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

Controllable Motion Generation via Diffusion Modal Coupling

Luobin Wang*,1, Hongzhan Yu*,1, Chenning Yu[1], Sicun Gao[1], Henrik Iskov Christensen[1]

***Abstract*— Diffusion models are increasingly used in robotics to represent multi-modal distributions over system states and behaviors, but precise control of generated outcomes without degrading physical realism remains challenging. This paper introduces a controllable diffusion framework that (i) replaces the standard unimodal Gaussian prior with an explicit multi-modal prior, and (ii) enforces modal coupling between prior components and principal data modes through novel forward and reverse diffusion processes. Sampling is initialized directly from a selected prior mode aligned with task constraints, avoiding train–test mismatch and manifold drift commonly induced by post-hoc guidance. Empirical evaluations on motion prediction (Waymo Dataset) and multi-task control (Maze2D) show consistent improvements over guidance-based baselines in fidelity, diversity, and controllability. These results indicate that multi-modal priors with strong modal coupling provide a scalable basis for controllable motion generation in robotics. The official implementation is provided in https://github.com/RobinWangSD/Diffusion-Modal-Coupling/.**

I. INTRODUCTION

Diffusion models [1] provide expressive, high-fidelity generative capabilities well suited to robotics, where uncertainty and multi-modality are intrinsic. Recent applications span sensor simulation [2], [3], [4], trajectory generation [5], [6], and control policy learning [7], [8], [9], illustrating the utility of diffusion for modeling complex, high-dimensional behavior distributions. Despite this progress, controllability, i.e., the ability to produce samples that satisfy task objectives and domain constraints while remaining on the data manifold, remains a central obstacle. Unconstrained sampling can produce implausible plans or policies, thereby undermining the intended driving objectives [5]. Due to limited controllability, practitioners resort to generating large numbers of samples to capture desirable outcomes, which raises scalability concerns and complicates the mining of high-fidelity samples for large-scale synthetic generations.

Existing approaches typically inject constraints at inference, via constraint-based sampling or post-hoc guidance, to encode domain knowledge and task objectives into the reverse process [10], [11],. While such mechanisms improve alignment with target objectives, they introduce a train-test mismatch: external gradients modify the learned score field during sampling, perturbing the reverse dynamics and steering intermidiate states away from high-density regions of the data distribution. The resulting off-manifold drift degrades sample fidelity. Achieving fine-grained controllability without sacrificing data-manifold realism therefore remains an open problem (Figure 1 [a-c]).

*Equal Contribution, [1]University of California, San Diego

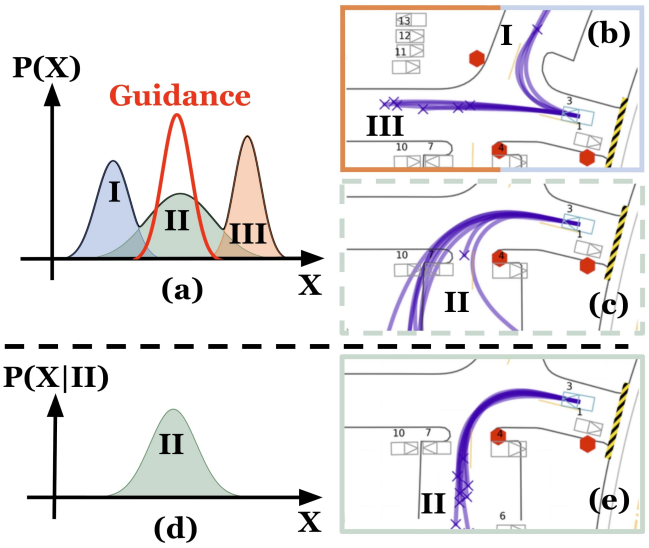

Fig. 1: High-level comparison of guidance-based approaches versus our proposed method. **(a)** A standard diffusion model fits a multi-modal data distribution (three modes). A guidance term (red) attempts to steer sampling toward a rare yet operationally critical mode (Mode II). **(b)** Standard (unguided) sampling concentrates on high-probability modes. **(c)** Guidance perturbs intermediate states off the well-trained data manifold, degrading fidelity. **(d)(e)** Our method couples each principal mode to a dedicated prior component, enabling direct, mode-aligned control at sampling while avoiding guidance-induced distribution mismatch.

We address this problem by introducing modal coupling via a multi-modal prior. Concretely, we replace the standard Gaussian prior with a Gaussian-mixture prior whose components are in one-to-one correspondence with principal modes of the data distribution (Figure 1 [d-e]). We derive modified forward and reverse diffusion processes that (i) map all data samples sharing the same label to a common, non-standard Gaussian during noising and (ii) recover the corresponding data mode when denoising is initialized from the associated prior component. At inference, controllability is obtained by selecting the prior mode consistent with the target constraints and running standard denoising, without post-hoc guidance. By removing external guidance terms, our method eliminates train–test mismatch and mitigates intermediate-step distribution drift, while retaining flexible controllability (Fig. 2).

Our current formulation assumes that the data modes subject to control are explicitly known, yet the diffusion model itself does not require conditioning on these mode labels. While this may appear restrictive, it provides a foundation for strong controllability over data with unknown

979-8-3315-8160-2/26/$31.00 ©2026 IEEE

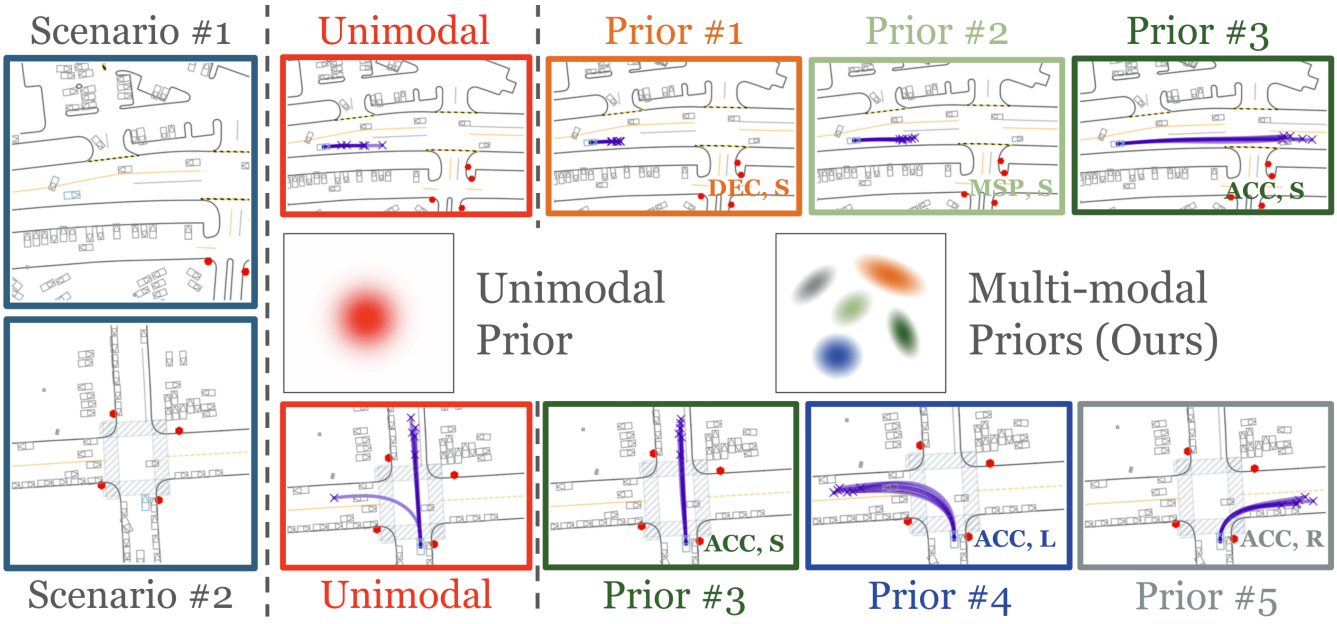

Fig. 2: High-level overview. Conventional diffusion models use a unimodal prior distribution and lack an intrinsic mechanism to select which trajectories are emphasized. We introduce a multi-modal prior and enforce strong modal coupling between prior and data via a novel diffusion process. The framework enables direct mode selection even with an unconditioned diffusion model, supporting precise and adaptive motion generation. In the figure, each prior component corresponds to one behavior. "ACC", "DEC" and "MSP" refer to speed modes (acceleration, deceleration, and maintaining speed), while "R", "L" and "S" represent steering modes (right, left, and straight).

key modes, wherein the central challenge shifts to accurately identifying the appropriate prior mode from target constraints. Nevertheless, by adopting a *multi-modal* prior distribution, strong *modal coupling*, and a carefully-designed *prior parametrization*, our method significantly outperforms guidance-based techniques in both fidelity and controllability. We validate these claims on the Waymo dataset for motion prediction, and in Maze2D for multi-task control. The paper is organized as follows. Section II and III review related work and background. We detail the proposed method in Section IV. Section V presents the experimental results, and Section VI concludes.

II. Related Work

The multi-modal nature of human behaviors poses a great challenge for predicting realistic trajectories and control sequences. Diffusion models have proven effective in capturing this multi-modality within driving scenarios while closely adhering to real-world behavior distributions. SceneDM [12] utilizes a diffusion-based framework to model joint-distributions of all agents in a scene. SceneDiffuser [13] employs a latent diffusion architecture derived from Bird's Eye View representations, whereas MotionDiffuser [5] demonstrates its capabilities of predicting realistic future trajectories that align with true data distribution via PCA-compressed trajectory representations. Additionally, VBD [14] jointly optimizes a motion predictor and a denoiser that share the same scene encoder, which further improves realism and versatility. However, achieving such realism and broad distribution coverage often requires drawing many random samples from a standard Gaussian prior that is unimodal. Our approach enhances the realism of generated trajectories by incorporating a multi-modal prior that more effectively captures distinct data modes.

The typical strategy to control diffusion-based generation is incorporating domain-specific objectives into the generation process. Classifier guidance [15] guides the diffusion model with a separate cost function that encodes the objectives during sampling. Recent works [5], [16], [17], [18] introduce either language-model–generated or analytical guidance functions to pursue realism or safety-critical objectives.

On the other hand, Classifier-free guidance [10] additionally optimizes a time-dependent conditional model to obtain guidance. This is widely adopted in other fields including text-to-image [19] and 3D objection [20] generation. However, the constraints imposed by guidance often degrade the realism of generated motion [16] due to distribution mismatch between denoising steps and corresponding diffusion forward steps. Our approach avoids this issue by coupling prior and data modes and constructing a shared probability path for both forward and reverse diffusion.

III. Background

Diffusion models are probabilistic generative models that synthesize new data by iteratively denoising an initial noise sample. They first define a forward stochastic process that progressively perturbs real data x_0 into pure noise:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}), \tag{1}$$

$$q(x_t|x_{t-1}) = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon_t, \tag{2}$$

where $x_{1:T}$ denotes the sequence of noised samples, ϵ_t is standard Gaussian noise, and β_t is forward variance. Let $\alpha_t := 1-\beta_t$ and $\bar{\alpha}_t := \prod_{i=1}^{t} \alpha_i$. The above forward process yields the closed-form marginal:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I). \tag{3}$$

Choosing a variance schedule $\beta_{1:T}$ with $\bar{\alpha}_T \to 0$ ensures the prior $x_T \sim \mathcal{N}(0, I)$, i.e., a standard Gaussian distribution.

The reverse process removes noise step by step via Gaussian transitions. Concretely,

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t), \sigma_t^2 I), \tag{4}$$

$$\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t, \tag{5}$$

where θ are the denoiser parameters trained by maximizing the evidence lower bound [1]: $\max_\theta \ -\log p_\theta(x_0|x_1) + \sum_t D_{KL}[q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)]$. A common reparameterization predicts noise:

$$\mu_\theta(x_t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)), \tag{6}$$

where ϵ_θ approximates the score $\nabla_{x_t} \log q(x_t)$ across noise scales. Alternatively, predicting the clean sample x_θ^0 gives the closed-form posterior mean:

$$\mu_\theta(x_t) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_\theta^0(x_t, t). \tag{7}$$

Controllability denotes the ability to enforce the specified constraints on generated samples. A common approach is guidance [10] which augments the learned score at each reverse step with a task-specific cost f:

$$\hat{\epsilon}_\theta(x_t, t) = \epsilon_\theta(x_t, t) + \omega\nabla_{x_t} f(x_t), \tag{8}$$

where ω controls the strength of the guidance. Here, f is flexible: it can encode hand-designed differentiable penalties (e.g., speed limits, action bounds, or label/style preferences) or be realized implicitly via a conditional score model, as in the popular classifier-free guidance approaches [21].

IV. Enhanced Diffusion Controllability via Modal Coupling

A common strategy for controlling diffusion-based generation is guidance. However, guidance is absent from diffusion model training and is applied only at sampling as a post-hoc modification (8). This train–test mismatch can induce distribution shift, pushing intermediate states off the high-fidelity data manifold. To address this, we propose adopting a multi-modal prior and pose the question:

> *if each prior mode is tightly coupled to a principal data mode at training, can we run denoising process from different prior modes for direct controllability without guidance?*

We show the answer is yes: by modifying the diffusion processes to accommodate a carefully-parameterized Gaussian-mixture prior, we can initialize the denoising/reverse process from a selected prior mode to target specific data constraints. Importantly, this eliminates reliance on post-hoc modifications, avoiding guidance-induced distribution mismatch.

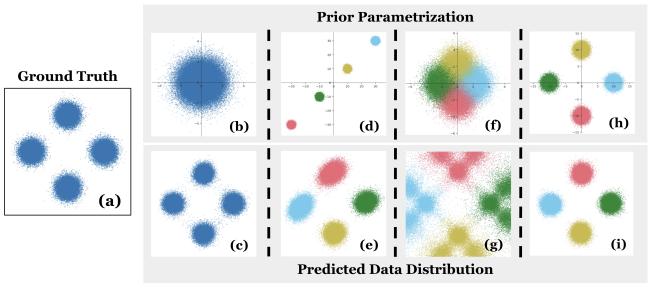

Fig. 3: 2D toy example. (a) The data distribution with four distinct modes. (b-c) Results from DDPM [1] show using unimodal prior yields spurious samples in the gaps between modes. (d-e) When prior means have large magnitude (i.e., lie far from the origin), the diffusion model struggles to recover realistic per-mode data distributions. (f-g) Insufficient separation between prior modes also prevents the model from accurately capturing the data distribution. (h-i) With a carefully designed prior parameterization that maintains clear separation between modes without introducing excessive values, our method produces substantially fewer spurious samples and further enables direct control over individual modes. Corresponding modes and samples share the same color.

A. Gaussian Mixture as a Multi-Model Prior

We assume data modes are explicitly known, and model the multi-modal prior as a Gaussian mixture model:

$$x_T \sim \sum_{i=1}^{k} r_i \cdot \mathcal{N}(\mu_i, \sigma_i^2 I), \tag{9}$$

where k is the number of modes, r_i is the proportion of data with mode label i, and μ_i, σ_i^2 are the mean and variance of the component i. Conditioned on a specific label L, the prior reduces to a unimodal Gaussian:

$$x_T|x_0 \sim \mathcal{N}(\mu_L, \sigma_L^2 I), \tag{10}$$

allowing us to explicitly account for different data modes while retaining a unimodal form given a specific label.

B. Modal Coupling

Effective training with a Gaussian mixture prior requires two conditions. First **(modal coupling)**: each prior mode must correspond one-to-one with a data mode. For label L, the forward process maps data with label L to $\mathcal{N}(\mu_L, \sigma_L^2 I)$, while initializing the reverse process from that specific prior mode recovers the data of label L. Second **(trajectory separation)**: throughout denoising, diffusion paths from distinct prior modes must remain sufficiently separated to preserve mode identity and avoid cross-mode interference. This section focuses on modal coupling, while the discussion of prior parameterization for trajectory separation is in Section IV-C.

We first define a forward noising process such that the terminal distribution satisfies a general Gaussian prior $x_T|x_0 \sim \mathcal{N}(\mu, \sigma^2 I)$.

Lemma 1. *Let* $\eta_t := 1 + \sum_{m=1}^{t-1}\left(\sqrt{\prod_{n=m+1}^{t}\alpha_n}\right)$, *and consider the forward noising process*

$$q(x_t|x_{t-1}) = \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\sigma\epsilon_t + \frac{\mu}{\eta_T}. \tag{11}$$

where $\epsilon_t \sim \mathcal{N}(0, I)$. Then, for any step t,

$$q(x_t|x_0) = \mathcal{N}(x_t|\sqrt{\bar{\alpha}_t}x_0 + \frac{\eta_t \mu}{\eta_T}, (1-\bar{\alpha}_t)\sigma^2 I). \quad (12)$$

Under the standard assumption that $\bar{\alpha}_T \to 0$ as T grows large, it follows that $q(x_T|x_0) = \mathcal{N}(x_T; \mu, \sigma^2 I)$.

Proof. See Appendix VI-A. It shows that a *constant* shift term μ/η_T at each forward step and noise scaling by σ yield the desired terminal prior, enabling a matched reverse process.

Lemma 2. *For the diffusion model with the forward process defined in (11), the reverse process is:*

$$p(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}|\mu(x_t), \beta(x_t)), \quad (13)$$

where $\beta(x_t) = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t\sigma^2$, and

$$\mu(x_t) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\bar{\alpha}_t}\underline{\mathbf{x_0}} \quad (14)$$
$$+ \frac{\eta_{t-1}(1-\alpha_t) - \sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)\eta_T} \cdot \mu,$$
$$= \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\mu}{\eta_T} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\sigma\underline{\boldsymbol{\epsilon_t}}). \quad (15)$$

Proof. See Appendix VI-B. This completes the derivation of the modified forward and reverse processes, making each data mode directly associated with a non-standard Gaussian prior component. We now introduce the training objective. Given a labeled dataset $\mathcal{X}$ and pre-defined prior parameters $\{\mu_{1:k}, \sigma_{1:k}\}$, we construct noisy samples via (12) and train the clean-prediction model x_θ^0 with:

$$\min_\theta \underset{\substack{t\in[1,T]\\ \epsilon\sim N(0,I)}}{E}\left[\sum_{(x_0,L)\in\mathcal{X}} \|x_\theta^0(\hat{x}_t(x_0, L, \epsilon), t) - x_0\|^2\right], \quad (16)$$

$$\hat{x}_t(x_0, L, \epsilon) = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\sigma_L\epsilon + \frac{\eta_t\mu_L}{\eta_T}. \quad (17)$$

Switching the reparameterization of μ_θ to noise prediction is straightforward but omitted for performance considerations.

At sampling, we can draw the prior sample x_T from the mixture (9) to cover all modes, or fix a component (10) to target specific constraints.

C. Prior Parametrization

Ensuring trajectory separation is crucial for reliable mode identification, which in turn underpins controllability. Without it, reverse trajectories from different prior components may overlap, causing mode indistinguishability and loss of control (Figure 3). We address this by carefully parameterizing the Gaussian mixture prior.

To place the Gaussian means, we draw on the concept of placing *evenly spaced* points on a high-dimensional sphere. Let d be the data dimension, and k the number of modes. If $k \le d+1$, which is common in complex data distributions specialized by diffusion models, the means can be located at the vertices of a $(k-1)$-simplex embedded in $\mathbb{R}^d$ [22]. Let $e_{1:k} \subset \mathbb{R}^k$ be the standard basis vectors and $\mathbb{1}_k$ the all-ones vector. With sphere radius $\delta > 0$, define

$$w_i = \delta \cdot \sqrt{\frac{k}{k+1}} \cdot (e_i - \mathbb{1}_k \cdot \frac{1}{k}) \quad (18)$$

$$\mu_i = [w_i^1, ..., w_i^k, 0, ..., 0]^T \in \mathbb{R}^d, \quad (19)$$

where w_i^j denotes the j-th component of w_i. One can verify that $\|\mu_i\|^2 = \delta$ for all i, and the pairwise distance d is:

$$d = \delta \cdot \sqrt{2 + 2/(k-1)}. \quad (20)$$

Finally, we choose the Gaussian variances so that c-level confidence ellipsoids do not overlap. That is, for all $i \in [1, k]$,

$$\sigma_i\sqrt{\mathcal{X}_{d,c}^2} \le d, \quad (21)$$

where $\mathcal{X}_{d,c}^2$ is the c quantile of the chi-square distribution with d degress of freedom [23].

V. Experiments

We begin with the Waymo dataset, showing that our approach produces realistic and feasible motions while preserving controllability, in contrast to post-hoc guidance baselines. We then show that our method extends naturally to a multi-task setting by treating each task as a distinct mode. Unlike prior diffusion-based planners that train a separate model per task [24], [25], a single model suffices for our approach while maintaining strong per-task performance.

A. Waymo - Controllable Motion Prediction

We evaluate on the Waymo Open Motion Dataset [26], under single-agent setting with explicitly-defined mode structure. Following [27], ground-truth actions L are deterministically labeled along two independent axes: steering (left, right, straight) and speed (accelerate, decelerate, maintain speed) with a differentiable classifier $L = g(x)$. The task is to generate an 8-s ego control trajectory consistent with the queried action label while respecting map topology and contextual agent interactions. We predict 2-D controls over 80 future steps (dimension 160), yielding $k = 9$ action modes from the Cartesian product of the two axes.

Versatile Behavior Diffusion (VBD) [14] is adopted as the model backbone. From VBD, we reuse: (i) the query-centric Transformer encoder for scene-level interaction encoding, and (ii) diffusion timesteps embeddings for in-context conditioning. Our instantiation departs from VBD in two ways: (1) we flatten the control sequence as a single vector representation [18], and (2) during denoising, we apply cross-attention from the ego control-sequence embeddings to the scene context to capture ego-environment interactions.

We compare against two guidance baselines. **Classifier Guidance** (*CG*) applies the deterministic classifiers (used for labeling) to yield gradients with respect to the unconditioned score network [16]. **Classifier-Free Guidance** (*CFG*) trains additional conditional models with shared parameters [21].

Method		minADE (↓)	minFDE (↓)	Collision (↓)	OffRoad (↓)	ACC[ST] (↑)	ACC[SP] (↑)	ACC (↑)	Inference Time (↓)
CG	$\eta = 1$	2.614	5.999	0.054	0.127	0.882	0.968	0.746	2.717
	$\eta = 10$	2.862	6.655	0.057	0.160	0.916	0.987	0.902	2.703
	$\eta = 100$	4.681	12.358	0.059	0.275	**0.976**	0.996	**0.970**	2.712
CFG	$\lambda = 0.9$	3.994	10.194	0.070	0.162	0.878	0.833	0.746	1.535
	$\lambda = 1$	3.034	7.138	0.069	0.159	0.877	0.977	0.864	1.566
	$\lambda = 1.05$	2.687	6.055	0.065	0.157	0.885	0.996	0.883	1.544
	$\lambda = 1.1$	3.099	7.506	0.062	0.184	0.887	**1.000**	0.887	1.547
Ours	$\delta = 0.5$	2.506	5.104	0.053	0.121	0.895	0.999	0.894	0.899
	$\delta = 1.0$	**2.266**	**4.757**	**0.043**	**0.092**	0.914	**1.000**	0.914	0.892
	$\delta = 2.0$	2.525	5.303	0.489	0.101	0.907	0.998	0.904	0.906
	$\delta = 4.0$	2.772	5.697	0.060	0.135	0.895	0.993	0.888	**0.864**
	EM	2.578	5.500	0.060	0.140	0.897	0.995	0.893	0.891

TABLE I: Quantitative results on the Waymo Open Motion Dataset under reference driving modes. Driving modes comprise steering and speed categories. minADE/minFDE are averaged across steering–speed modes, mitigating the effect of imbalanced mode frequencies.

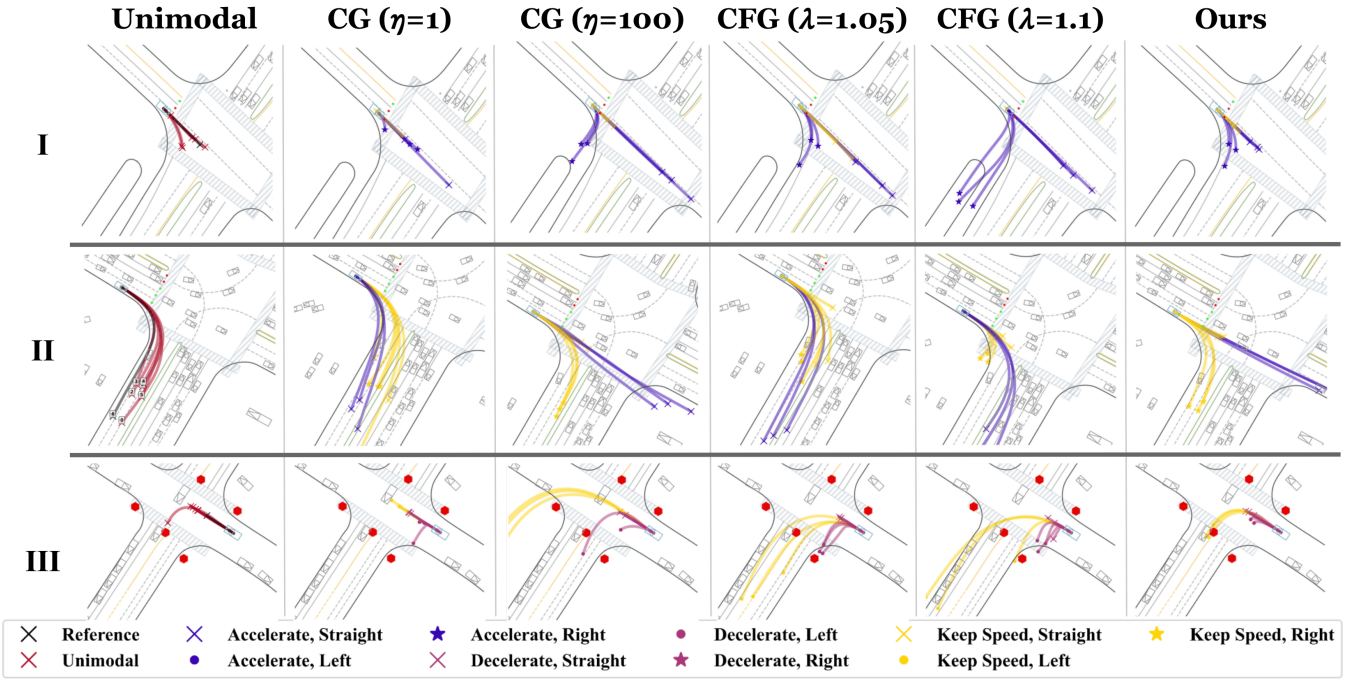

Fig. 4: Qualitative results for motion prediction. In standard diffusion, trajectories are sampled randomly from a unimodal prior, offering no inherent controllability. CG applies guidance to steer generation, but it relies heavily on the guidance influence factor, making it difficult to balance sample fidelity against controllability. CFG blends unconditional and conditional outputs for guidance, which limits controllability and fidelity when the target lies off the reference data manifold. Our method integrates modal coupling with a multi-modal prior distribution, yielding notable improvements in both sample fidelity and controllability.

Both methods apply guidance in a post-hoc manner, with standard formulations given as follows:

$$\textbf{CG} \quad \hat{\epsilon}_\theta(x_t, t) = \epsilon_\theta(x_t, t) + \eta \nabla_{x_t} g(\mu_\theta(x_t)),$$
$$\textbf{CFG} \quad \hat{\epsilon}_\theta(x_t, t) = \epsilon_\theta(x_t, t) + \lambda(\epsilon_\theta(x_t, t, L) - \epsilon_\theta(x_t, t)),$$

Unless stated otherwise, we set per-mode scales $\sigma_i = 1$ for $i \in \{1, \ldots, k\}$. We also evaluate an **empirical-means** (*EM*) initialization that sets the multi-modal prior's means and standard deviations from training-set statistics.

Next, we describe the evaluation metrics used in our study. We report *Average Displacement Error (ADE)*, i.e., the mean Euclidean distance between predicted and ground-truth positions, and *Final Displacement Error (FDE)*, i.e., the terminal position error. To assess physical plausibility, we compute *Collision* and *Off-Road* rates[28], defined as the fraction of rollouts that intersect another agent or leave the drivable area, respectively. We introduce accuracy (*ACC*) statistics to

Method		Collision (↓)	OffRoad (↑)	ACC (↑)
CG	$\eta = 1$	0.050	0.125	0.595
	$\eta = 10$	0.060	0.135	0.830
	$\eta = 100$	0.060	0.205	**0.960**
CFG	$\lambda = 0.9$	0.080	0.215	0.725
	$\lambda = 1$	0.075	0.195	0.875
	$\lambda = 1.05$	0.060	0.185	0.900
	$\lambda = 1.1$	0.060	0.195	0.910
Ours	$\delta = 0.5$	0.060	0.115	0.915
	$\delta = 1.0$	**0.040**	**0.110**	0.930
	$\delta = 2.0$	0.055	0.120	0.920
	$\delta = 4.0$	0.070	0.140	0.925
	EM	0.060	0.180	0.905

TABLE II: Quantitative validation for trajectory synthesis controllability, using 8 sampled trajectories per scenario. The test set consists of 200 scenarios, each manually labeled with potentially feasible ego-agent futures that deviate from the dataset's original trajectories.

evaluate control fidelity, while further decomposing it into per-axis statistics: *ACC[ST]* (steering) and *ACC[SP]* (speed).

1) Motion Prediction Accuracy: Table I summarizes the prediction accuracy results, evaluated against ground-truth labels of the ego-agent's future behaviors. While the focus of this work is on controllability, we demonstrate that coupling diffusion model training on explicitly labeled datasets with a multi-modal prior distribution substantially improves predictive performance.

Comparison to guidance baselines. For both guidance approaches (CG, CFG), the guidance coefficient (η for CG; λ for CFG) is sensitive to tuning and enforces a control-fidelity trade-off. Since these methods operate with a unimodal prior, they offer limited mode-specific learning and coverage. Even at their best settings, our method outperforms them: a mode-coupled prior allocates capacity per mode and enables direct, label-aligned control, yielding more robust predictions. Computationally, CG requires per-step backpropagation through the diffusion model, while CFG doubles forward passes. Both introduce higher inference cost than our single reverse process from the selected prior component.

Ablation on δ and EM prior. Setting $\delta = 1.0$ consistently outperforms $\delta = 0.5$, improving inter-component separation in the prior and reducing overlap among reverse-diffusion trajectories. Excessively large δ degrades performance: as in the 2-D toy example (Figure 3), components displaced far from the origin yield implausible trajectory distributions. Under the EM configuration, steering means concentrate near zero because left/right-turn trajectories contain extended straight segments. Speed statistics are less skewed but still exhibit partial mode collapse. Although EM preserves cross-mode action-sequence correlations, the increased component overlap reduces separability and harms accuracy.

Method		U-Maze	Medium	Large
Diffuser	U-Maze	113.9	N/A	N/A
	Medium	N/A	**121.5**	N/A
	Large	N/A	N/A	**123.0**
Ours, $k = 3$, $\delta = 30$		**119.5**	121.4	120.9

TABLE III: Quantitative evaluations on Maze2D. We adopt the baseline performance from [29], using normalized accumulative reward returns as the evaluation metric. Notably, the baseline trains a separate diffusion model for each layout (or mode), unlike in Table I where our comparisons focus on multi-modal data modeling and the baseline there is a single model handling various modes.

2) Controllable Trajectory Synthesis: Table II evaluates realism and feasibility when generation is conditioned on driving modes that deviate from the dataset reference. We manually annotate feasible future modes: e.g., if the ego is in a lane that permits turning while the reference continue straight, the turn is marked feasible. The evaluation set contains 200 scenarios and will be released with the code. Visualizations of generated trajectories using various controllability methods on manually labeled scenarios are presented in Figure 4.

Under out-of-distribution conditioning, our method continues to produce feasible, on-road trajectories, as reflected in the low collision and off-road rates. CG/CFG often fail to realize off-reference modes with realistic generations: e.g., intersection scenes frequently go off-road after crossing. While achieving superior sample realism, the proposed method also maintains stable, high accuracy. These results underscore the limitations of guidance-based control (particularly post-hoc guidance) that our approach is designed to overcome.

B. Maze2d - Multi-task Control

We extend our method method to multi-task control. Rather than confining control-level constraints to a single task, we can also view each task itself as a distinct data mode. We illustrate this with evaluations in Maze2D, where a single, task-agnostic model is trained to perform long-horizon path planning across three maze configurations.

We define the diffusion model to predict 384-step state-control trajectories conditioned on the initial and goal positions. With 3-D states and 2-D controls, the target distribution is 1920-dimensional. We set $\sigma_i = 1$ for $i \in [1, 3]$, and choose $\delta = 30$. As shown in Table III, the unified model matches baseline performance across all tested mazes. Note that the goal here is not to surpass the baseline, but to demonstrate that casting tasks as modes establishes modal coupling over a multi-modal prior, enabling diverse tasks to be handled by a single model without task-specific conditioning achitectures.

VI. Conclusion

This paper presents a novel framework that enables fine-grained control over diffusion models while preserving high-fidelity sample generation. By aligning the sampling process

with key data modes from the outset, our method avoids the distribution drift common in post-hoc guidance approaches. Experimental evaluations show that the proposed method consistently outperforms existing techniques on both quantitative and qualitative measures. Moreover, this work lays a strong foundation for future research aimed at relaxing the assumption of explicit known data modes, thereby advancing towards more controllable diffusion models.

APPENDIX

A. Proof of Lemma 1

The proof for (12) proceeds by induction. We begin with the base case using the proposed forward process (11):

$$x_1 = \sqrt{\alpha_1}x_0 + \sqrt{1-\alpha_1}\sigma\epsilon_1 + \frac{\mu}{\eta_T} \tag{22}$$

$$\sim \mathcal{N}(\sqrt{\bar{\alpha}_1}x_0 + \frac{\eta_1\mu}{\eta_T}, (1-\bar{\alpha}_1)\sigma^2 I) \tag{23}$$

The derivation from (22) to (23) is based on the fact that $\alpha_1 = \bar{\alpha}_1$ and $\eta_1 = 1$ by definition. Next, we assume that for an arbitrary $t \in [2, T]$, it holds true that:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}x_0 + \sqrt{1-\bar{\alpha}_{t-1}}\sigma\epsilon_{t-1} + \frac{\eta_{t-1}\mu}{\eta_T}. \tag{24}$$

From the forward process (11), we have:

$$x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\sigma\epsilon_t + \frac{\mu}{\eta_T} \tag{25}$$

$$= \sqrt{\alpha_t}\Big(\sqrt{\bar{\alpha}_{t-1}}x_0 + \sqrt{1-\bar{\alpha}_{t-1}}\sigma\epsilon_{t-1} + \frac{\eta_{t-1}\mu}{\eta_T}\Big) \tag{26}$$

$$+ \sqrt{1-\alpha_t}\sigma\epsilon_t + \frac{\mu}{\eta_T}$$

$$= \sqrt{\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}\eta_{t-1}+1}{\eta_T}\cdot\mu \tag{27}$$

$$+ \Big(\sqrt{\alpha_t(1-\bar{\alpha}_{t-1})}\sigma\epsilon_{t-1} + \sqrt{1-\alpha_t}\sigma\epsilon_t\Big).$$

Note that by the definition of η_t:

$$\eta_t = 1 + \sum_{m=1}^{t-1}\left(\sqrt{\prod_{n=m+1}^{t}\alpha_n}\right) \tag{28}$$

$$= 1 + \sum_{m=1}^{t-1}\left(\sqrt{\prod_{n=m+1}^{t-1}\alpha_n}\cdot\sqrt{\alpha_t}\right) \tag{29}$$

$$= 1 + \sqrt{\alpha_t}\cdot\Big[1 + \sum_{m=1}^{t-2}\left(\sqrt{\prod_{n=m+1}^{t-1}\alpha_n}\right)\Big] \tag{30}$$

$$= 1 + \sqrt{\alpha_t}\eta_{t-1}. \tag{31}$$

Meanwhile, to handle the last term in (27), we essentially merge two zero-mean Gaussian distributions with distinct variances. That is, merging $\mathcal{N}(0, \sigma_1^2 I)$ and $\mathcal{N}(0, \sigma_2^2 I)$ leads to the new distribution $\mathcal{N}(0, (\sigma_1^2 + \sigma_2^2)I)$. Here, the merged standard deviation is:

$$\alpha_t(1-\bar{\alpha}_{t-1})\sigma^2 + (1-\alpha_t)\sigma^2 = (1-\bar{\alpha}_t)\sigma^2. \tag{32}$$

Substituting everything back into (27), we have:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \frac{\eta_t}{\eta_T}\mu + \sqrt{1-\bar{\alpha}_t}\sigma\epsilon^*, \tag{33}$$

where ϵ^* denotes an arbitrary standard Gaussian sample. This concludes the proof of Lemma 1.

B. Proof of Lemma 2

First, the reverse probability is tractable only when conditioned on x_0. By Bayes' theorem, we have:

$$p(x_{t-1}|x_t, x_0) = q(x_t|x_{t-1}, x_0)\cdot\frac{q(x_{t-1}|x_0)}{q(x_t|x_0)}. \tag{34}$$

Then by Lemma 1:

$$p(x_{t-1}|x_t, x_0) \tag{35}$$

$$\propto \exp(-\frac{1}{2}\Big[\frac{(x_t - \sqrt{\alpha_t}x_{t-1} - \mu/\eta_T)^2}{(1-\alpha_t)\sigma^2} \tag{36}$$

$$+ \frac{(x_{t-1} - \sqrt{\bar{\alpha}_{t-1}}x_0 - (\eta_{t-1}/\eta_T)\mu)^2}{(1-\bar{\alpha}_{t-1})\sigma^2}$$

$$- \frac{(x_t - \sqrt{\bar{\alpha}_t}x_0 - (\eta_t/\eta_T)\mu)^2}{(1-\bar{\alpha}_t)\sigma^2}\Big])$$

$$\propto \exp(-\frac{1}{2}\Big[\Big(\frac{\alpha_t}{(1-\alpha_t)\sigma^2} + \frac{1}{(1-\bar{\alpha}_{t-1}\sigma^2)}\Big)x_{t-1}^2 \tag{37}$$

$$+ 2\Big(\frac{-\sqrt{\alpha_t}x_t + \sqrt{\alpha_t}\mu/\eta_T}{(1-\alpha_t)\sigma^2}$$

$$+ \frac{-\sqrt{\bar{\alpha}_{t-1}}x_0 - (\eta_{t-1}/\eta_T)\mu}{(1-\bar{\alpha}_{t-1})\sigma^2}\Big)x_{t-1}\Big])$$

From (36) to (37), the constant terms that do not involve x_{t-1} are all omitted. Following the standard Gaussian density function, the mean and variance of $p(x_{t-1}|x_t, x_0)$ can be parameterized as $\mathcal{N}(\tilde{\mu}, \tilde{\beta})$ where:

$$\tilde{\beta} = 1/\Big(\frac{\alpha_t}{(1-\alpha_t)\sigma^2} + \frac{1}{(1-\bar{\alpha}_{t-1}\sigma^2)}\Big) \tag{38}$$

$$= \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\cdot(1-\alpha_t)\sigma^2. \tag{39}$$

Then we derive $\tilde{\mu}$ as follows:

$$\tilde{\mu}(x_t, x_0) = -\Big(\frac{-\sqrt{\alpha_t}x_t + \sqrt{\alpha_t}\mu/\eta_T}{(1-\alpha_t)\sigma^2} \tag{40}$$

$$+ \frac{-\sqrt{\bar{\alpha}_{t-1}}x_0 - (\eta_{t-1}/\eta_T)\mu}{(1-\bar{\alpha}_{t-1})\sigma^2}\Big)\cdot\tilde{\beta}_t$$

$$= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}x_0 \tag{41}$$

$$+ \frac{\eta_{t-1}(1-\alpha_t) - \sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)\eta_T}\cdot\mu$$

Furthermore, we can parameterize x_0 in terms of x_t and ϵ_t based on (33):

$$\tilde{\mu}_t(x_t) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t \tag{42}$$
$$+ \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t} \cdot \frac{x_t - \sqrt{1-\bar{\alpha}_t}\sigma\epsilon_t - (\eta_t/\eta_T)\mu}{\sqrt{\bar{\alpha}_t}}$$
$$+ \frac{\eta_{t-1}(1-\alpha_t) - \sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)\eta_T} \cdot \mu$$
$$= \frac{1}{\sqrt{\alpha_t}}\Big(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\sigma\epsilon_t\Big) \tag{43}$$
$$+ \frac{\mu}{(1-\bar{\alpha}_t)\eta_T} \cdot \Big(\eta_{t-1}(1-\alpha_t)$$
$$- \sqrt{\alpha_t}(1-\bar{\alpha}_{t-1}) - \frac{(1-\alpha_t)\eta_t}{\sqrt{\alpha_t}}\Big)$$
$$= \frac{1}{\sqrt{\alpha_t}}\Big(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\sigma\epsilon_t\Big) \tag{44}$$
$$+ \frac{\mu}{(1-\bar{\alpha}_t)\eta_T} \cdot \frac{-(1-\bar{\alpha}_t)}{\sqrt{\alpha_t}}$$
$$= \frac{1}{\sqrt{\alpha_t}}\Big(x_t - \frac{\mu}{\eta_T} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\sigma\epsilon_t\Big) \tag{45}$$

Last three equations are due to $\eta_t = 1 + \sqrt{\alpha_t}\eta_{t-1}$.

REFERENCES

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[2] Haoxi Ran, Vitor Guizilini, and Yue Wang. Towards realistic scene generation with lidar diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14738–14748, 2024.

[3] Anthony Zhou, Zijie Li, Michael Schneier, John R Buchanan Jr, and Amir Barati Farimani. Text2pde: Latent diffusion models for accessible physics simulation. *arXiv preprint arXiv:2410.01153*, 2024.

[4] Luobin Wang, Runlin Guo, Quan Vuong, Yuzhe Qin, Hao Su, and Henrik Christensen. A real2sim2real method for robust object grasping with neural surface reconstruction. In *2023 IEEE 19th International Conference on Automation Science and Engineering (CASE)*, pages 1–8, 2023.

[5] Chiyu Jiang, Andre Cornman, Cheolho Park, Benjamin Sapp, Yin Zhou, Dragomir Anguelov, et al. Motiondiffuser: Controllable multi-agent motion prediction using diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9644–9653, 2023.

[6] Brian Yang, Huangyuan Su, Nikolaos Gkanatsios, Tsung-Wei Ke, Ayush Jain, Jeff Schneider, and Katerina Fragkiadaki. Diffusion-es: Gradient-free planning with diffusion for autonomous driving and zero-shot instruction following. *arXiv preprint arXiv:2402.06559*, 2024.

[7] Guangyao Zhou, Sivaramakrishnan Swaminathan, Rajkumar Vasudeva Raju, J Swaroop Guntupalli, Wolfgang Lehrach, Joseph Ortiz, Antoine Dedieu, Miguel Lázaro-Gredilla, and Kevin Murphy. Diffusion model predictive control. *arXiv preprint arXiv:2410.05364*, 2024.

[8] Hongzhan Yu, Chiaki Hirayama, Chenning Yu, Sylvia Herbert, and Sicun Gao. Sequential neural barriers for scalable dynamic obstacle avoidance. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11241–11248, 2023.

[9] Chenning Yu, Hongzhan Yu, and Sicun Gao. Learning control admissibility models with graph neural networks for multi-agent navigation. In Karen Liu, Dana Kulic, and Jeff Ichnowski, editors, *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 934–945. PMLR, 14–18 Dec 2023.

[10] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[11] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852, 2023.

[12] Zhiming Guo, Xing Gao, Jianlan Zhou, Xinyu Cai, and Botian Shi. Scenedm: Scene-level multi-agent trajectory generation with consistent diffusion models. *arXiv preprint arXiv:2311.15736*, 2023.

[13] Ethan Pronovost, Meghana Reddy Ganesina, Noureldin Hendy, Zeyu Wang, Andres Morales, Kai Wang, and Nick Roy. Scenario diffusion: Controllable driving scenario generation with diffusion. *Advances in Neural Information Processing Systems*, 36:68873–68894, 2023.

[14] Zhiyu Huang, Zixu Zhang, Ameya Vaidya, Yuxiao Chen, Chen Lv, and Jaime Fernández Fisac. Versatile behavior diffusion for generalized traffic agent simulation. *arXiv preprint arXiv:2404.02524*, 2024.

[15] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[16] Ziyuan Zhong, Davis Rempe, Danfei Xu, Yuxiao Chen, Sushant Veer, Tong Che, Baishakhi Ray, and Marco Pavone. Guided conditional diffusion for controllable traffic simulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3560–3566, 2023.

[17] Brian Yang, Huangyuan Su, Nikolaos Gkanatsios, Tsung-Wei Ke, Ayush Jain, Jeff Schneider, and Katerina Fragkiadaki. Diffusion-es: Gradient-free planning with diffusion for autonomous driving and zero-shot instruction following, 2024.

[18] Yinan Zheng, Ruiming Liang, Kexin ZHENG, Jinliang Zheng, Liyuan Mao, Jianxiong Li, Weihao Gu, Rui Ai, Shengbo Eben Li, Xianyuan Zhan, and Jingjing Liu. Diffusion-based planning for autonomous driving with flexible guidance. In *The Thirteenth International Conference on Learning Representations*, 2025.

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[20] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9298–9309, October 2023.

[21] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[22] Harold Scott Macdonald Coxeter. *Regular polytopes*. Courier Corporation, 1973.

[23] Edwin B Wilson and Margaret M Hilferty. The distribution of chi-square. *Proceedings of the National Academy of Sciences*, 17(12):684–688, 1931.

[24] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[25] Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning*, 2022.

[26] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9710–9719, 2021.

[27] Shuhan Tan, Boris Ivanovic, Xinshuo Weng, Marco Pavone, and Philipp Kraehenbuehl. Language conditioned traffic generation. In *7th Annual Conference on Robot Learning*, 2023.

[28] Nico Montali, John Lambert, Paul Mougin, Alex Kuefler, Nicholas Rhinehart, Michelle Li, Cole Gulino, Tristan Emrich, Zoey Yang, Shimon Whiteson, et al. The waymo open sim agents challenge. *Advances in Neural Information Processing Systems*, 36:59151–59171, 2023.

[29] Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022.