting sketch and text control. To establish robust backbones for controllable simulation, we first convert the open-loop VBD and SMART models into closed-loop simulators via additional training, denoted as **VBD-CL** and **SMART-tiny-CLSFT** [24].

Our adapters consistently improve controllability across both backbones while maintaining high realism. For spatially grounded signals (**Sketch** and **Latent**), VBD achieves roughly **65%** relative mADE reduction, while SMART reaches up to **80%**. Importantly, Meta scores remain comparable to or surpass the unconditional base models, confirming these gains do not degrade simulation realism.

Compared to ProSim [20], which is trained on the full dataset, our adapters achieve comparable or stronger performance while using only **1%** of the training data and keeping the backbone models frozen. While natural language inherently

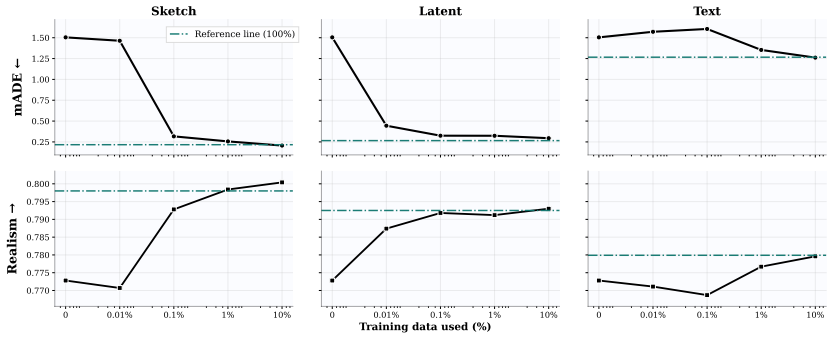

Fig. 5: Sample Efficiency on WOSAC. Evaluated using the autoregressive backbone. We report controllability (mADE ↓, top) and realism (Meta Score ↑, bottom) as a function of training data size. The unconditional base model corresponds to the 0% data point. Our adapters achieve strong controllability with minimal supervision, surpassing the base model with only **0.01%** (Latent) and **0.1%** (Sketch) data while maintaining comparable realism. Performance saturates around 10% data, matching a full-data LoRA finetuning.

provides sparser spatial grounding compared to explicit trajectories or behavior codes, it still yields consistent controllability gains and preserves stable closed-loop interactions, and better realism compared to ProSim. [3]

4.3 Sample Efficiency and Data Scaling

In this section, we focus on evaluating the **sample efficiency** of our proposed adaptation framework. Adapters are trained on logarithmically spaced subsets of the Waymo dataset (50 to 50,000 scenarios; 0.01% $\sim$ 10% data). Figure 5 illustrates how controllability (mADE ↓) and distributional realism (Meta Score ↑) evolve as the amount of training data increases. For reference, we additionally report the performance obtained by fine-tuning the base model with LoRA on the full 100% dataset using the same control adapters.

Data Efficiency Across Modalities. Our adapters demonstrate strong sample efficiency across control modalities. Even with extremely small amounts of paired data, they are able to learn effective controllable behaviors while preserving the generative prior of the frozen base model. Among the evaluated modalities, **Latent** control is the most data-efficient: with only **0.01%** of the training data, it already achieves performance close to the reference model trained with the full dataset. The **Sketch** modality also performs strongly under limited supervision, approaching the reference performance with merely **0.1%** of the data. In contrast, the **Text** modality exhibits a delayed scaling curve, requiring $\sim$1% data to outperform the baseline. Because natural language provides abstract guidance

[3] Results for ProSim are obtained from the official checkpoint released by the authors. Additional details are provided in the supplementary material.

Table 2: Counterfactual Rollouts. Control adapters enable diverse counterfactual rollouts while maintaining closed-loop driving quality. Sketch and latent controls achieve strong controllability (low Control ADE and high maneuver success), while all modalities increase trajectory diversity (Coverage) compared to the unconditional simulator without degrading safety metrics or PDMScore.

Method	Control ADE (m)	Success (%) ↑	Coverage ↑	Coll. ↓	Offroad ↓	PDMScore ↑
Unconditional	0.0000	–	201.02	0.0563	0.0252	78.40
Sketch CF	0.3727	70.61	**287.17**	0.0836	0.0252	79.30
Latent CF	0.4542	**73.89**	254.41	0.0774	0.0251	**80.33**
Text CF	2.4446	69.84	282.57	**0.0520**	**0.0227**	79.56

and relies on noisy auto-generated captions (e.g., ProSim-Instruct-520k), text control primarily serves as high-level intent biasing rather than precise trajectory supervision.

Realism Preservation and Saturation. Importantly, extremely low-data regimes do not compromise simulation realism. While Meta scores for Sketch and Text dip slightly at 0.01%, they rapidly recover and match the base model by 0.1% data, while Latent control improves realism from the start. Performance across all modalities saturates around 10% data, where Sketch, Latent and Text adapters achieve mADE comparable to the **reference performance** obtained by LoRA fine-tuning on the full 100% dataset (See Fig. 5). This suggests that effective controllable simulation can be achieved without large-scale paired annotations.

4.4 Counterfactual and Long-Tail Scenario Generation

Having established the sample efficiency of our control adaptation framework, we next evaluate its practical utility for scenario generation using the context-aware scenario transfer pipeline described in Sec. 3.5. Specifically, we generate counterfactual and long-tail scenarios within existing traffic scenes.

(i) **Counterfactual Rollouts:** We select 250 query agents and retrieve context-matched agents representing common driving intentions (*turning, lane changing, accelerating, decelerating, cruising, stopping*). For each query agent, we select up to four distinct intentions with the highest context-match scores and transfer the corresponding behavior annotations (*sketch* or *latent*) to the query agent in the original scene to generate counterfactual rollouts.

Evaluation Metrics. For counterfactual generation, we evaluate three aspects: (i) **controllability**, i.e., how well the transferred behavior controls the target agent, (ii) **behavioral diversity**, and (iii) **driving realism**. Controllability is measured using **Control ADE**, which computes the displacement error between the generated trajectory and the transferred control behavior (e.g., sketch or latent trajectory). We also report the maneuver **Success Rate**, defined as the fraction of rollouts that successfully execute the same high-level maneuver in each scenario. To evaluate diversity, **Coverage** measures the spatial distributional spread of generated trajectories, computed as the number of grid cells

Table 3: Long-Tail Scenario Generation. Context-aware retrieval significantly improves intent alignment and driving quality when transferring rare behaviors to new contexts. Compared to random context selection, context matching reduces collisions and improves closed-loop driving performance.

Strategy	Control ADE ↓	Traj ADE	Coll. ↓	Offroad ↓	PDMScore ↑
Context Random	3.8723	16.33	0.3988	0.0612	30.95
Context Match	**1.6994**	11.30	**0.2561**	**0.0393**	**51.67**

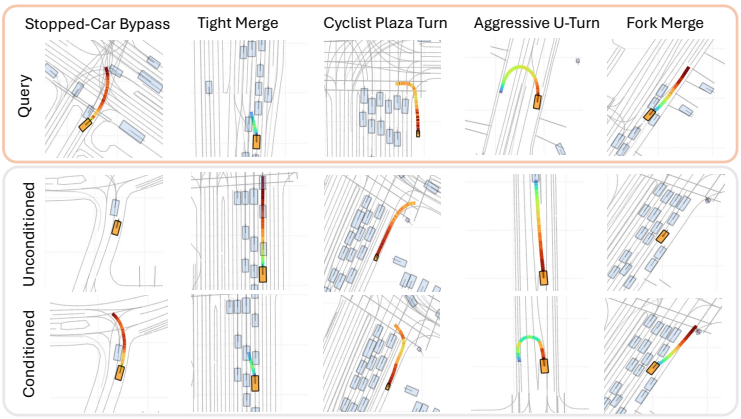

Fig. 6: Qualitative results of long-tail scenario generation. Top: Source scenarios providing the target long-tail behavioral latents. **Middle:** Default base model predictions in the retrieved matching contexts, showing only nominal driving. **Bottom:** Our framework successfully injects the queried intent, enabling agents to execute complex, safety-critical maneuvers realistically in novel contexts.

whose kernel density estimate (KDE) exceeds a predefined threshold [6]. Driving realism is evaluated using **Collision**, **Offroad**, and the closed-loop **PDMScore** [5]. We do not report WOSAC metrics in this setting, as they measure similarity to ground-truth trajectories, which is not appropriate for counterfactual generation where trajectories intentionally deviate from the recorded future.

Table 2 shows that our control adapters enable effective counterfactual generation across modalities. Sketch and latent controls achieve strong controllability, reflected by low Control ADE and high maneuver success rates (70%). At the same time, driving realism is preserved, with all modalities maintaining stable safety metrics and achieving PDM scores comparable to or higher than the unconditional baseline. Finally, the generated trajectories exhibit greater diversity: conditioning on different high-level behaviors increases trajectory coverage compared to the unconditional baseline, indicating that our framework can explore multiple plausible futures within the same scene (see supplementary material for qualitative results).

(ii) **Long-Tail Synthesis:** Beyond generating alternative futures within a scene, we also evaluate whether our framework can synthesize rare behaviors in new contexts. We use the SMART model's likelihood from [24] to identify the top

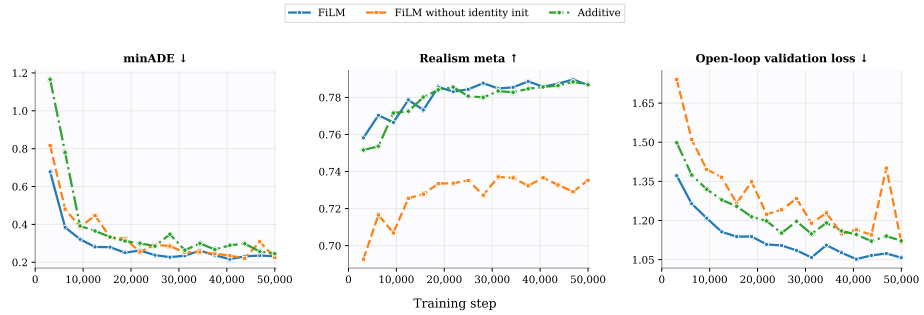

Fig. 7: Ablation on Modulation Mechanisms. Multiplicative FiLM (Blue) converges faster and achieves lower mADE than additive modulation (Green), reaching strong performance within fewer optimization steps. Identity initialization is crucial for preserving the base model's realism prior (Center) at the start of training.

0.1% hardest-to-predict scenarios in the dataset. From these scenarios, we curate eight long-tail behavior categories, including *plaza turning, meandering, threading stopped cars, multi-lane weaving, unsignalized crossing, fork-merge turning, tight merging,* and *aggressive U-turns*, as shown in Fig. 6.

Similar to the counterfactual setup, for each behavior, we retrieve the top 30 context-matched candidate agents and transfer the corresponding behavior latent $\mathbf{z}_{\text{target}}$ to the selected agent while leaving surrounding agents unconditioned. As a baseline, we compare against randomly selecting agents within the scene to receive the transferred behavior. Since long-tail behaviors intentionally deviate from nominal driving, we additionally report **Traj ADE** to quantify the extent to which the generated trajectory diverges from the original unconditional simulation.

Table 3 highlights the importance of context-aware transfer. Random context transfer often places the target behavior in incompatible environments, leading to significantly higher collision rates (0.3988) and degraded driving performance. In contrast, **Context Match** better aligns the injected behavior with compatible traffic contexts, reducing collisions (0.2561) and improving the PDMScore.

Nevertheless, transferring long-tail behaviors across scenarios remains inherently challenging, as these maneuvers (e.g., tight merges) are rare and highly constrained. We observe that many collision cases result from the trade-off between enforcing the conditioned long-tail behavior and preserving safe interactions with surrounding agents. Finally, the large trajectory deviation (**Traj ADE** = 11.30m) indicates that the model successfully overrides the nominal unconditional behavior to execute the targeted long-tail interactions.

4.5 Ablation Studies

We conduct additional ablations to validate our core design choices regarding adapter modulation design and conditioning strategies. Unless otherwise specified, models are trained on a 5,000-scenario subset (∼1% of WOMD) utilizing

Table 4: Static vs. Dynamic Control Conditioning. Latent control is highly effective under both conditioning strategies. Dynamic conditioning slightly improves control precision (mADE), while static conditioning maintains competitive performance and strong simulation realism.

Representation	Time Emb.	Meta ↑	Kinematic ↑	Interactive ↑	Map ↑	mADE ↓
Sketch Control Modality						
Dynamic	×	0.7799	0.5261	0.7921	0.9091	**0.3970**
Static	✓	0.7719	0.5083	0.7823	0.9091	1.0598
Latent Control Modality						
Dynamic	×	0.7848	0.5272	0.8027	0.9089	**0.3151**
Static	✓	0.7880	0.5113	0.8087	0.9195	0.4842

the autoregressive backbone to ensure rapid iteration while preserving statistical significance.

Adapter Modulation Design. We study the architectural design of the control adapter using the Sketch modality. Figure 7 compares our proposed **FiLM with Identity Initialization** against two baselines: FiLM with random initialization and additive modulation. The additive baseline injects control signals via bias addition, i.e., $\mathbf{h}^{(l)}_{\text{add}} = \mathbf{h}^{(l)} + \boldsymbol{\beta}^{(l)}(z_i)$.

As shown in the center plot of realism in Fig. 7, **identity initialization** is critical. Randomly initialized adapters significantly disrupt the pretrained model early in training, causing a severe drop in realism, whereas identity-initialized FiLM preserves the base model's behavior from the start. Moreover, FiLM is substantially more efficient: while additive modulation fails to converge after 50k steps (∼9 hours on a single RTX 4090), our FiLM adapter reaches better mADE in under 12k steps (∼2.2 hours). These results suggest that identity-initialized multiplicative modulation provides a stronger inductive bias for control-adapter fine-tuning.

Control Conditioning Strategy. We compare two conditioning strategies for Sketch and Latent signals: (i) **Static**, where the future trajectory is encoded once and injected with a time embedding across steps, and (ii) **Dynamic**, where the control signal is re-encoded at each timestep to follow the receding horizon.

Table 4 shows that **latent control is effective**, achieving strong controllability under both strategies while maintaining high realism (e.g., Meta = 0.7880, Map = 0.9195 for static latent). While dynamic conditioning achieves slightly lower mADE (0.3151 vs. 0.4842), **static conditioning remains competitive** and avoids enforcing strict timestep alignment in closed-loop interactions. This flexibility is desirable in closed-loop simulation, where agents must adapt their behavior in response to interactions with surrounding agents.

5 Conclusion

We present a data-efficient control adaptation framework that enables controllable traffic simulation across multiple conditioning modalities, including sketch,

latent behavior codes, and natural language. Built on top of pretrained generative traffic models, our lightweight adapters can be applied to both **autoregressive** and **diffusion-based** architectures without modifying the backbone models. Experiments show that our approach achieves strong controllability while preserving simulation realism using only **1% of the training data**. Beyond standard control evaluations, our framework enables **counterfactual scenario generation** and **long-tail behavior synthesis**, producing diverse yet realistic traffic interactions. Future work includes planner-in-the-loop evaluation and extending the framework to support vulnerable road user simulation.

References

1. Bergamini, L., Ye, Y., Scheel, O., Chen, L., Hu, C., Del Pero, L., Osiński, B., Grimmett, H., Ondruska, P.: Simnet: Learning reactive self-driving simulations from real-world observations. In: ICRA (2021)
2. Chang, W.J., Pittaluga, F., Tomizuka, M., Zhan, W., Chandraker, M.: Safe-sim: Safety-critical closed-loop traffic simulation with diffusion-controllable adversaries. In: ECCV (2024)
3. Chang, W.J., Zhan, W., Tomizuka, M., Chandraker, M., Pittaluga, F.: Langtraj: Diffusion model and dataset for language-conditioned trajectory simulation. In: ICCV (2025)
4. Chen, P.L., Kung, C.H., Chang, C.H., Chiu, W.C., Chen, Y.T.: Controllable collision scenario generation via collision pattern prediction. In: ICRA (2026)
5. Dauner, D., Hallgarten, M., Li, T., Weng, X., Huang, Z., Yang, Z., Li, H., Gilitschenski, I., Ivanovic, B., Pavone, M., Geiger, A., Chitta, K.: Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In: NeurIPS (2024)
6. Ding, W., Cao, Y., Ding Zhao, C.X., Pavone, M.: Bits: Bi-level imitation for traffic simulation. In: ICRA (2023)
7. Ding, W., Cao, Y., Zhao, D., Xiao, C., Pavone, M.: Realgen: Retrieval augmented generation for controllable traffic scenarios. In: ECCV (2024)
8. Ettinger, S., Cheng, S., Caine, B., Liu, C., Zhao, H., Pradhan, S., Chai, Y., Sapp, B., Qi, C.R., Zhou, Y., Yang, Z., Chouard, A., Sun, P., Ngiam, J., Vasudevan, V., McCauley, A., Shlens, J., Anguelov, D.: Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In: ICCV (2021)
9. Feng, L., Li, Q., Peng, Z., Tan, S., Zhou, B.: Trafficgen: Learning to generate diverse and realistic traffic scenarios. In: ICRA (2023)
10. Gulino, C., Fu, J., Luo, W., Tucker, G., Bronstein, E., Lu, Y., Harb, J., Pan, X., Wang, Y., Chen, X., Co-Reyes, J.D., Agarwal, R., Roelofs, R., Lu, Y., Montali, N., Mougin, P., Yang, Z., White, B., Faust, A., McAllister, R., Anguelov, D., Sapp, B.: Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research. In: NeurIPS (2023)
11. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022)
12. Huang, Z., Zhang, Z., Vaidya, A., Chen, Y., Lv, C., Fisac, J.F.: Versatile behavior diffusion for generalized traffic agent simulation. arXiv preprint arXiv:2404.02524 (2024)

13. Igl, M., Kim, D., Kuefler, A., Mougin, P., Shah, P., Shiarlis, K., Anguelov, D., Palatucci, M., White, B., Whiteson, S.: Symphony: Learning realistic and diverse agents for autonomous driving simulation. In: ICRA (2022)
14. Jiang, C.M., Bai, Y., et al.: Scenediffuser: Efficient and controllable driving simulation initialization and rollout. In: NeurIPS (2024)
15. Montali, N., Lambert, J., Mougin, P., Boone, A., Boulton, P., Lu, Y., Devin, C., Huguet, R., Dasari, J., Sapp, B., et al.: The waymo open sim agents challenge. In: NeurIPS (2023)
16. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.C.: Film: Visual reasoning with a general conditioning layer. In: AAAI (2018)
17. Rempe, D., Philion, J., Guibas, L.J., Fidler, S., Litany, O.: Generating useful accident-prone driving scenarios via a learned traffic prior. In: CVPR (2022)
18. Rowe, L., Girgis, R., Gosselin, A., Paull, L., Pal, C., Heide, F.: Scenario dreamer: Vectorized latent diffusion for generating driving simulation environments. In: CVPR (2025)
19. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
20. Tan, S., Ivanovic, B., Chen, Y., Li, B., Weng, X., Cao, Y., Krähenbühl, P., Pavone, M.: Promptable closed-loop traffic simulation. In: CoRL (2024)
21. Tan, S., Ivanovic, B., Weng, X., Pavone, M., Krähenbühl, P.: Language conditioned traffic generation. In: CoRL (2023)
22. Wu, W., Feng, X., Gao, Z., Kan, Y.: Smart: Scalable multi-agent real-time motion generation via next-token prediction. In: NeurIPS (2024)
23. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)
24. Zhang, Z., Karkus, P., Igl, M., Ding, W., Chen, Y., Ivanovic, B., Pavone, M.: Closed-loop supervised fine-tuning of tokenized traffic models. In: CVPR (2025)
25. Zhao, T., Zhao, L., Eskenazi, M., Black, A.W.: Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In: ACL (2017)
26. Zhong, Z., Rempe, D., Chen, Y., Ivanovic, B., Cao, Y., Xu, D., Pavone, M., Ray, B.: Language-guided traffic simulation via scene-level diffusion. In: CoRL (2023)
27. Zhong, Z., Rempe, D., Xu, D., Chen, Y., Veer, S., Che, T., Ray, B., Pavone, M.: Guided conditional diffusion for controllable traffic simulation. In: ICRA (2023)
28. Zhou, Z., Haibo, H., Chen, X., Wang, J., Guan, N., Wu, K., Li, Y.H., Huang, Y.K., Xue, C.J.: Behaviorgpt: Smart agent simulation for autonomous driving with next-patch prediction. In: NeurIPS (2024)