

Backbone Architecture Design for Diffusion-Based UAV Trajectory Planning in Dynamic Wireless Environments

Minjoo Kim and Soohyun Park

Department of Computer Science, Sookmyung Women's University, Seoul, Republic of Korea

E-mails: {minjoo, soohyun.park}@sookmyung.ac.kr

Abstract—Diffusion models have shown strong potential for trajectory generation in robot navigation, yet the effect of denoising backbone architecture remains underexplored for UAV planning in dynamic wireless environments. This paper evaluates six backbone architectures within a unified diffusion-based subpath planner for UAV dynamic obstacle avoidance. The comparison covers validation loss, trajectory smoothness, inference latency, parameter count, training stability, and closed-loop navigation success. The results show that backbone selection creates meaningful trade-offs across these metrics. UNet-1D provides the smoothest and most parameter-efficient trajectories, U-ViT achieves the lowest validation loss, and attention-based models remain competitive in navigation success. These findings suggest that diffusion backbones should be selected according to planning horizon, smoothness requirements, and deployment constraints rather than architectural novelty alone.

Index Terms—Diffusion model, UAV trajectory planning, dynamic obstacle avoidance, denoising backbone, reinforcement learning

I. INTRODUCTION

Unmanned aerial vehicles (UAVs) are increasingly used as aerial communication platforms, including temporary base stations, mobile relays, sensing nodes, and emergency network infrastructure. In dynamic wireless environments, the UAV trajectory affects not only collision safety but also communication quality, service continuity, and energy consumption. A poorly planned route may increase link outage, delay task completion, or consume excessive battery energy even if the UAV eventually reaches the goal. Therefore, UAV trajectory planning must generate feasible, smooth, and computationally efficient paths while remaining robust to moving obstacles and spatially varying wireless conditions. This requirement is especially important in urban or disaster-response scenarios, where ground infrastructure may be partially unavailable and the UAV must maintain connectivity while moving through cluttered three-dimensional space. In such settings, the motion planner cannot be evaluated only by whether it reaches the destination. The generated trajectory must also avoid aggressive waypoint changes, remain compatible with low-level flight control, and support repeated replanning under limited onboard computation.

Reinforcement learning (RL) has been widely adopted for UAV navigation, offering reactive control policies that

adapt to environmental changes [1], [2]. However, RL-based planners often struggle under high obstacle density, as the policy must simultaneously reason about global navigation goals, local collision avoidance, energy-efficient motion, and wireless service constraints. When the policy directly outputs continuous control commands, small errors in local decisions can accumulate into oscillatory motion or unsafe detours. One practical way to reduce this burden is to provide the RL controller with a short-horizon subpath that summarizes a feasible local direction of travel. The controller can then focus on tracking and local correction, while the subpath planner captures the distribution of smooth waypoint sequences.

Diffusion models, originally developed for image synthesis [3], have recently emerged as a powerful alternative for trajectory generation in robotics. By learning a distribution over trajectories and generating waypoints through iterative denoising, diffusion-based planners provide smooth and multi-modal subpaths [4], [5]. In UAV navigation, such subpaths can complement a high-level RL policy by providing structured waypoint guidance in cluttered and wireless-aware environments. This is particularly useful when the planner must balance short-horizon obstacle avoidance with smooth motion and deployment constraints. Unlike deterministic regression models, diffusion planners can represent multiple plausible local maneuvers around obstacles and recover a trajectory by gradually removing noise from an initially random sample. This makes the denoising network a central component of the planner rather than a minor implementation detail.

A critical yet underexplored design choice in diffusion-based trajectory planners is the backbone architecture of the denoising network. Existing works commonly adopt UNet-1D [4], while recent sequence models such as Transformers [6], U-ViT [7], TCN [8], and Mamba [9] offer alternative inductive biases. Whether these architectures improve UAV trajectory generation remains unclear, particularly for short-horizon planning with smoothness and real-time constraints. Architectures that are powerful for long text or vision sequences do not necessarily provide the best inductive bias for a short sequence of UAV waypoints. For example, global attention can model pairwise dependencies across all waypoints, while convolutional filters emphasize local temporal continuity. State-space models are designed for efficient long-range sequence processing, but the benefit of this property may

Soohyun Park is the corresponding author (e-mail: soohyun.park@sookmyung.ac.kr).

be limited when the planning horizon contains only a small number of waypoints.

In this paper, we investigate how the denoising backbone affects diffusion-based UAV trajectory planning. Six architectures are evaluated within the same hierarchical planning framework using training, trajectory-quality, efficiency, and navigation metrics. The principal contributions are as follows:

- Six denoising backbones are compared under a controlled UAV trajectory planning setting with identical data, context encoding, and sampling configuration.
- The results quantify trade-offs among validation loss, trajectory smoothness, inference cost, parameter count, and navigation success in dynamic obstacle environments.
- The analysis provides practical guidance for selecting diffusion backbones for short-horizon UAV planning, where model capacity, local smoothness, and deployment efficiency must be considered together.

II. RELATED WORK

A. Diffusion Models for Trajectory Planning

Diffusion probabilistic models were introduced by Ho et al. [3] as an iterative denoising framework for generative modeling. Janner et al. [4] extended diffusion models to offline reinforcement learning by formulating trajectory planning as a denoising problem over state-action sequences. Decision Diffuser [5] further incorporated conditional generation for decision making. These works established diffusion models as effective trajectory generators, but the effect of denoising backbone design has received less attention in UAV navigation, where short horizons, obstacle density, and deployment constraints jointly affect performance. Most trajectory-diffusion studies emphasize conditioning strategies, sampling procedures, or reward-guided generation. By contrast, the internal architecture used to predict denoising corrections is often inherited from prior work without systematic comparison. For UAV navigation, this omission is nontrivial because the denoiser must generate waypoint sequences that are both geometrically feasible and smooth enough for execution. The present study focuses on this architectural dimension while keeping the dataset, context representation, and sampling procedure fixed.

B. Backbone Architectures for Sequence Modeling

Neural sequence backbones embody different inductive biases. UNet-1D adapts the encoder-decoder U-Net structure [10] to temporal sequences. TCN uses dilated convolutions for temporal modeling [8]. Transformers model pairwise dependencies through self-attention [6], while U-ViT adapts ViT-style blocks to diffusion architectures [7]. Mamba introduces selective state-space modeling for efficient long-sequence processing [9]. The suitability of these architectures for short-horizon UAV trajectory denoising remains an empirical question. Convolutional models impose local temporal structure and naturally favor smooth changes between adjacent waypoints. Attention-based models provide flexible global interactions but introduce larger parameter counts and weaker

locality unless positional and architectural biases are carefully tuned. State-space models offer efficient sequence processing with recurrent dynamics, but their optimization behavior can differ from that of convolutional denoisers in small-data or short-horizon regimes. These distinctions motivate a controlled comparison under the same UAV planning task.

C. UAV Navigation with Reinforcement Learning

Model-free RL methods such as SAC [1] and TD3 [2] have been applied to continuous-control UAV navigation. Hierarchical approaches decompose navigation into high-level subgoal selection and low-level control [11]. In wireless settings, UAV mobility is also tied to link quality, service continuity, and energy consumption. This work follows a hierarchical structure in which a diffusion planner generates short-horizon subpaths and an SAC controller executes waypoint tracking under dynamic obstacle and channel-context inputs. This decomposition is useful because wireless UAV navigation contains both strategic and local components. The strategic component involves moving toward the goal while maintaining usable communication context, whereas the local component requires smooth motion around nearby obstacles. The diffusion planner is therefore evaluated as a local trajectory generator embedded within a closed-loop navigation system rather than as an isolated offline prediction model.

III. EXPERIMENTAL SETUP

A. UAV Navigation Scenario

Experiments are conducted in a simulated 3D workspace of dimensions $40 \times 40 \times 12$ m. The workspace contains randomly placed static spherical obstacles and dynamic obstacles with randomized initial positions and constant velocities. The UAV must reach a goal position from a random start while avoiding collisions. Two difficulty tiers are considered:

- **Standard:** 3 or 6 simultaneous dynamic obstacles, used for framework-level validation.
- **Dense:** 10 or 15 simultaneous dynamic obstacles, used as the main stress test for backbone comparison.

The dense setting is designed to stress short-horizon replanning. As the number of moving obstacles increases, a locally feasible direction can become invalid within a few time steps, requiring the planner to generate subpaths that are smooth but still responsive to changes in nearby obstacle positions. This setting also reflects wireless UAV operation in crowded environments, where detours around obstacles may alter the local channel context and increase mission time.

B. Planning Framework

A hierarchical planning framework is adopted. A high-level SAC policy manages task-level navigation and periodically queries a diffusion-based subpath planner. The planner receives a context vector composed of the current UAV state (9-dimensional), obstacle encoding (40-dimensional), a local channel-quality patch (500-dimensional), and an urgency signal (3-dimensional). Given this context, the planner generates $H = 10$ waypoints through DDIM sampling [12] with $T =$

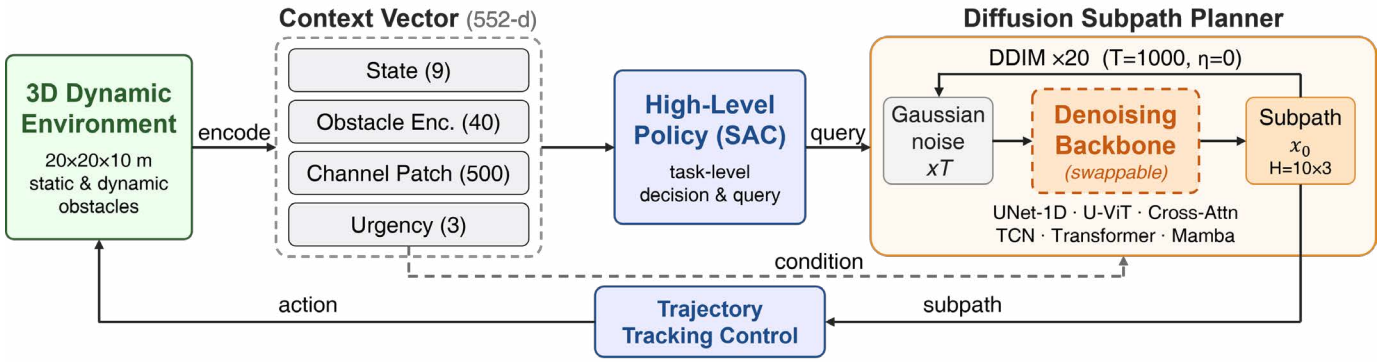


Fig. 1. Overview of the hierarchical UAV navigation framework. A high-level SAC policy observes the environment through a shared context vector (state, obstacle encoding, channel patch, and urgency signal) and queries the diffusion subpath planner, which denoises a Gaussian sample into an $H=10$ waypoint subpath via DDIM sampling. A trajectory tracking controller executes the subpath in the environment. Only the denoising backbone (dashed) is varied across experiments; all other components are held fixed.

1000 diffusion steps and 20 inference steps. The generated waypoints act as a compact local plan rather than a complete episode-level trajectory. This design allows the system to replan as obstacles move and as the UAV enters different regions of the workspace. The same context vector is used for every backbone so that differences in performance can be attributed to the denoising architecture rather than to changes in environmental information.

Only the denoising backbone is varied across experiments. The training dataset, optimizer, learning-rate schedule, context encoding, and DDIM configuration are held fixed. Each backbone is trained for 200 epochs on 50,000 trajectory samples collected from the same environment distribution. This controlled protocol is important because diffusion training can be sensitive to implementation details such as sampling steps, conditioning dimensionality, and optimizer settings. Holding these components fixed isolates the architectural effect and makes the resulting comparison more interpretable.

C. Evaluation Metrics

Each backbone is evaluated using validation loss, trajectory smoothness, inference latency, parameter count, navigation success rate, and training stability. Trajectory smoothness is measured by mean absolute acceleration across consecutive waypoints. Navigation success is the fraction of episodes in which the UAV reaches the goal without collision. A backbone is marked unstable when the final validation loss exceeds the best validation loss by more than 50%. Validation loss measures the offline denoising objective, whereas navigation success measures closed-loop behavior after the generated subpaths interact with the SAC controller and dynamic environment. These two quantities are not expected to be identical, since a model with low denoising error may still generate locally irregular waypoints, and a model with slightly higher loss may remain effective after feedback control. For this reason, the evaluation emphasizes trade-offs across metrics rather than ranking backbones by a single scalar value.

IV. BACKBONE ARCHITECTURES

Six backbone architectures are evaluated as the denoising network within the diffusion trajectory planner. Each architecture receives the same context vector and produces a flattened waypoint sequence of dimension $H \times 3$.

A. UNet-1D

UNet-1D adapts the encoder-decoder architecture of U-Net [10] to 1D temporal sequences. The encoder downsamples the noisy trajectory through 1D convolutional residual blocks, while the decoder upsamples with skip connections from corresponding encoder stages. This structure combines coarse trajectory representation with local waypoint refinement. In the diffusion planner, this structure allows the model to estimate both broad subpath geometry and fine local corrections during denoising. Because each convolutional filter operates over neighboring waypoints, UNet-1D directly reflects the local continuity prior of physical motion.

B. U-ViT

U-ViT [7] replaces the convolutional encoder-decoder with a Vision Transformer backbone while retaining skip connections between shallow and deep transformer layers. The input trajectory is processed through multi-head self-attention to model global temporal relations. This architecture can capture interactions between non-adjacent waypoints and may reduce denoising error when global sequence consistency is important. However, it generally requires more parameters than a compact convolutional backbone.

C. Cross-Attention Transformer

The Cross-Attention Transformer processes the noisy trajectory and context vector through separate encoders and combines them via cross-attention layers. Trajectory tokens attend to context tokens during denoising. This design explicitly exposes the trajectory representation to environmental context at intermediate layers. It is therefore a useful baseline for testing whether stronger context fusion improves short-horizon planning.

TABLE I
BACKBONE CHARACTERISTICS AND TRAINING METRICS

Backbone	Val. Loss ↓	Smooth. ↓	Params	Infer. (ms) ↓	Stable
UNet-1D	0.1878	0.834	2.6M	3.00	✓
U-ViT	0.1749	1.218	7.1M	2.81	✓
CrossAttn	0.1860	1.533	7.2M	3.02	✓
TCN	0.1984	0.897	5.4M	2.74	✓
Transformer	0.2216	1.998	6.7M	2.57	✓
Mamba	0.2183	2.220	7.7M	3.26	×

D. Temporal Convolutional Network

TCN [8] consists of dilated causal convolutions with residual connections. The dilation factor increases across layers, allowing the network to capture temporal dependencies without recurrence or attention. Compared with UNet-1D, TCN uses a feed-forward convolutional sequence model without the same encoder-decoder reconstruction structure. It provides a convolutional alternative for separating the effect of locality from the effect of U-shaped denoising.

E. Transformer

The Transformer backbone [6] applies multi-head self-attention over the waypoint sequence. Each waypoint can attend to every other waypoint, enabling global dependency modeling. For short waypoint sequences, this global access is computationally inexpensive, but it may provide less explicit bias toward local smoothness than temporal convolution.

F. Mamba

Mamba [9] introduces a selective state-space model with input-dependent transition parameters. It processes sequences through a linear scan and is designed for efficient long-sequence modeling. Its inclusion tests whether recent state-space sequence models transfer effectively to diffusion-based waypoint generation. Since the evaluated horizon is short, the experiment also examines whether long-sequence efficiency translates into practical planning gains in this regime.

V. RESULTS

A. Training Efficiency and Model Characteristics

Table I summarizes the training-phase metrics. UNet-1D achieves the lowest trajectory smoothness score (0.834) and the smallest parameter count (2.6M), while U-ViT obtains the lowest validation loss (0.1749). Mamba shows the highest smoothness penalty and unstable validation-loss behavior. The distinction between validation loss and smoothness is important. U-ViT obtains the best denoising loss, but UNet-1D produces smoother waypoint sequences, indicating that the loss value alone does not fully capture trajectory quality for UAV execution. TCN also achieves a low smoothness score, which suggests that convolutional locality is beneficial for this task. However, UNet-1D combines this locality with a smaller parameter count and a denoising-oriented encoder-decoder structure.

TABLE II
NAVIGATION SUCCESS RATE IN DENSE DYNAMIC ENVIRONMENTS

Backbone	Dyn-10		Dyn-15		Avg
	Slow	Fast	Slow	Fast	
UNet-1D	0.733	0.633	0.667	0.700	0.683
U-ViT	0.667	0.667	0.633	0.700	0.667
Mamba	0.633	0.667	0.600	0.733	0.658
Transformer	0.567	0.633	0.600	0.667	0.617
CrossAttn	0.600	0.600	0.533	0.533	0.567
TCN	0.567	0.567	0.500	0.633	0.567

TABLE III
ABLATION STUDY: EFFECT OF SUBPATH PLANNING AND PREDICTION
(AVG. SUCCESS RATE)

Backbone	w/ Pred	w/o Pred	No Subpath
UNet-1D	0.683	0.692	0.467
U-ViT	0.667	0.650	0.358
Mamba	0.658	0.667	0.358
Transformer	0.617	0.658	0.325
CrossAttn	0.567	0.542	0.375
TCN	0.567	0.592	0.475

B. Navigation Performance in Dense Dynamic Environments

Table II reports navigation success rates across four dense dynamic scenarios. UNet-1D achieves the highest average success rate (0.683), followed by U-ViT (0.667) and Mamba (0.658). The gap between the top-performing backbones is moderate, showing that validation loss alone does not determine closed-loop navigation performance. The dense dynamic results also show that no backbone dominates every scenario. For example, Mamba performs competitively in the Dyn-15 fast setting despite its unstable training behavior, while UNet-1D provides the best average performance and the best result in Dyn-10 slow and Dyn-15 slow. This pattern indicates that closed-loop success depends on the interaction among generated subpaths, controller behavior, and obstacle motion rather than on offline denoising quality alone.

C. Ablation Study

To assess the contribution of diffusion subpath planning, the full system is compared against a variant in which diffusion-generated waypoints are replaced by direct interpolation toward the goal. The effect of dynamic obstacle prediction is also examined. Results are shown in Table III. Removing the subpath planner reduces success across all backbones, while prediction provides a smaller additional effect. The No Subpath variant consistently performs worse, confirming that the diffusion-generated local plan contributes substantially to navigation robustness. The effect of prediction is smaller and varies by backbone, suggesting that the primary benefit comes from structured subpath generation rather than from the prediction module alone. This supports the central focus of the paper, which is to analyze the denoising backbone used inside the subpath planner.

D. Discussion

The results show that backbone choice should be evaluated as a trade-off rather than by a single metric. UNet-1D provides the best smoothness and parameter efficiency, U-ViT achieves the lowest validation loss, and attention-based models remain competitive in navigation success. Thus, offline denoising accuracy, trajectory regularity, and closed-loop control performance are related but not interchangeable. This distinction is practically important for UAV deployment. A planner selected only by validation loss may generate trajectories that are accurate under the training objective but less favorable for smooth flight. Conversely, a lightweight model with a slightly higher loss may be preferable when the generated subpaths are easier for the controller to track.

The smoothness result is consistent with the local structure of UAV motion. For a waypoint sequence $\mathbf{x}_{0:H-1}$, a second-order smoothness penalty can be written as

$$\mathcal{J}_{\text{smooth}}(\mathbf{x}) = \sum_{t=1}^{H-2} \|\mathbf{x}_{t+1} - 2\mathbf{x}_t + \mathbf{x}_{t-1}\|_2^2. \quad (1)$$

Its gradient depends on a banded finite-difference operator, so the corresponding correction is local in time. This property matches the inductive bias of 1D convolutional filters and helps explain why UNet-1D generates smoother subpaths in the short-horizon setting. The same observation also explains why a larger global sequence model is not automatically advantageous. When the dominant correction is local, additional capacity for long-range interaction may improve representational flexibility without improving the motion prior most relevant to the task.

The planning horizon in this work is $H = 10$. At this scale, a compact convolutional encoder-decoder can cover most of the waypoint sequence while preserving local detail through skip connections. Architectures designed for long-range sequence modeling, such as Transformer and Mamba, may require longer horizons or richer temporal dependencies to fully exploit their capacity. The instability observed in Mamba further suggests that increasing sequence-modeling capacity does not necessarily improve diffusion trajectory planning under limited horizon and finite simulation data. This does not mean that attention or state-space backbones are unsuitable for UAV planning in general. Their advantages may become clearer in longer-horizon planning, multi-agent coordination, or tasks where the planner must remember events over extended temporal windows. The present results instead indicate that, for short-horizon subpath denoising, local smoothness and stable optimization are more influential than architectural expressiveness alone.

Overall, the study points to a practical design guideline. For short-horizon UAV subpath generation, a backbone should be selected by considering smoothness, robustness, and computational cost together. UNet-1D is attractive when smooth and lightweight planning is prioritized, whereas attention or state-space models may be preferable in settings that emphasize longer temporal reasoning. This guideline is relevant

to dynamic wireless environments, where frequent replanning requires reliable local plans at low computational cost.

VI. CONCLUSION

This paper evaluated six diffusion backbone architectures for UAV trajectory planning in dynamic wireless environments. The results show that backbone design affects validation loss, smoothness, training stability, model size, and closed-loop navigation in different ways. UNet-1D provides the strongest smoothness and parameter-efficiency trade-off, while other backbones remain competitive under selected metrics. These findings suggest that diffusion backbone selection should be guided by planning horizon, motion smoothness requirements, and deployment constraints. The study also shows that diffusion subpath planning is a meaningful component of the hierarchical navigation system, since removing it substantially reduces success across backbones. Future work will extend the analysis to longer horizons, physical UAV experiments, and reward formulations that more directly couple trajectory planning with wireless quality-of-service objectives.

ACKNOWLEDGMENT

This research was supported by MSIT (Ministry of Science and ICT), Korea, under ITRC (Information Technology Research Center) (IITP-2026-RS-2024-00436887) supervised by IITP (Institute for Information & Communications Technology Planning & Evaluation).

REFERENCES

- [1] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018.
- [2] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018.
- [3] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
- [4] M. Janner, Y. Du, J. B. Tenenbaum, and S. Levine, "Planning with diffusion for flexible behavior synthesis," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2022.
- [5] A. Ajay, Y. Du, A. Gupta, J. Tenenbaum, T. Jaakkola, and P. Agrawal, "Is conditional generative modeling all you need for decision-making?" in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023.
- [6] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017.
- [7] F. Bao, S. Nie, K. Xue, Y. Cao, C. Li, H. Su, and J. Zhu, "All are worth words: A vit backbone for diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023.
- [8] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.
- [9] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.
- [10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2015, pp. 234–241.
- [11] A. Levy, G. Konidaris, R. Platt, and K. Saenko, "Learning multi-level hierarchies with hindsight," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.
- [12] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.