# Compositional Diffusion Models for Autonomous Drone Racing

Fangguo Zhao[1,†], Zhouheng Li[1,†], Yunhao Luo[2], Hanbing Zhang[1], Xin Guan[1], Lei Xie[1], and Shuo Li[1]
[1]Zhejiang University [2]University of Michigan
[†]Contributed equally to this work.

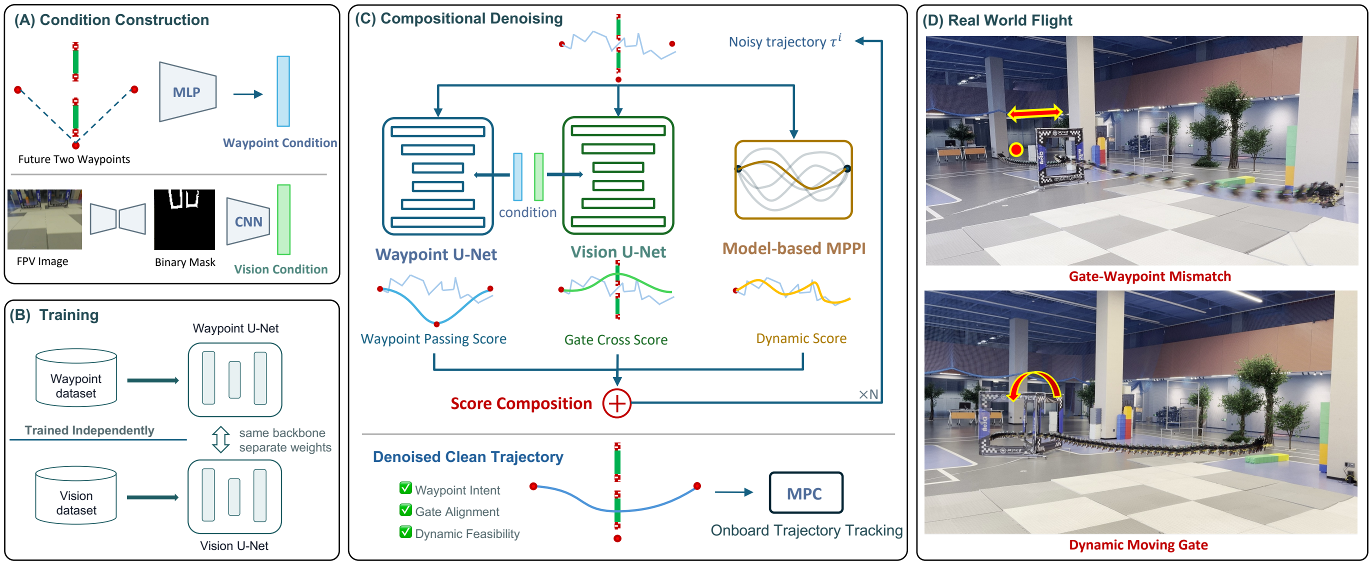


Fig. 1. **Compositional diffusion planning for autonomous drone racing.** (A) Runtime construction of waypoints and vision conditions from future waypoints and onboard FPV observations. (B) The two diffusion models use the same U-Net backbone but are independently trained with separate datasets and weights. (C) During compositional denoising, the noisy trajectory $\tau^i$ is jointly guided by a waypoint-conditioned U-Net, a vision-conditioned U-Net, and a model-based MPPI module, which provide waypoint-passing, gate-crossing, and dynamic-feasibility scores, respectively. The scores are composed at each denoising step and applied iteratively for $N$ steps. (D) The real-world flight trajectory of two challenging racing tasks.

***Abstract*—Autonomous drone racing probes the limits of agile flight by coupling high-speed perception, planning, and control. However, existing systems often depend on fixed track layouts, known gate poses, predefined trajectories, or track-specific training. These assumptions limit zero-shot deployment on real-world tracks with unseen geometries, gate offsets, orientations, and motion. We present a real-time multimodal, physics-informed energy-based generative planner for flexible-track racing without known global gate poses. Specifically, the planner decomposes racing into two complementary constraints. Waypoints encode global topological progress, while onboard vision provides local geometric guidance for gate traversal. We train waypoint- and vision-conditioned energy-based diffusion planners independently, and compose them online at the score level. An MPPI-derived model-based score further guides denoising with dynamic feasibility costs. This guidance biases sampled paths toward executable quadrotor motion without offline trajectory optimization. In simulation and real-world flights, the planner transfers to unseen tracks under a single, tuning-free planning configuration and supports high-agility flight through moving gates. Score-level composition balances rapid waypoint progress with precise visual gate interaction. It also uses waypoint intent to disambiguate visually symmetric multi-gate scenes. Model-based guidance improves the consistency and executability of sampled trajectories under aggressive flight. These results indicate that independently trained modal skills can be composed into a transferable racing planner with multimodal decision-making and dynamics-aware execution.**

## I. Introduction

Autonomous quadrotors have reached highly agile flight performance, enabling tasks that were previously difficult for small aerial robots Hanover et al. [14]. Autonomous drone racing is a representative benchmark for this capability. It requires a quadrotor to perceive the scene, plan ahead, and execute aggressive motion at high speed. In a race, the drone must follow an ordered route while passing through a sequence of gates. Waypoints define the global progress of the track. Gates impose local geometric constraints on where and how the vehicle should pass. Recent systems have approached, and in some cases surpassed, champion-level human performance on known and fixed racing tracks Song et al. [32], Kaufmann et al. [17]. However, fast lap time on a known course is only one part of racing ability.

A skilled human pilot can adapt to a new track after limited familiarization. The pilot reuses familiar subskills, such as straights, turns, and gate approaches, and connects them according to the new layout. During flight, first-person visual feedback helps the pilot adjust each gate approach. At the same time, the pilot uses an internal understanding of the downstream route to prepare the next maneuver Pfeiffer and Scaramuzza [27]. This suggests a harder autonomy target. A racing drone should not depend only on a fixed reference trajectory or a policy specialized to one course. It should convert route intent and onboard visual observation into dynamically feasible motion on unseen waypoint–gate layouts.

Existing racing systems do not fully provide this capability. Many methods rely on predefined waypoints, known gate poses, offline trajectory optimization, or track-specific training. These assumptions make the problem tractable, but they reduce flexibility. They also become restrictive when gates appear in new positions, with new orientations, or with motion. Vision-based systems reduce some dependence on prior maps, but aggressive first-person flight introduces blur, occlusion, partial observations, and fast viewpoint changes. These effects make it difficult to rely only on explicit gate pose estimation or on a single end-to-end visual policy.

The key challenge is therefore not only perception, planning, or control in isolation. It is the online coordination of heterogeneous information. Waypoints provide route-level intent and temporal progress. Onboard vision provides local geometric evidence for the current gate. Dynamics determine whether a candidate path can be executed by the quadrotor. A general racing planner must combine these sources during flight while remaining fast enough for agile motion.

We propose a real-time multimodal energy-based diffusion planning framework for zero-shot autonomous drone racing, as shown in Fig. 1. The framework decomposes racing into two task-level diffusion experts. A waypoint-conditioned expert captures global route progress through future waypoints. A vision-conditioned expert captures local gate constraints from onboard observations. The two experts are trained independently, so each modality can use the data source best suited to its role. During flight, their scores are composed online during denoising. This allows route guidance and visual gate correction to act on the same candidate trajectory.

To improve physical feasibility, we further add model-based diffusion guidance using Model Predictive Path Integral (MPPI) rollouts. This guidance injects dynamic costs into the denoising process. It regularizes composed samples that may fall outside the training distribution, while preserving the learned task priors. The resulting planner combines global waypoint progress, local visual gate interaction, and dynamics-aware execution in a single real-time planning process.

The main contributions of this paper are as follows:

1) We formulate zero-shot waypoint–gate drone racing as a compositional trajectory-generation problem. The framework does not rely on predefined reference trajectories, known global gate poses, or track-specific training and fine-tuning.
2) We introduce a multimodal energy-based diffusion planner that composes separately trained waypoint- and vision-conditioned models at the score level. This design enables online fusion of heterogeneous modalities and allows the planner to combine long-horizon route progress with local gate alignment on unseen layouts and gate-moving environments.
3) We develop an MPPI-guided model-based denoising mechanism that adds dynamic-feasibility scores during reverse diffusion. This mechanism regularizes out-of-distribution composed samples without replacing the learned task prior.
4) We evaluate the framework on unseen tracks with diverse gate configurations, including moving and multiple gates. The results demonstrate that compositional diffusion planning improves flexible and task-valid racing by coupling waypoint progress, visual gate interaction, and physics-aware execution.

## II. Related Work

### A. Drone Racing

*1) Trajectory Optimization and Reference-based Drone Racing:* Many autonomous racing systems formulate the task as reference-trajectory tracking Hanover et al. [14]. Polynomial trajectory generation produces smooth paths through predefined waypoints Mellinger and Kumar [25], Loianno et al. [22], Wang et al. [35], while nonlinear optimization and point-mass planning can compute time-optimal references when the track is known Foehn et al. [11], Shen et al. [31], Romero et al. [30]. These methods are accurate under fixed geometry, but they often optimize a surrogate tracking objective rather than the racing task itself Song et al. [32].

Learning-based planners and environment-conditioned controllers reduce this dependence on fixed references Zhou et al. [41], Wang et al. [34]. Sampling-based optimal control has also enabled reference-free multi-waypoint agile flight Zhao et al. [38]. However, these methods still simplify racing as waypoint traversal and do not explicitly model gate-level visual interaction, which is needed for unseen gates under onboard perception.

*2) Gate-aware and Vision-based Racing:* Gate-aware planning methods incorporate gate geometry through constraints at gate locations Qin et al. [28], Wang et al. [35], Han et al. [13], racing corridors Krinner et al. [18], or analytical spatial guidance Zhao et al. [39]. These methods improve local gate interaction, but they usually assume that gate poses, track geometry, or desired passing states are known, static, or reliably recoverable. Vision-based racing reduces the need for external track information by estimating gate poses from onboard images using geometric solvers or learned models Falanga et al. [10], Li et al. [19, 20], Kaufmann et al. [17], Bosello et al. [3], Azhari et al. [1], Bahnam et al. [2]. Such systems can achieve strong monocular racing performance when the environment is well represented by training or mapping Bahnam et al. [2]. However, motion blur, occlusion, partial gate observations, and rapid viewpoint changes make explicit pose estimation brittle during aggressive first-person flight. End-to-end visual and depth-based policies avoid some geometric assumptions by mapping visual input directly to actions or planning outputs Kaufmann et al. [16], Geles et al. [12], Xing et al. [36], Yu et al. [37], Zhao et al. [40]. This improves robustness in some settings, but the learned behavior may still depend on layouts seen during training. Generalizable monocular racing on unseen waypoint–gate configurations therefore remains challenging.

### B. Compositional Diffusion Model and Model-based Guidance

Diffusion policies have shown strong potential for robot action generation and motion planning Chi et al. [8]. They can represent multimodal solution distributions and provide a natural mechanism for composition. In energy-based diffusion planning, score fields can be composed during sampling using classifier-free guidance Du et al. [9]. Each score can encode a constraint, objective, or subtask, allowing denoising to revise candidate trajectories under multiple guidance signals.

Compositional diffusion has been explored in image generation Liu et al. [21] and path generation Luo et al. [23], Mao et al. [24]. Recent diffusion policies for manipulation also show cross-modal conditioning and generalization capabilities Cao et al. [5], Chen et al. [7]. These results suggest that diffusion models are suitable for planning problems requiring online coordination between multiple guidance sources Urain et al. [33].

Drone racing, however, poses a less aligned composition problem. In many prior settings, conditional models are trained on the same trajectory distribution or tightly coupled datasets Chen et al. [7]. Their conditions differ, but their trajectory support is largely shared. In our setting, the waypoint prior represents route progress, while the vision prior represents local gate interaction. These priors can come from heterogeneous data and may not be directly compatible when composed.

Dynamic feasibility further complicates this composition. Diffusion planners are learned from data and do not guarantee that generated trajectories satisfy system dynamics. Under score composition, this issue can be amplified because the combined score may push samples outside the support of either expert. Prior work addresses this problem by adding optimization or projection during diffusion Bouvier et al. [4], Römer et al. [29], or by connecting sampling-based optimal control with diffusion through model-based diffusion Pan et al. [26] and JM2D Jung et al. [15]. Our method follows this direction but targets compositional drone racing: it composes independently trained waypoint and vision experts at inference time, and adds MPPI-based guidance to bias the denoising process toward executable quadrotor motion.

## III. Method

### A. Vision–Waypoint Multimodal Compositional Planning

We formulate the trajectory planning of drone racing as conditional trajectory sampling over a horizon-$H$ state sequence $\boldsymbol{\tau}_{1:H} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_H]$, where $H = 40$ and $\boldsymbol{x}_t = [p_x, p_y, p_z, v_x, v_y, v_z, q_w, q_x, q_y, q_z]^\top \in \mathbb{R}^n$ which is the position, velocity and orientation states. The planner is conditioned jointly on a visual observation $o_v$ and a waypoint-level goal $g$. Instead of learning a single joint model $p_\theta(\boldsymbol{\tau}_{1:H} \mid o_v, g)$, we train two experts separately: a vision-conditioned planner $p_v(\boldsymbol{\tau}_{1:H} \mid o_v)$ and a waypoint-conditioned planner $p_g(\boldsymbol{\tau}_{1:H} \mid g)$. This factorization reduces the dependence on large paired image–waypoint datasets and makes the two modalities easier to compose at test time.

### B. Energy-parameterized Unimodal Diffusion Experts.

Our method uses a diffusion style model for each planner. For either modality, the clean trajectory $\boldsymbol{\tau}^0$ is corrupted by the forward diffusion process

$$\boldsymbol{\tau}^i = \sqrt{\bar{\alpha}_i}\boldsymbol{\tau}^0 + \sqrt{1-\bar{\alpha}_i}\boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \tag{1}$$

where $i$ is the reverse-diffusion index, $\alpha_i$ is the noise coefficient and $\bar{\alpha}_i = \prod_{j=1}^{i} \alpha_j$. The denoising backbone is a one-dimensional conditional U-Net over the trajectory horizon, with sinusoidal diffusion-step embeddings and FiLM-modulated residual blocks adopted from Chi et al. [8]. The gate observation $o_v$ and the future two-waypoint $g$ states are encoded into a condition vector. The two experts use separate condition encoders.

Rather than predicting noise directly, each expert is parameterized as an energy-based denoiser. Given a condition $c \in \{o_v, g\}$, the network outputs an intermediate vector field $F_\theta(\boldsymbol{\tau}^i, i, c)$, from which we define the scalar energy

$$E_\theta(\boldsymbol{\tau}^i, i, c) = \frac{1}{2}\left\|F_\theta(\boldsymbol{\tau}^i, i, c)\right\|_2^2. \tag{2}$$

The predicted noise is the gradient of this energy scalar,

$$\boldsymbol{\varepsilon}_\theta(\boldsymbol{\tau}^i, i, c) = \nabla_{\boldsymbol{\tau}^i} E_\theta(\boldsymbol{\tau}^i, i, c), \tag{3}$$

which makes the denoiser a conservative vector field induced by a scalar potential Chao et al. [6]. The corresponding score is therefore proportional to $-\boldsymbol{\varepsilon}_\theta(\boldsymbol{\tau}^i, i, c)/\sqrt{1-\bar{\alpha}_i}$. The vector $F_\theta$ is not used as the denoising direction directly; only the gradient of its induced scalar energy is used.

### C. Training Objective.

Each expert is trained independently on their own condition type and datasets. For every minibatch, the clean trajectory is normalized, a random diffusion step $i$ is sampled, and Gaussian noise is injected according to the forward process (1). The denoiser predicts $\boldsymbol{\varepsilon}_\theta(\boldsymbol{\tau}^i, i, c)$, and the diffusion objective is the masked mean-squared error

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\boldsymbol{\tau}^0, \boldsymbol{\varepsilon}, i}\left[\left\|\left(\boldsymbol{\varepsilon}_\theta(\boldsymbol{\tau}^i, i, c) - \boldsymbol{\varepsilon}\right)\right\|_2^2\right], \tag{4}$$

We additionally reconstruct the corresponding condition, which yields the full training loss

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \mathcal{L}_{\text{rec}}. \tag{5}$$

For the vision expert, $\mathcal{L}_{\text{rec}}$ reconstructs the input binary mask image; for the waypoint expert, it reconstructs the waypoint condition. This auxiliary term regularizes the modality encoder, while $\mathcal{L}_{\text{diff}}$ trains the energy-gradient denoiser itself.

*D. Compositional Inference.*

At test time, the two experts are evaluated on the same noisy trajectory $\boldsymbol{\tau}^i$. Each expert is queried with both a null condition and its modality-specific condition, yielding unconditional and conditional denoising vectors. The composed denoising direction is

$$\boldsymbol{\varepsilon}^i_{\text{comp}} = \boldsymbol{\varepsilon}^i_{\emptyset} + \omega_g \left(\boldsymbol{\varepsilon}^i_g - \boldsymbol{\varepsilon}^i_{\emptyset}\right) + \omega_v \left(\boldsymbol{\varepsilon}^i_v - \boldsymbol{\varepsilon}^i_{\emptyset}\right), \tag{6}$$

where $\omega_g$ and $\omega_v$ are the waypoint and vision guidance weights. This composition can be interpreted as a product-of-experts-style score combination. When the guidance weights are nonnegative and the base distribution is shared, the corresponding density takes the following tempered product form.

$$\begin{aligned} p_{\text{comp}}(\boldsymbol{\tau}_{1:H} \mid o_v, g) \propto\ & p_{\emptyset}(\boldsymbol{\tau}_{1:H})^{1-\omega_g-\omega_v} \\ & \times p_g(\boldsymbol{\tau}_{1:H} \mid g)^{\omega_g} p_v(\boldsymbol{\tau}_{1:H} \mid o_v)^{\omega_v}. \end{aligned} \tag{7}$$

The weights can be adjusted flexibly during inference. The composed denoising vector is then passed to a DDIM scheduler for update,

$$\boldsymbol{\tau}^{i-1} = \text{DDIMStep}\left(\boldsymbol{\tau}^i, \boldsymbol{\varepsilon}^i_{\text{comp}}, i\right), \tag{8}$$

which yields a task-consistent proposal trajectory before model-based feasibility refinement.

*E. Model-Based Feasibility Guidance with MPPI*

Although the composed denoising direction $\boldsymbol{\varepsilon}^i_{\text{comp}}$ provides a task-consistent trajectory update, the resulting clean estimate may still be inconsistent with quadrotor dynamics. We therefore introduce a dynamics-aware guidance term based on MPPI rollouts. The role of this module is not to replace the learned diffusion prior, but to convert the current diffusion prediction into a nearby dynamically realizable trajectory and inject it back as an additional denoising score.

At reverse diffusion step $i$, the composed denoiser first predicts the clean trajectory

$$\hat{\boldsymbol{\tau}}^0_i = \frac{\boldsymbol{\tau}^i - \sqrt{1-\bar{\alpha}_i}\boldsymbol{\varepsilon}^i_{\text{comp}}}{\sqrt{\bar{\alpha}_i}}. \tag{9}$$

This trajectory is unnormalized and used as a reference for model-based rollout evaluation. The first state of $\hat{\boldsymbol{\tau}}^0_i$ defines the rollout initial state, and the remaining states define the tracking reference for the MPPI horizon.

We sample $N$ action sequences from a Gaussian proposal

$$\mathbf{u}^{n,i}_{1:H} \sim q^u_i(\mathbf{u}_{1:H}) = \mathcal{N}\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}^u_i\right), \qquad n = 1, \ldots, N, \tag{10}$$

where $\boldsymbol{\mu}_i$ is the current MPPI action-sequence mean. The covariance of the action proposal is tied to a diffusion-style noise schedule defined in the normalized action space inspired by model based diffusion Pan et al. [26]. Specifically, we define

$$\alpha^u_j = 1 - \beta^u_j, \qquad \bar{\alpha}^u_i = \prod_{j=1}^{i} \alpha^u_j, \tag{11}$$

where $\beta^u_j$ is a prescribed action-space noise schedule. The proposal covariance is then set to

$$\boldsymbol{\Sigma}^u_i = \sigma^2_i \mathbf{I}, \qquad \sigma^2_i = 1 - \bar{\alpha}^u_i. \tag{12}$$

This schedule is used only to modulate the exploration scale of MPPI rollouts. Early reverse steps use a broader action proposal, while later steps use a more concentrated proposal around the updated action mean.

Each sampled action sequence is rolled out through the quadrotor dynamics. Here, a trajectory $\boldsymbol{\tau}$ denotes a sequence of states, i.e., $\boldsymbol{\tau} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_H]$, and $\boldsymbol{x}_t$ denotes the state at time step $t$. For brevity, we omit the diffusion step index $i$ in the rollout dynamics. The rollout dynamics are

$$\boldsymbol{x}^n_{t+1} = f(\boldsymbol{x}^n_t, \boldsymbol{u}^n_t), \qquad t = 0, \ldots, H-1. \tag{13}$$

The rollout cost is computed as a state-tracking objective with respect to the diffusion reference trajectory $\hat{\boldsymbol{\tau}}^0_i$:

$$J^i_n = \sum_{t=0}^{H-1} \ell\left(\boldsymbol{x}^n_{t+1}, \hat{\boldsymbol{x}}^0_{i,t+1}\right), \tag{14}$$

where

$$\ell(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \|\boldsymbol{p} - \hat{\boldsymbol{p}}\|^2_{Q_p} + \|\boldsymbol{v} - \hat{\boldsymbol{v}}\|^2_{Q_v} + \|\boldsymbol{e}_q(\boldsymbol{q}, \hat{\boldsymbol{q}})\|^2_{Q_q}. \tag{15}$$

Here $\boldsymbol{e}_q$ denotes the quaternion attitude error. The MPPI weights are then computed as

$$w^i_n = \frac{\exp\left(-\tilde{J}^i_n/\lambda\right)}{\sum_{s=1}^{N} \exp\left(-\tilde{J}^i_s/\lambda\right)}, \qquad \lambda > 0, \tag{16}$$

where $\tilde{J}^i_n$ denotes the batch-normalized rollout cost used for numerical stability. The optimal action-sequence mean is updated by

$$\boldsymbol{\mu}^{\text{opt}}_i = \sum_{n=1}^{N} w^i_n \mathbf{u}^{n,i}_{1:H}. \tag{17}$$

Rolling out $\boldsymbol{\mu}^{\text{opt}}_i$ through the dynamics yields a dynamically feasible trajectory $\boldsymbol{\tau}^{0,i}_{\text{MB}}$. The MPPI action mean is then updated with $\boldsymbol{\mu}^{\text{opt}}_i$ for the next reverse step, and shifted between consecutive replanning cycles in a receding-horizon manner.

Finally, we convert the dynamically feasible trajectory into a diffusion-space score. All variables in this conversion are represented in the normalized diffusion trajectory space:

$$\boldsymbol{s}^i_{\text{MB}} = \frac{-\boldsymbol{\tau}^i + \sqrt{\bar{\alpha}_i}\boldsymbol{\tau}^{0,i}_{\text{MB}}}{1 - \bar{\alpha}_i}. \tag{18}$$

The equivalent noise correction is

$$\boldsymbol{\varepsilon}^i_{\text{MB}} = -\sqrt{1-\bar{\alpha}_i}\boldsymbol{s}^i_{\text{MB}}. \tag{19}$$

The final denoising direction is

$$\boldsymbol{\varepsilon}^i_{\text{total}} = \boldsymbol{\varepsilon}^i_{\text{comp}} + \omega_{\text{MB}} \boldsymbol{\varepsilon}^i_{\text{MB}}, \tag{20}$$

where $\omega_{\text{MB}}$ controls the strength of the model-based feasibility guidance. This additive guidance biases denoising toward trajectories that are both consistent with the composed diffusion prior and executable under the quadrotor dynamics.

TABLE I
ONLINE RACING PERFORMANCE ACROSS UNSEEN TRACK DIFFICULTIES.

| Method | Success rate ↑ | | | | Gate pass ↑ | Pass error ↓ | Safety margin ↑ | Lap time ↓ |
|---|---|---|---|---|---|---|---|---|
| | Easy (%) | Medium (%) | Hard (%) | Overall (%) | Overall (%) | Overall (m) | Overall (m) | Overall (s) |
| **CDP (ours)** | 91.3 | **77.3** | **74.7** | **81.1** | **90.8** | 0.158 | 0.359 | 9.36 |
| MoE-DP | 78.0 | 72.7 | **74.7** | 75.1 | 90.2 | **0.145** | **0.370** | 10.67 |
| Waypoint-only DP | 78.0 | 7.3 | 62.7 | 49.3 | 49.0 | 0.266 | 0.250 | 7.70 |
| VG-MPPI | **96.7** | 76.0 | 30.7 | 67.8 | 66.9 | 0.214 | 0.304 | **5.87**$^{\dagger}$ |

## IV. RESULTS

We organize the results around three challenges raised by cross-modal score composition: zero-shot transfer, cross-modal complementarity, and dynamics-aware feasibility.

1) First, we ask whether independently trained waypoint and vision experts can be composed online to race through track layouts that were never seen during training. We address this through unseen-track racing tests that transfer across new static route geometries, gate offsets, and gate orientations.
2) Second, we ask whether waypoint intent and visual geometry provide distinct but complementary constraints within the same control decision. The double gate task isolates this question by creating visually ambiguous multi-gate scenes in which waypoint conditioning resolves the intended route and visual conditioning supports precise gate traversal.
3) Third, we ask whether score composition perturbs the learned trajectory distribution or compromises dynamic feasibility. The model-based denoising study addresses this issue by examining whether composed samples can be regularized towards dynamically feasible, model-supported quadrotor trajectories.

### A. Compositional Drone Racing on Unseen Tracks

We first evaluated whether the proposed compositional diffusion planner can generalize gate-valid racing behavior to unseen tracks. The composed diffusion planner, denoted as CDP, combines waypoint conditioning for route-level progress with image conditioning for local gate alignment. The waypoint-conditioned score provides a global racing scaffold, while the vision-conditioned score corrects trajectories near gates whose poses are not known in advance.

We compared CDP with two learned ablations and one optimization-based baseline. The waypoint-only ablation removes visual conditioning and tests whether route guidance alone is sufficient for gate traversal. The no-compose ablation, denoted as Mixture of Experts (MoE-DP), replaces simultaneous score composition with switching between single-modality behaviors. We also compare against a vision-guided MPPI controller Zhao et al. [40] (VG-MPPI), which uses depth images to reshape MPPI samples and represents a different online control paradigm.

The evaluation used 15 unseen tracks generated from waypoint sequences, including five easy, five medium, and five hard layouts. Gates were placed with varying positions and orientations, and could either coincide with or deviate from the nominal waypoint path. During execution, waypoint coordinates were provided as route guidance, but gate poses were not available. Each method was evaluated under the same track and speed protocol.

We recomputed all metrics from the recorded trajectories. Success rate and gate-pass rate measure task validity. Pass error and safety margin are computed in the local gate frame and are count-weighted over available gate-crossing statistics. Lap time measures traversal efficiency. For VG-MPPI, lap time is averaged only over completed rollouts, since many failed rollouts do not produce valid lap-time measurements. Therefore, lap time should be interpreted jointly with success rate and gate-pass rate.

Table I shows that CDP achieves the strongest task validity on unseen tracks. CDP obtains the highest success rate and gate-pass rate, with 81.1% successful rollouts and 90.8% valid gate passages. VG-MPPI records the shortest completed-lap time, but its lower success and gate-pass rates indicate that local visual optimization can produce fast motion while missing required gates. This trade-off shows that speed alone is not sufficient for valid racing on unseen waypoint–gate layouts.

The ablations further clarify the role of score composition. Waypoint-only DP is fast, but its low gate-pass rate shows that route guidance alone cannot correct for gate displacement, orientation, or aperture-level alignment. MoE-DP achieves slightly lower pass error and higher safety margin than CDP on successful crossings, but it is less reliable over complete rollouts. This suggests that switching between single-modality behaviors can recover accurate local crossings in some cases, but simultaneous score composition better preserves full-track task validity.

Together, these results show that compositional diffusion planning improves online racing reliability by coupling global route progress with local visual gate alignment. CDP is not always the fastest method, but it achieves the strongest task validity across unseen tracks. The proposed composition therefore provides the most reliable trade-off among success rate, gate-pass rate, gate-centered traversal, and traversal efficiency.

### B. Waypoint Guidance for Multimodal Gate Traversal

The previous experiments showed that visual guidance can correct waypoint-guided flight when gate geometry varies. We next evaluate the complementary question: whether waypoint guidance can resolve branch ambiguity when the visual scene contains multiple feasible gate apertures.

We designed a symmetric double-gate task with two horizontally separated apertures in front of the quadrotor. Since both apertures were physically traversable, the vision-conditioned observation alone did not specify a unique route. The task therefore tested whether a coarse waypoint command could select the intended branch while visual guidance preserved local gate-centered alignment. The two apertures were separated by 2.0 m, and the waypoint was placed on the desired side rather than at the gate center. This required the waypoint-conditioned expert to choose the correct branch and the vision-conditioned expert to refine the final crossing.

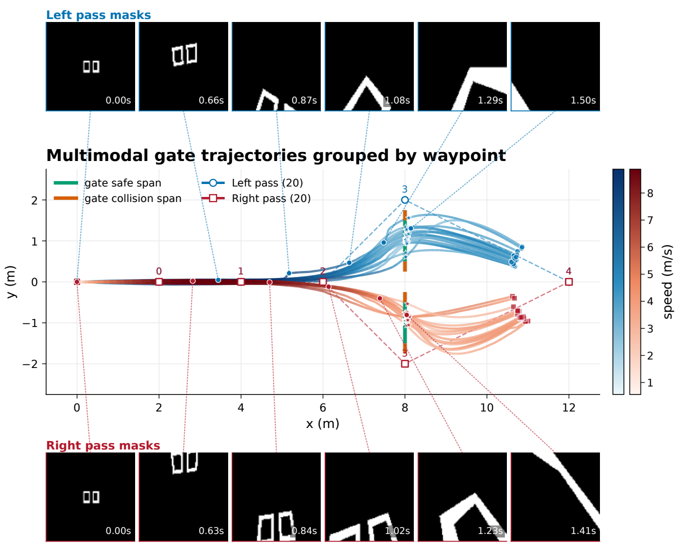


Fig. 2. **Waypoint conditioning resolves visual ambiguity in a symmetric double-gate scene.** Both apertures are visually feasible, but the waypoint selects the intended branch. Overlaid trajectories separate according to waypoint intent, while visual feedback refines the final crossing.

Figure 2 shows that waypoint conditioning consistently determines the selected branch. Left-commanded trajectories cross the gate plane on the positive local-$y$ side, whereas right-commanded trajectories cross on the negative side. This separation indicates that waypoint guidance provides stable route intent even when both gates are visually valid.

At the same time, waypoint guidance alone is not sufficient for safe traversal. Because the branch waypoint is intentionally offset from the aperture center, pure waypoint tracking would bias the vehicle toward the route cue rather than the safe gate span. In the composed planner, waypoint conditioning selects the branch, and visual conditioning corrects the local crossing geometry. This combination achieves collision-free traversal in 39 of 40 rollouts, with similar gate-centered accuracy on both branches.

These results highlight the complementary roles of the two modalities. Waypoints provide controllable route intent when perception admits multiple feasible choices, while visual feedback enforces local alignment at the selected gate. Together with the unseen-track experiments, this result shows that the two modalities address different sources of ambiguity: visual guidance corrects geometric variation, whereas waypoint guidance resolves branch ambiguity.

### C. Dynamic Feasibility of Cross-modal Composition

We next evaluated whether cross-modal score composition remains compatible with quadrotor dynamics under aggressive and out-of-distribution racing conditions. Composing independently trained waypoint and vision experts can move samples outside the support of either model. The resulting trajectory may satisfy task-level guidance while becoming locally inconsistent with quadrotor dynamics. We therefore add an MPPI-based correction during denoising to bias composed samples toward executable motion.

We varied the waypoint and vision guidance weights and swept the model-based correction weight $\omega_{\mathrm{MB}}$ in (20). As shown in Fig. 3, the benefit of model-based guidance depends strongly on the composition regime. We measure dynamic feasibility using cumulative admissibility error (CAE) Bouvier et al. [4], which is defined as the distance between a predicted trajectory and its closest admissible projection. Balanced multimodal compositions benefit most from moderate correction, whereas waypoint-dominant, vision-only, or strongly amplified compositions show weaker gains. The conditional energies at the final denoising step provide an explanation for this dependency. We compute each expert's energy under its own condition and use it as an indicator of how well the generated trajectory remains within that expert's learned support. When one modality dominates the composition, denoising stays relatively close to the dominant expert's support. In contrast, when the waypoint and vision guidance weights increase together, the conditional energies of both experts rise. This suggests that aggressive cross-modal composition can push samples away from both learned supports. Therefore, model-based correction is most beneficial when the two diffusion experts provide balanced task guidance rather than strongly conflicting scores.

The qualitative comparison in Fig. 4 supports the same conclusion. Without model-based guidance, the rollout contains sharper local trajectory changes, and some body-frame action candidates show abrupt or inconsistent directions. With moderate guidance, the trajectory becomes smoother and the action candidates remain more forward-consistent. The rollout preserves the intended cross-modal behavior while becoming more compatible with quadrotor dynamics.

These results show that model-based denoising acts as a bounded feasibility regularizer rather than an external controller. It is most effective when the composed diffusion score already provides a plausible task-level trajectory and the correction weight remains moderate. If the correction is too strong, or if the composition is dominated by one expert, the model-based term can compete with the learned prior and reduce the benefit of the composition. Overall, dynamic feasibility improves most consistently when cross-modal composition and physics-informed regularization remain balanced.

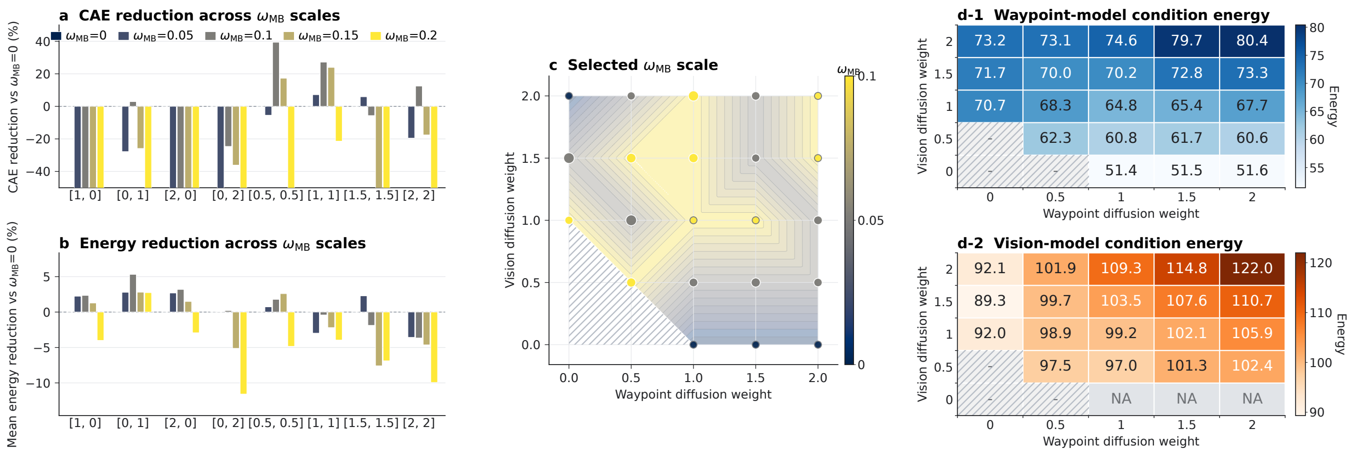


Fig. 3. **The benefit of model-based guidance depends on cross-modal composition weights. (a-b)** The figure compares changes in final CAE introduced in Bouvier et al. [4] and energy after applying MPPI guidance across diffusion weight combinations. Positive bars denote improved dynamic feasibility relative to the unguided setting, whereas negative bars denote degradation. **(c)** Selected model-based guidance weight $\omega_{\mathrm{MB}}$ that yields the largest improvement for each waypoint–vision diffusion weight combination. **(d)** Heat-map visualization of the condition energies induced by compositional diffusion: **(d-1)** the waypoint-model condition energy and **(d-2)** the vision-model condition energy.

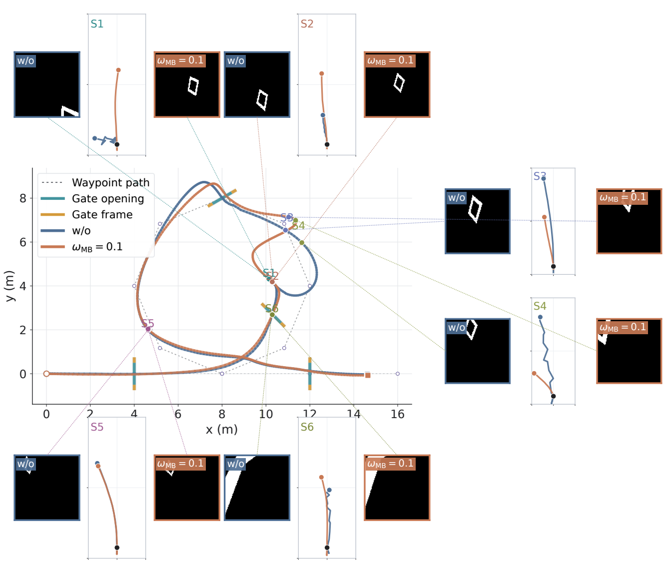


Fig. 4. **Physically informed denoising improves trajectory feasibility while preserving intended behavior.** For the diffusion setting $[1, 1]$, rollouts without MPPI guidance and with moderate MPPI guidance are compared. Coloured markers indicate matched states selected from the flight trajectory. For each state, the saved segmentation mask and corresponding body-frame action candidate are shown.

## V. Conclusion

This paper presented a real-time multimodal energy-based diffusion planner for zero-shot autonomous drone racing. The proposed framework decomposes racing into two complementary trajectory-generation skills. A waypoint-conditioned expert provides route-level progress, while a vision-conditioned expert provides local gate-traversal guidance from onboard observations. By composing their scores during denoising, the planner converts coarse route intent and visual gate evidence into a unified racing trajectory. An MPPI-based model-guided term further regularizes the composed samples toward dynamically feasible quadrotor motion.

Experiments on unseen tracks, multi-gate scenes, and model-based denoising ablations validate the role of each component. On unseen tracks, compositional diffusion planning improves task-valid racing by coupling global waypoint progress with local gate alignment. In symmetric double-gate scenes, waypoint conditioning resolves visual branch ambiguity, while visual conditioning maintains safe aperture traversal. The feasibility analysis further shows that moderate model-based guidance improves trajectory consistency without replacing the learned cross-modal prior.

These results indicate that independently trained modal skills can be composed into a transferable racing planner for flexible waypoint–gate environments. The method does not require predefined reference trajectories, known global gate poses, or track-specific retraining. At the same time, the experiments show that composition and model-based correction must remain balanced, especially under aggressive motion and moving gates. This suggests that compositional diffusion with physics-aware denoising is a practical route toward generalizable agile flight with onboard perception.

## VI. Supplementary Material

### A. Training and Inference

All training and deployment experiments were conducted on an NVIDIA RTX 3090 GPU. Each diffusion expert was trained independently, and training one expert took approximately 12 hours. During inference, we used DDIM sampling for fast trajectory generation. For single-model inference, the number of DDIM denoising steps was set to 8. For compositional inference, the number of DDIM denoising steps was set to 4 to reduce runtime while preserving online planning performance. The complete planning pipeline runs at approximately 16 Hz during deployment.

## REFERENCES


[1] Maulana Bisyir Azhari, Donghun Han, Je In You, Sungjun Park, and David Hyunchul Shim. Drift-corrected monocular vio and perception-aware planning for autonomous drone racing. *arXiv preprint arXiv:2512.20475*, 2025.

[2] Stavrow A Bahnam, Robin Ferede, Till M Blaha, Anton E Lang, Erin Lucassen, Quentin Missinne, Aderik EC Verraest, Christophe De Wagter, and Guido CHE de Croon. Monorace: Winning champion-level drone racing with robust monocular ai. *arXiv preprint arXiv:2601.15222*, 2026.

[3] Michael Bosello, Flavio Pinzarrone, Sara Kiade, Davide Aguiari, Yvo Keuter, Aaesha AlShehhi, Gyordan Caminati, Kei Long Wong, Ka Seng Chou, Junaid Halepota, Fares Alneyadi, Jacopo Panerati, and Giovanni Pau. On your own: Pro-level autonomous drone racing in uninstrumented arenas. *IEEE Robotics and Automation Letters*, 11(3):2674–2681, March 2026. ISSN 2377-3774. doi: 10.1109/lra.2026.3653405. URL http://dx.doi.org/10.1109/LRA.2026.3653405.

[4] Jean-Baptiste Bouvier, Kanghyun Ryu, Kartik Nagpal, Qiayuan Liao, Koushil Sreenath, and Negar Mehr. Ddat: Diffusion policies enforcing dynamically admissible robot trajectories. In *Robotics: Science and Systems (RSS)*, 2025.

[5] Jiahang Cao, Yize Huang, Hanzhong Guo, Rui Zhang, Mu Nan, Weijian Mai, Jiaxu Wang, Hao Cheng, Jingkai Sun, Gang Han, et al. Compose your policies! improving diffusion-based or flow-based robot policies via test-time distribution-level composition. *arXiv preprint arXiv:2510.01068*, 2025.

[6] Chen-Hao Chao, Wei-Fang Sun, Bo-Wun Cheng, and Chun-Yi Lee. On investigating the conservative property of score-based generative models. *arXiv preprint arXiv:2209.12753*, 2022.

[7] Haonan Chen, Jiaming Xu, Hongyu Chen, Kaiwen Hong, Binghao Huang, Chaoqi Liu, Jiayuan Mao, Yunzhu Li, Yilun Du, and Katherine Driggs-Campbell. Multi-modal manipulation via multi-modal policy consensus. *arXiv preprint arXiv:2509.23468*, 2025.

[8] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

[9] Yilun Du, Conor Durkan, Robin Strudel, Joshua B Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Sussman Grathwohl. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. In *International conference on machine learning*, pages 8489–8510. PMLR, 2023.

[10] Davide Falanga, Elias Mueggler, Matthias Faessler, and Davide Scaramuzza. Aggressive quadrotor flight through narrow gaps with onboard sensing and computing using active vision. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5774–5781, 2017. doi: 10.1109/ICRA.2017.7989679.

[11] Philipp Foehn, Angel Romero, and Davide Scaramuzza. Time-optimal planning for quadrotor waypoint flight. *Science Robotics*, 6(56):eabh1221, 2021.

[12] Ismail Geles, Leonard Bauersfeld, Angel Romero, Jiaxu Xing, and Davide Scaramuzza. Demonstrating agile flight from pixels without state estimation. *arXiv preprint arXiv:2406.12505*, 2024.

[13] Zhichao Han, Zhepei Wang, Neng Pan, Yi Lin, Chao Xu, and Fei Gao. Fast-racing: An open-source strong baseline for $\mathrm{SE}(3)$ planning in autonomous drone racing. *IEEE Robotics and Automation Letters*, 6(4):8631–8638, 2021. doi: 10.1109/LRA.2021.3113976.

[14] Drew Hanover, Antonio Loquercio, Leonard Bauersfeld, Angel Romero, Robert Penicka, Yunlong Song, Giovanni Cioffi, Elia Kaufmann, and Davide Scaramuzza. Autonomous drone racing: A survey. *IEEE Transactions on Robotics*, 40:3044–3067, 2024. doi: 10.1109/TRO.2024.3400838.

[15] Wonsuhk Jung, Utkarsh Aashu Mishra, Nadun Ranawaka Arachchige, Yongxin Chen, Danfei Xu, and Shreyas Kousik. Joint model-based model-free diffusion for planning with constraints. In *9th Annual Conference on Robot Learning*, 2025. URL https://openreview.net/forum?id=E9t1ekt6W9.

[16] Elia Kaufmann, Mathias Gehrig, Philipp Foehn, René Ranftl, Alexey Dosovitskiy, Vladlen Koltun, and Davide Scaramuzza. Beauty and the beast: Optimal methods meet learning for drone racing. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 690–696. IEEE, 2019.

[17] Elia Kaufmann, Leonard Bauersfeld, Antonio Loquercio, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. Champion-level drone racing using deep reinforcement learning. *Nature*, 620(7976):982–987, 2023.

[18] Maria Krinner, Angel Romero, Leonard Bauersfeld, Melanie Zeilinger, Andrea Carron, and Davide Scaramuzza. Mpcc++: Model predictive contouring control for time-optimal flight with safety constraints. *arXiv preprint arXiv:2403.17551*, 2024.

[19] Shuo Li, Michaël MOI Ozo, Christophe De Wagter, and Guido CHE de Croon. Autonomous drone race: A computationally efficient vision-based navigation and control strategy. *Robotics and Autonomous Systems*, 133:103621, 2020.

[20] Shuo Li, Erik van der Horst, Philipp Duernay, Christophe De Wagter, and Guido CHE de Croon. Visual model-predictive localization for computationally efficient autonomous racing of a 72-g drone. *Journal of Field Robotics*, 37(4):667–692, 2020.

[21] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European confer-*

*ence on computer vision*, pages 423–439. Springer, 2022.

[22] Giuseppe Loianno, Chris Brunner, Gary McGrath, and Vijay Kumar. Estimation, control, and planning for aggressive flight with a small quadrotor with a single camera and imu. *IEEE Robotics and Automation Letters*, 2(2):404–411, 2017. doi: 10.1109/LRA.2016.2633290.

[23] Yunhao Luo, Chen Sun, Joshua B Tenenbaum, and Yilun Du. Potential based diffusion motion planning. *arXiv preprint arXiv:2407.06169*, 2024.

[24] Wule Mao, Zhouheng Li, Yunhao Luo, Yilun Du, and Lei Xie. Rapid and safe trajectory planning over diverse scenes through diffusion composition, 2025. URL https://arxiv.org/abs/2507.04384.

[25] Daniel Mellinger and Vijay Kumar. Minimum snap trajectory generation and control for quadrotors. In *2011 IEEE International Conference on Robotics and Automation*, pages 2520–2525. IEEE, 2011.

[26] Chaoyi Pan, Zeji Yi, Guanya Shi, and Guannan Qu. Model-based diffusion for trajectory optimization, 2024. URL https://arxiv.org/abs/2407.01573.

[27] Christian Pfeiffer and Davide Scaramuzza. Human-piloted drone racing: Visual processing and control. *IEEE Robotics and Automation Letters*, 6(2):3467–3474, 2021. doi: 10.1109/LRA.2021.3064282.

[28] Chao Qin, Maxime S.J. Michet, Jingxiang Chen, and Hugh H.-T. Liu. Time-optimal gate-traversing planner for autonomous drone racing. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8693–8699, 2024. doi: 10.1109/ICRA57147.2024.10610148.

[29] Ralf Römer, Alexander von Rohr, and Angela Schoellig. Diffusion predictive control with constraints. In *Proceedings of the 7th Annual Learning for Dynamics & Control Conference*, pages 791–803. PMLR, 2025.

[30] Angel Romero, Sihao Sun, Philipp Foehn, and Davide Scaramuzza. Model predictive contouring control for time-optimal quadrotor flight. *IEEE Transactions on Robotics*, 38(6):3340–3356, 2022.

[31] Yuyang Shen, Jin Zhou, Danzhe Xu, Fangguo Zhao, Jinming Xu, Jiming Chen, and Shuo Li. Aggressive trajectory generation for a swarm of autonomous racing drones. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7436–7441, 2023. doi: 10.1109/IROS55552.2023.10341844.

[32] Yunlong Song, Angel Romero, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. Reaching the limit in autonomous racing: Optimal control versus reinforcement learning. *Science Robotics*, 8(82):eadg1462, 2023.

[33] Julen Urain, Ajay Mandlekar, Yilun Du, Nur Muhammad, Danfei Xu, Katerina Fragkiadaki, Georgia Chalvatzaki, Jan Peters, et al. A survey on deep generative models for robot learning from multimodal demonstrations. *IEEE Transactions on Robotics*, 42:60–79, 2025.

[34] Hongze Wang, Jiaxu Xing, Nico Messikommer, and Davide Scaramuzza. Environment as policy: Learning to race in unseen tracks. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11333–11339, 2025. doi: 10.1109/ICRA55743.2025.11127376.

[35] Qianhao Wang, Dong Wang, Chao Xu, Alan Gao, and Fei Gao. Polynomial-based online planning for autonomous drone racing in dynamic environments. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1078–1085, 2023. doi: 10.1109/IROS55552.2023.10342456.

[36] Jiaxu Xing, Angel Romero, Leonard Bauersfeld, and Davide Scaramuzza. Bootstrapping reinforcement learning with imitation for vision-based agile flight. *arXiv preprint arXiv:2403.12203*, 2024.

[37] Feng Yu, Yang Su, Yu Hu, Yang Deng, Linzuo Zhang, and Danping Zou. Mastering diverse, unknown, and cluttered tracks for robust vision-based drone racing. *IEEE Robotics and Automation Letters*, 11(2):2090–2097, 2026. doi: 10.1109/LRA.2025.3643267.

[38] Fangguo Zhao, Xin Guan, and Shuo Li. Rethinking reference trajectories in agile drone racing: A unified reference-free model-based controller via mppi. *arXiv preprint arXiv:2509.14726*, 2025.

[39] Fangguo Zhao, Jiahao Mei, Jin Zhou, Yuanyi Chen, Jiming Chen, and Shuo Li. Gate-aware online planning for two-player autonomous drone racing. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8536–8542, 2025. doi: 10.1109/ICRA55743.2025.11127454.

[40] Fangguo Zhao, Hanbing Zhang, Zhouheng Li, Xin Guan, and Shuo Li. Vision-guided mppi for agile drone racing: Navigating arbitrary gate poses via neural signed distance fields, 2026. URL https://arxiv.org/abs/2603.07199.

[41] Jin Zhou, Jiahao Mei, Fangguo Zhao, Jiming Chen, and Shuo Li. Online motion planning for quadrotor multi-point navigation using efficient imitation learning-based strategy. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5386–5393, 2025. doi: 10.1109/IROS60139.2025.11246459.