

ARTICLE

Predictive and Explainable UAV Navigation via World Models, Structured Future Reasoning and LLM-Based Explanation

Rihem Farkh¹, Ghislain Oudinet², Alaeddine Moussa³ and Yasser Fouad^{4,*}

¹LabISEN, KLaIM, ISEN Méditerranée, Toulon, France

²LabISEN, VISION-AD, ISEN Méditerranée, Toulon, France

³Laboratoire LIS UMR CNRS 7020, Aix Marseille Université, Marseille, France

⁴Department of Applied Mechanical Engineering, College of Applied Engineering, Muzahimiyah Branch, King Saud University, Riyadh, Saudi Arabia

*Corresponding Author: Yasser Fouad. Email: yfouad@ksu.edu.sa

Received: 14 May 2026; Accepted: 13 August 2026; Published: 21 October 2026

ABSTRACT: Autonomous UAV navigation in safety-critical environments requires more than accurate prediction of future states; it also requires the ability to interpret action-conditioned futures as actionable safety risk. This paper presents a predictive and explainable UAV navigation framework that combines a latent world model with a Future Interpretation Module (FIM), a safety-constrained decision layer, and a hybrid explanation module. The world model predicts action-conditioned latent futures, while the FIM converts these rollouts into interpretable descriptors including time-to-collision, minimum vertical clearance, route-commitment risk, and risk trend. These descriptors are used jointly for decision-making, safety filtering, and explanation grounding. In a delayed-hazard diagnostic benchmark with 150 episodes per agent, the FIM policy achieved 100.0% success and 0.0% collisions, compared with 97.3% success and 2.7% collisions for the prediction-only world-model baseline. In descriptor-level ablation, Full FIM achieved the highest success rate (71.3%) while eliminating collisions and route traps, whereas individual descriptors reduced collisions but often produced route-trap failures. Across six cross-environment validation scenarios and 900 Full-FIM episodes, the method achieved 98.0% success and 0.0% collisions. Because the prediction-only baseline was also near saturation in this broad suite, these cross-environment results are interpreted as robustness evidence rather than as a claim of uniform superiority. Overall, the results show that FIM improves diagnostic delayed-risk reasoning and provides robust collision-free behavior across the evaluated suite, while broad superiority over all prediction-only baselines remains outside the scope of the current evidence. Secondary comparisons with adapted YUME, Matrix-Game, NWM, and ANWM remain diagnostic references under a unified three-action UAV interface, not complete re-evaluations of those methods in their native settings.

KEYWORDS: UAV navigation; world models; future interpretation; risk-aware planning; explainable artificial intelligence; autonomous systems

1 Introduction

Unmanned aerial vehicles (UAVs) are increasingly used in inspection, surveillance, logistics, emergency response, warehouse monitoring, and infrastructure maintenance. In these applications, the navigation problem is not limited to obstacle avoidance in static scenes; it also involves reasoning under delayed hazards, occlusions, narrow passages, route traps, and external disturbances such as wind. A UAV may face situations in which the geometrically shortest path is not the safest path, or in which the immediate scene appears

benign even though a future obstacle will soon block the route. These cases are difficult because the decision cannot be made from the current observation alone.

Many learning-based navigation systems remain strongly reactive: they map the current observation to an action or short-horizon policy update. Even when future prediction is available, the system may still lack an explicit intermediate representation that explains how predicted futures relate to safety. This creates a gap between forecasting and interpreting: a system may predict what is likely to happen next, but still fail to answer operationally important questions such as whether the predicted route is becoming unsafe, whether clearance is shrinking too quickly, or whether the vehicle is already committing to a trap.

This work addresses that gap through a modular framework that separates prediction, interpretation, decision-making, and explanation. At time step t , the UAV observation is denoted by $o_t = (x_t, e_t)$ where x_t is the RGB image and e_t contains proprioceptive and inertial measurements such as velocity, altitude, orientation, and IMU-derived motion cues. The observation is encoded into a latent state, a learned world model generates action-conditioned future rollouts, and the Future Interpretation Module (FIM) transforms those rollouts into interpretable risk descriptors. A safety-constrained decision layer uses these descriptors to choose among candidate actions, and a hybrid explanation module verbalizes the resulting rationale without directly controlling the vehicle.

The contribution is framed cautiously. FIM is not presented as a new time-to-collision or clearance metric. Its novelty lies in the structured intermediate representation that converts action-conditioned future rollouts into interpretable risk descriptors and uses them jointly for decision-making, safety filtering, and explanation grounding.

- A future interpretation module that converts action-conditioned future rollouts into structured risk descriptors for decision-making and explanation.
- A decoupled prediction-interpretation-control architecture for UAV navigation under delayed and temporally ambiguous risk.
- A safety-constrained decision module that combines progress objectives with predicted collision, time-to-collision, clearance, risk trend, and route-commitment descriptors.
- A grounded explanation module that verbalizes the controller trace without giving the language model direct control authority.

The experimental evaluation includes explicit discussion of baseline-adaptation limitations, uncertainty reporting, descriptor-level ablation, robustness validation across six environments, and cross-environment visual analyses of trajectories, risk traces, and aggregate metrics. The paper does not claim general superiority over foundation models such as YUME, Matrix-Game, NWM, or ANWM in their native problem settings. Instead, these methods are used as adapted diagnostic references under a unified UAV action interface to assess whether structured future-risk interpretation provides added value within the proposed benchmark. Accordingly, the broad cross-environment suite is interpreted primarily as evidence of robust, explainable, and safety-constrained navigation rather than as proof of uniform numerical superiority over all prediction-only baselines.

2 Related Work

2.1 Predictive UAV Navigation and World Models

Predictive control and learned world modeling have become central to autonomous navigation because they allow candidate actions to be evaluated before execution. Latent dynamics models and recurrent state-space models have shown that compact predictive representations can support planning in partially observed

environments [1,2]. World-model and Dreamer-style approaches motivate the predictive backbone of this framework, but they generally do not expose a structured intermediate representation of future safety.

Recent world-model and vision-language navigation systems can generate or reason about future observations, but they often use predicted futures directly for policy selection or visual scoring. UAV and robot navigation studies, including agile drone racing and self-supervised navigation, further show the importance of coupling learned perception with task-directed action selection [3,4]. However, these systems rarely expose an explicit descriptor-level representation that explains whether an action was selected because it preserved clearance, avoided a delayed route trap, or simply appeared visually closer to the target.

2.2 Risk-Aware Planning, Model Predictive Control, and Safety Constraints

Risk-aware planning, nonlinear model predictive control, and control barrier functions have long used quantities related to collision probability, clearance, time-to-collision, and safety margins [5–8]. This paper does not claim novelty for these metrics individually. Instead, the contribution is to expose them as a reusable interpretation layer over learned action-conditioned rollouts. This differs from pipelines in which risk terms are embedded inside a single optimization objective without being available for modular explanation and auditing.

2.3 Explainable and Neuro-Symbolic Robotics

Explainability in robotics has increasingly moved beyond feature attribution toward task-level rationales grounded in the robot internal decision process [9,10]. For safety-critical autonomy, an explanation is useful only if it is tied to the internal logic that actually produced the action. This motivates the hybrid explanation design used here, in which the language model verbalizes structured FIM traces rather than acting as a free decision-maker [11,12].

2.4 Broader Navigation under Dynamic and Cluttered Conditions

Navigation under dynamic and cluttered conditions has also been studied beyond UAV systems, including cooperative maritime and underwater robotics. Recent work on multi-USV cooperative navigation and heterogeneous AUV path planning highlights challenges that are closely related to the present problem, such as dynamic obstacle avoidance, route-risk evaluation, cooperative path selection, and navigation through cluttered environments [13,14]. Although these studies address different vehicle platforms, they are relevant because they emphasize structured reasoning under uncertainty rather than purely reactive motion selection. This broader literature supports the motivation for separating prediction, risk interpretation, and decision-making in safety-critical autonomous navigation. The present framework extends this perspective to UAV navigation by applying structured future-risk interpretation to action-conditioned world-model rollouts.

2.5 Generative AI in UAV-Enabled Secure and Intelligent Transportation Systems

Generative artificial intelligence is also being applied beyond onboard navigation, including secure authentication, privacy preservation, and adaptive services in UAV-enabled smart-city transportation. Abdusalomov et al. proposed a Generative-AI-assisted Risk-Adaptive Authentication framework for UAV-enabled intelligent transportation systems, in which contextual risk estimates are used to adapt authentication strength while reducing computation and communication overhead compared with heavier cryptographic baselines [15]. This direction is relevant to the present study because safety-critical UAV autonomy requires both intelligent onboard decision-making and secure interaction with networked transportation infrastructure. Whereas that work addresses authentication and privacy in UAV-enabled transportation networks, the present framework focuses on action-conditioned future-risk interpretation for

onboard navigation. Together, these efforts position generative-AI methods within a broader UAV ecosystem in which perception, prediction, secure communication, and human-understandable explanations must co-exist.

3 System Overview

At each time step t , the UAV receives the multimodal observation $o_t = (x_t, e_t)$, where x_t is the RGB image and e_t is the proprioceptive/inertial state vector. The goal is to select an action a_t that advances the UAV toward its task objective while respecting safety constraints under uncertainty.

The architecture contains four conceptual layers: prediction, interpretation, decision, and explanation. The prediction layer encodes multimodal observations and generates action-conditioned future rollouts. The interpretation layer, implemented by the FIM, converts these rollouts into structured risk descriptors. The decision layer filters and scores candidate actions using those descriptors. The explanation layer verbalizes the selected action, rejected alternatives, and dominant risk causes from the controller trace. Fig. 1 summarizes this four-layer organization, and Fig. 2 details the flow from predicted latent states through FIM descriptors, safety-constrained selection, and grounded explanation.

This modular separation is intentional. It allows the predictive backbone to be replaced independently of the interpretation layer, exposes interpretable quantities for auditing, and reduces the risk that explanations become detached from the actual controller logic.

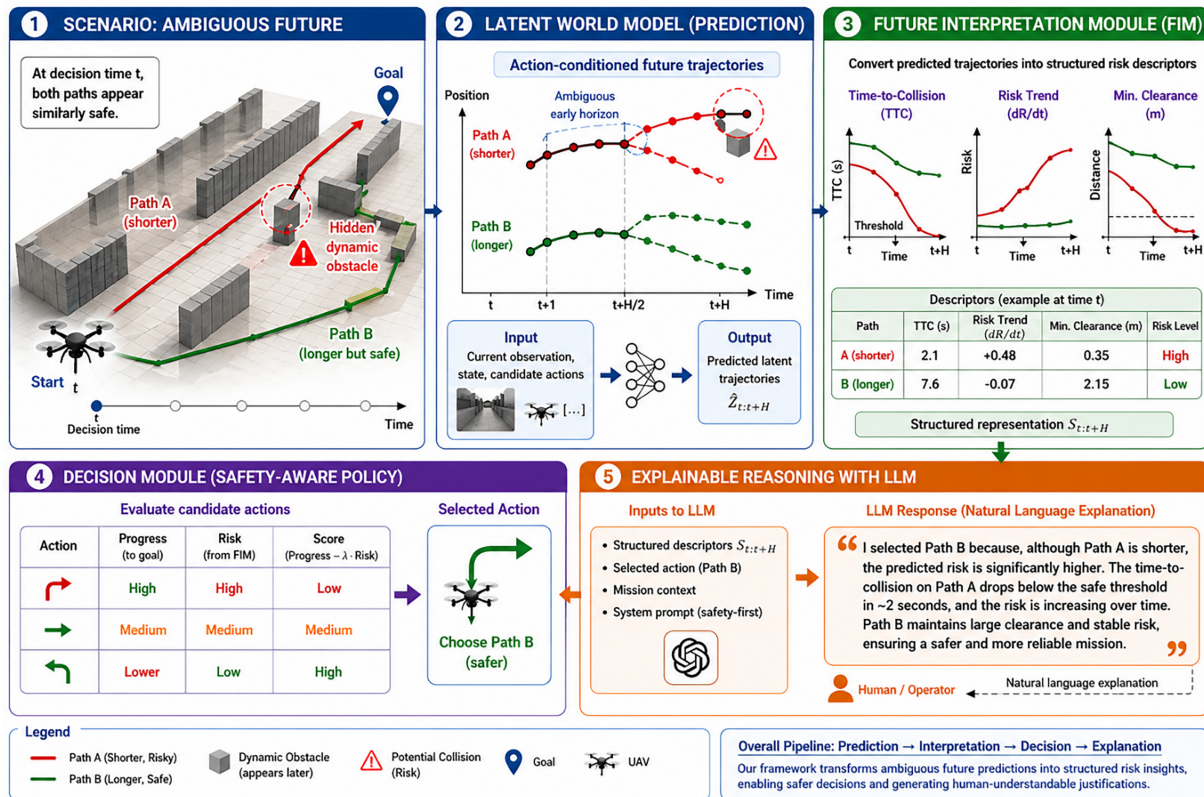


Figure 1: Simplified overview of the predictive and explainable UAV navigation framework.

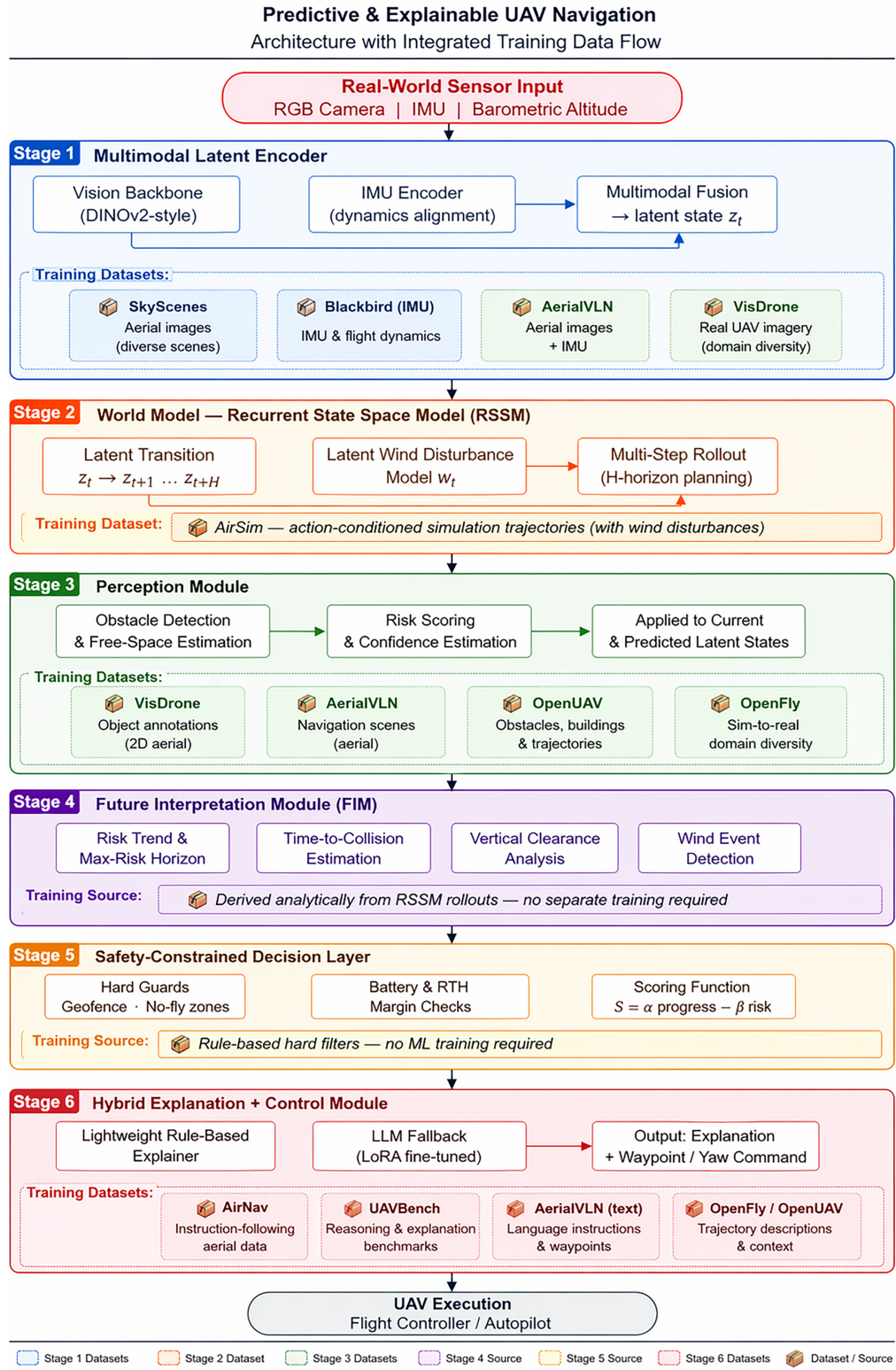


Figure 2: Detailed architecture from latent prediction to FIM reasoning, safety-constrained decision-making, and grounded explanation.

4 Methodology

The proposed framework integrates latent world modeling with structured future interpretation and grounded explanation. The methodology retains the mathematical structure of the proposed predictive-FIM framework, but the notation is normalized so that observed quantities, predicted quantities, risk descriptors, and explanation variables are clearly distinguished. The strongest methodological claim is not that individual risk metrics are new, but that action-conditioned rollouts are converted into an explicit intermediate representation used jointly for decision-making, safety filtering, and explanation grounding.

4.1 Multimodal Latent Representation

At time step t , the UAV receives a multimodal observation composed of an RGB image and proprioceptive/inertial measurements:

$$o_t = (x_t, e_t) \quad (1)$$

here, x_t denotes the image observation and e_t contains state information such as velocity, attitude, altitude, and inertial cues. A learned encoder maps the observation into a compact latent state:

$$z_t = f_\phi(o_t) \quad (2)$$

To support temporal prediction under uncertainty, the latent state is decomposed into a deterministic recurrent component h_t and a stochastic component s_t :

$$z_t = (h_t, s_t) \quad (3)$$

This representation captures both scene structure and vehicle dynamics while keeping the downstream interpretation layer independent of the visual backbone.

4.2 Latent World Model

The latent world model predicts future states conditioned on previous latent states, actions, and disturbances. The deterministic state evolves according to

$$h_t = f_h(h_{t-1}, s_{t-1}, a_{t-1}, w_{t-1}) \quad (4)$$

where a_{t-1} is the previous action and w_{t-1} is a latent disturbance variable. The prior stochastic state and posterior training state are modeled as

$$s_t \sim p_\theta(s_t | h_t) \quad (5)$$

$$s_t \sim q_\phi(s_t | h_t, o_t) \quad (6)$$

Environmental disturbances, including wind-like perturbations and unmodeled motion effects, are represented as latent variables:

$$w_t \sim p_\psi(w_t | h_t) \quad (7)$$

This disturbance-aware formulation is used to improve robustness, but it is still evaluated cautiously because simulator disturbances cannot fully reproduce real aerodynamic effects.

Transition Dynamics and Latent Rollout

For each candidate action, future states are recursively predicted over a horizon H . Predicted quantities are marked with hats to distinguish them from observed or target states:

$$\hat{h}_{t+k}^a = f_h(\hat{h}_{t+k-1}^a, \hat{s}_{t+k-1}^a, a_{t+k-1}, \hat{w}_{t+k-1}) \quad (8)$$

$$\hat{s}_{t+k}^a \sim p_\theta(s_{t+k} | \hat{h}_{t+k}^a) \quad (9)$$

$$\hat{z}_{t+k}^a = (\hat{h}_{t+k}^a, \hat{s}_{t+k}^a), k = 1, \dots, H \quad (10)$$

The action-conditioned latent rollout is therefore

$$\tau_t(a) = \{\hat{z}_{t+1}^a, \hat{z}_{t+2}^a, \dots, \hat{z}_{t+H}^a\} \quad (11)$$

This rollout is the predictive substrate used by the FIM. The decision module does not act directly on raw future images; it acts on structured descriptors extracted from these predicted futures.

4.3 Training Objective for the World Model

Training balances one-step prediction accuracy, multi-step temporal consistency, latent regularization, and disturbance robustness. The one-step prediction loss is

$$\mathcal{L}_{\text{pred}} = \mathbb{E}[\|z_{t+1} - \hat{z}_{t+1}\|_2^2] \quad (12)$$

To reduce long-horizon drift, a multi-step consistency term is used:

$$\mathcal{L}_{\text{multi}} = \sum_{k=1}^H \|z_{t+k} - \hat{z}_{t+k}\|_2^2 \quad (13)$$

The stochastic latent space is regularized using a Kullback-Leibler term between the posterior and prior distributions:

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}(q_\phi(s_t | h_t, o_t) \| p_\theta(s_t | h_t)) \quad (14)$$

To explicitly expose disturbance sensitivity during training, the model compares predictions with and without the disturbance variable:

$$\hat{z}_{t+1}^{\text{wind}} = F_\theta(z_t, a_t, w_t) \quad (15)$$

$$\hat{z}_{t+1}^{\text{no-wind}} = F_\theta(z_t, a_t, 0) \quad (16)$$

$$\mathcal{L}_{\text{wind}} = \|\hat{z}_{t+1}^{\text{wind}} - z_{t+1}\|_2^2 + \gamma \|\hat{z}_{t+1}^{\text{wind}} - \hat{z}_{t+1}^{\text{no-wind}}\|_2^2 \quad (17)$$

The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_1 \mathcal{L}_{\text{multi}} + \lambda_2 \mathcal{L}_{\text{KL}} + \lambda_3 \mathcal{L}_{\text{wind}} \quad (18)$$

This objective is used to train the predictive backbone. The FIM is then applied as an interpretation layer over the predicted rollouts, rather than being treated as a replacement for predictive learning.

4.4 Latent Perception Function

Semantic and geometric quantities are estimated from both observed and predicted latent states through a latent perception function:

$$p_t = \psi(z_t), \hat{p}_{t+k}^a = \psi(\hat{z}_{t+k}^a) \quad (19)$$

When supervised labels y_t are available, the perception head is trained using

$$\mathcal{L}_{\text{perc}} = \|\psi(z_t) - y_t\|_2^2 \quad (20)$$

where y_t encodes task-relevant quantities such as obstacle likelihood, free-space indicators, clearance proxies, semantic regions, or distance-related supervision. This design lets the same perception function be applied consistently to observed states and imagined futures.

4.5 Future Interpretation Module

The FIM converts predicted trajectories into structured descriptors. These descriptors are not introduced as new safety metrics individually; their role is to form a reusable, auditable bridge between learned future prediction and safety-constrained decision-making.

Remark 1: *The novelty of FIM is not the invention of individual safety metrics such as TTC, clearance, or risk trend. Its contribution is to organize these quantities as an explicit intermediate representation over action-conditioned predicted futures, and to reuse that representation consistently for decision selection, safety filtering, and explanation grounding.*

$$\Sigma_t(a) = g(\hat{p}_{t+1:t+H}^a) \quad (21)$$

In Eq. (21), g is a deterministic, non-learned descriptor-extraction operator. It maps the action-conditioned predicted perception sequence over the horizon to one structured descriptor vector by applying Eqs. (22)–(25) together with the collision and route-commitment tests. Eq. (26) lists the components of that same vector and therefore does not introduce a second definition.

The descriptor set includes maximum predicted risk, risk trend, time-to-collision, vertical clearance, predicted collision, and route-commitment risk. The maximum predicted risk is

$$R_{\text{max}}(a) = \max_{1 \leq k \leq H} r_{t+k}^a \quad (22)$$

where r_{t+k}^a is the predicted collision-risk score estimated from $\hat{p}_{t+k}^a$. The temporal risk trend is computed as

$$R_{\text{trend}}(a) = \frac{1}{H-1} \sum_{k=2}^H (r_{t+k}^a - r_{t+k-1}^a) \quad (23)$$

The time-to-collision descriptor is the first predicted horizon step at which the predicted collision-risk score exceeds threshold τ_r :

$$\text{TTC}(a) = \min\{k\Delta t \mid r_{t+k}^a > \tau_r\} \quad (24)$$

with $\text{TTC}(a)$ set to infinity if no collision-risk threshold crossing occurs within H . Minimum vertical clearance is represented as

$$C_{\text{vert}}(a) = \min_{1 \leq k \leq H} d_{t+k}^{\text{vert},a} \quad (25)$$

The full FIM summary is then

$$\Sigma_t(a) = \{R_{\max}(a), R_{\text{trend}}(a), \text{TTC}(a), C_{\text{vert}}(a), I_{\text{coll}}(a), R_{\text{route}}(a)\} \quad (26)$$

here, $I_{\text{coll}}(a)$ is the predicted collision indicator and $R_{\text{route}}(a)$ captures route-commitment or route-trap risk. This explicit descriptor vector is the central intermediate representation used by the safety layer and by the explanation module.

4.6 Safety-Constrained Decision Model

The admissible action set is defined by safety constraints over the FIM descriptors:

$$\mathcal{A}_{\text{safe}} = \{a \in \mathcal{A} \mid R_{\max}(a) \leq \tau_r, C_{\text{vert}}(a) \geq \tau_c, I_{\text{coll}}(a) = 0\} \quad (27)$$

Task progress is computed from the predicted rollout:

$$P(a) = h(\tau_t(a)) \quad (28)$$

Each admissible action is scored by combining progress with explicit risk and route-commitment penalties:

$$S(a) = \alpha P(a) - \beta R_{\max}(a) - \eta R_{\text{route}}(a) - \kappa [\tau_c - C_{\text{vert}}(a)] \quad (29)$$

For Eq. (29), the diagnostic implementation uses $\alpha = 1.0$, $\beta = 2.0$, $\kappa = 6.0$, and a state-dependent route-commitment coefficient $\eta = 60.0$ before the divider and $\eta = 80.0$ after route commitment. These are fixed implementation constants rather than learned parameters; Table 1 reports them explicitly. Their sensitivity is not evaluated in the present study.

The selected action is

$$a_t^* = \arg \max_{a \in \mathcal{A}_{\text{safe}}} S(a) \quad (30)$$

If $\mathcal{A}_{\text{safe}}$ is empty, the controller falls back to the predefined emergency or hover/stop behavior. This decision model is deliberately transparent: the selected action can be traced to progress, maximum risk, clearance, predicted collision, and route-commitment terms.

4.7 Explanation Model

The explanation module maps the selected action and its FIM trace to a human-readable rationale:

$$E_t^{\text{rule}} = \mathcal{E}(a_t^*, \Sigma_t(a_t^*)) \quad (31)$$

The rule-based explanation path is deterministic and low-latency. It is used whenever the FIM trace is sufficient to generate a concise grounded rationale.

4.8 LLM-Based Explanation

For high-risk, ambiguous, or semantically complex situations, the system can invoke an LLM to verbalize the structured trace. The LLM is not part of the flight controller and has no authority to select actions. Its input is restricted to structured quantities:

$$u_t = \{z_t, \Sigma_t(a_t^*), a_t^*\} \quad (32)$$

$$\text{Prompt}_t = \text{Template}(u_t) \quad (33)$$

$$E_t^{\text{LLM}} = \text{LLM}(\text{Prompt}_t) \quad (34)$$

If the explanation model is fine-tuned, the token-level cross-entropy loss is

$$\mathcal{L}_{\text{LLM}} = - \sum_j \log P(E_{t,j} | E_{t,<j}, \text{Prompt}_t) \quad (35)$$

During inference, the explanation source is selected by a hybrid rule:

$$E_t = \begin{cases} E_t^{\text{LLM}}, & R_{\max}(a_t^*) > \tau_{\text{LLM}} \text{ or } \text{ambiguity}_t = 1, \\ E_t^{\text{rule}}, & \text{otherwise.} \end{cases} \quad (36)$$

This preserves real-time control because the LLM is triggered only when needed and only verbalizes an already computed controller trace. The safety-critical action remains determined by the world model, FIM descriptors, and safety-constrained decision module.

5 Data, Training, and Implementation Details

5.1 Data Sources and Training Strategy

The three data streams are not merged into one corpus. AirSim [16] supplies action-conditioned RGB, proprioceptive-state, action, and next-state trajectories for the encoder and predictive backbone. VisDrone, AerialVLN, OpenUAV, and OpenFly supply optional aerial-domain data for perception robustness. Structured FIM controller traces supply prompt-response pairs only for explanation adaptation. The Isaac Sim episodes reported in Section 6 are evaluation data and are not used to train the explanation model.

Training is stage-wise. First, a convolutional image encoder and an MLP for inertial inputs are optimized with the GRU-based latent dynamics model on temporally ordered simulator sequences. Second, the lightweight feed-forward perception head is trained from simulator-derived geometric labels and optional aerial-domain supervision. Third, the explanation model is adapted from serialized FIM traces. The FIM itself is a deterministic inference-time layer and introduces no additional learned parameters or joint end-to-end training stage.

The documented predictive-training configuration uses Adam with a learning rate of 1×10^{-4} , batches of 32 sequences, sequence lengths of 20–50 steps, gradient clipping, KL balancing, rollout-horizon curriculum from 5 to 10 steps, and mixed-precision training on one GPU. The 18-step value in Table 1 is the diagnostic planning horizon used at inference, not the training rollout horizon. Explanation adaptation uses the separate LoRA settings reported in Section 5.3 and has no action-selection authority.

Reproducibility boundary: the archived experimental package does not contain the post-filtering trajectory counts, the frozen train/validation/test split manifest, or a fixed epoch count. These quantities therefore cannot be verified from the submitted evidence and are not inferred here. Full learning-stage replication requires release of the preprocessing manifest, split identifiers, checkpoint-selection rule, and trained checkpoints; the reported benchmark outcomes remain reproducible only from the supplied evaluation configurations and logs.

5.2 FIM Decision Parameters

Table 1 summarizes the decision parameters used in the diagnostic implementation. Clearance is reported with two distinct meanings. The controller uses predicted clearance as a soft planning margin and penalty term, whereas the tables report the realized episode-level closest approach along the executed trajectory. Therefore, realized minimum clearance values can be below the soft planning margin without implying collision; collisions are determined by the simulator contact flag and collision-radius proxy.

Table 1: FIM decision parameters used in the diagnostic implementation.

Parameter	Value	Notes
Planning horizon H	18 steps	Diagnostic benchmark horizon
Candidate action set	FORWARD, LEFT, RIGHT	Discrete proxy action interface
Action velocities	FORWARD = [1.25, 0, 0], LEFT = [1.10, 0.65, 0], RIGHT = [1.10, -0.95, 0] m/s	Isaac corridor benchmark
Simulation time step Δt	0.25 s	Benchmark setting
Collision radius	0.35 m	Benchmark collision proxy
TTC definition/safety use	Risk threshold 0.70; no hard time cut-off	First collision-risk threshold crossing defines TTC; inverse-TTC remains a continuous term
Clearance thresholds	0.35 m collision proxy; 0.60 m soft predicted-clearance planning margin	Reported minimum clearance is realized closest approach; collision is counted only when the simulator collision/contact criterion is triggered.
Route-trap lookahead	H-step rollout	Detects delayed blockage
Explanation control authority	None	LLM verbalizes traces only
Score weight α (progress)	1.0	Fixed coefficient in Eq. (29)
Score weight β (maximum risk)	2.0	Fixed coefficient in Eq. (29)
Score weight η (route commitment)	Pre-divider 60.0; post-commitment 80.0	State-dependent route-trap coefficient in Eq. (29)
Score weight κ (clearance)	6.0	Fixed coefficient on the clearance term in Eq. (29)

5.3 LLM Implementation Details

The current implementation uses a hybrid explanation module configured with Qwen/Qwen2.5-3B-Instruct for structured verbalization, TinyLlama/TinyLlama-1.1B-Chat-v1.0 for compact LoRA experiments, and sshleifer/tiny-gpt2 as a lightweight fallback path during testing. The LoRA configuration uses rank $r = 16$,

$\alpha = 32$, and dropout = 0.05. Prompts instruct the model to explain UAV decisions using only structured descriptors supplied by the FIM trace. Hardware used in development includes an Intel i7-10750H CPU and an NVIDIA RTX 3070 Laptop GPU. Runtime values are reported only where they were explicitly measured in the current logs.

5.4 Baseline Adaptation and Fairness Considerations

All methods were evaluated under the same Isaac Sim environment, action space, episode length, collision radius, and termination criteria. However, YUME, Matrix-Game, NWM, and ANWM [17–20] were not originally designed for the exact discrete UAV benchmark used here. Their native outputs were therefore adapted to a common three-action interface using a nearest-direction projection or equivalent benchmark wrapper. This unified interface ensured that all methods faced the same decision frequency, goal condition, and collision criteria.

This adaptation is necessarily imperfect. Some baselines were designed for continuous control, high-dimensional control, or richer vision-language interaction spaces. Projecting their outputs into a minimal discrete action set may reduce some of their original capabilities. The comparison is therefore not interpreted as a general claim of superiority over those methods in their native settings. The primary controlled comparison is between prediction-only WorldModel, descriptor-level ablations, and Full FIM using the same backbone, action space, and environment.

6 Experimental Evaluation

6.1 Diagnostic Ambiguous-Future-Risk Benchmark

The corridor experiment is retained as a diagnostic benchmark rather than the sole validation scenario. The environment contains a short upper route and a longer safer lower bypass. A delayed hazard enters the upper route at three ambiguity levels: easy = 4, medium = 8, and hard = 12 time steps. This creates route-trap behavior that cannot be solved reliably by myopic distance-to-goal scoring alone. Fig. 3 shows the Isaac Sim realization, while Fig. 4 presents the corresponding two-route schematic.

The reported metrics are defined as follows. Success rate (SR) is the fraction of episodes reaching the goal within the episode limit; collision rate (CR) is the fraction of episodes with simulator collision/contact events; risk anticipation accuracy (RAA) measures agreement between predicted and realized future-risk labels; decision optimality gap (DOG) is the cost gap between the selected action and the oracle reference action, so lower values are better; early risk detection time (ERDT) measures how early a future-risk event is detected; route-trap rate (RTR) reports episodes in which the selected route becomes blocked or non-recoverable before goal completion; and minimum clearance (MC) is the realized closest-approach statistic along the executed trajectory.

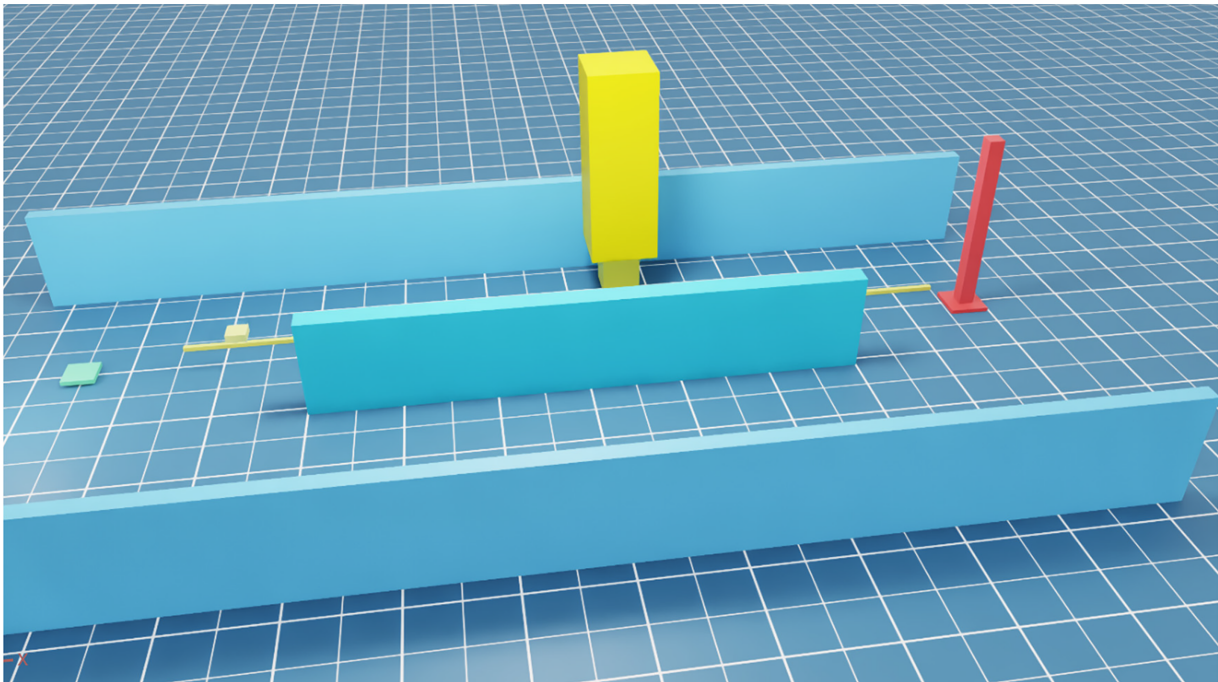


Figure 3: Isaac Sim realization of the ambiguous delayed-hazard benchmark.

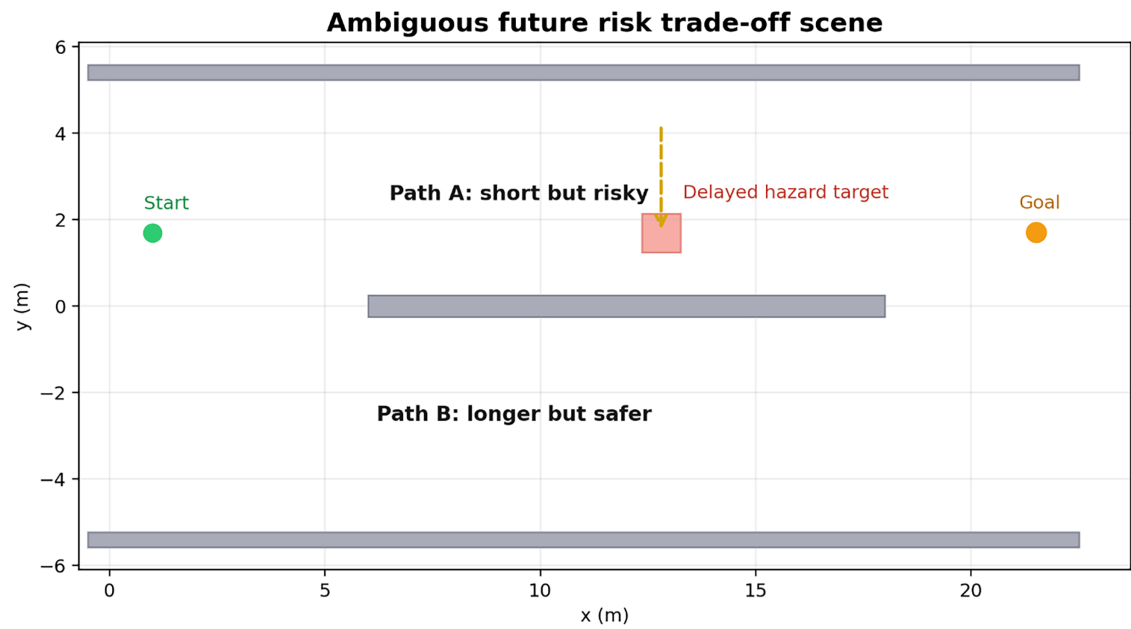


Figure 4: Schematic layout of the diagnostic corridor with short risky and longer safer routes.

6.2 Main Results with Statistical Uncertainty

For binary outcomes such as success rate and collision rate, Wilson 95% confidence intervals are reported. The updated diagnostic benchmark uses $N = 50$ matched episodes per difficulty level and per agent, giving 150 episodes per agent across the three ambiguity levels.

Across all 150 episodes per agent, the FIM policy achieves $SR = 1.000$ with Wilson 95% CI $[0.975, 1.000]$ and $CR = 0.000$ with Wilson 95% CI $[0.000, 0.025]$. The prediction-only WorldModel baseline is also strong in this diagnostic rerun, with $SR = 0.973$ and $CR = 0.027$. The observed benefit of FIM in Table 2 is therefore best interpreted as improved safety margin and decision cost rather than a large success-rate gap: mean minimum clearance increases from 0.495 to 0.577 m and DOG decreases from 9.611 to 7.051.

Table 2: Diagnostic corridor benchmark with Wilson 95% confidence intervals ($N = 50$ episodes per condition).

Difficulty	Agent	SR	95% CI (SR)	CR	95% CI (CR)	RAA	DOG	ERDT (s)
Easy	WorldModel	0.94	[0.838, 0.979]	0.06	[0.021, 0.162]	0.771	9.250	0.00
Easy	FIM (proposed)	1.00	[0.929, 1.000]	0.00	[0.000, 0.071]	0.801	7.398	0.00
Easy	Oracle	1.00	[0.929, 1.000]	0.00	[0.000, 0.071]	1.000	0.000	0.00
Medium	WorldModel	1.0	[0.929, 1.000]	0.00	[0.000, 0.071]	0.759	9.443	0.00
Medium	FIM (proposed)	1.00	[0.929, 1.000]	0.00	[0.000, 0.071]	0.821	6.294	0.00
Medium	Oracle	1.00	[0.929, 1.000]	0.00	[0.000, 0.071]	1.000	0.000	0.00
Hard	WorldModel	0.98	[0.895, 0.996]	0.02	[0.004, 0.105]	0.748	10.140	0.00
Hard	FIM (proposed)	1.00	[0.929, 1.000]	0.00	[0.000, 0.071]	0.810	7.461	0.00
Hard	Oracle	1.00	[0.929, 1.000]	0.00	[0.000, 0.071]	1.000	0.000	0.00

6.3 Adapted Baseline Comparison

The reviewer concern about asymmetric benchmarking is addressed by treating YUME, Matrix-Game, NWM, and ANWM as adapted reference methods rather than as native baselines. The supplied final CSV logs do not contain matched reruns of these methods under the revised cross-environment protocol. Therefore, Table 3 reports only the adaptation interface and interpretation limits, while the primary quantitative evidence is provided by the matched WorldModel/FIM/oracle diagnostic and ablation results in Tables 2 and 4.

Table 3: Adapted reference methods and fairness limitations under the unified UAV diagnostic interface.

Method	Native Scope/Output	Adapted Use and Interpretation in This Paper
YUME	Interactive world-generation or VLN-style generative future representation	Projected to the FORWARD, LEFT, and RIGHT diagnostic interface. Reported only as an adapted reference, not as evidence of native superiority.
Matrix-Game	Interactive world foundation model with learned action/world dynamics	Wrapped into the same three-action UAV interface and episode criteria. Interpreted only as diagnostic context because the wrapper is not its native control setting.

(Continued)

Table 3 (continued)

Method	Native Scope/Output	Adapted Use and Interpretation in This Paper
NWM	Pixel-space navigation world model	Action selection is projected to the benchmark commands. Used as a prediction-oriented reference, not as a full native re-evaluation.
ANWM	Aerial long-horizon world model for generation and navigation	Adapted to the same episode length and discrete action interface. Its native long-horizon output space is reduced by the wrapper.

Table 4: FIM ablation study over the delayed-hazard diagnostic setting (N = 150 episodes per method). SR, CR, RTR, and TO are mutually exclusive terminal outcome rates; rate columns are percentages.

Method	SR (%)	CR (%)	RTR (%)	TO (%)	MC (m)	ERDT (s)
WorldModel only	0.0	100.0	0.0	0.0	−0.177	0.000
WorldModel + TTC	37.3	0.0	62.7	0.0	0.175	0.018
WorldModel + clearance	36.0	0.0	64.0	0.0	0.213	0.012
WorldModel + risk trend	37.3	0.0	62.7	0.0	0.225	0.015
WorldModel + route penalty	0.0	100.0	0.0	0.0	−0.065	0.027
Full FIM	71.3	0.0	0.0	28.7	0.336	0.023

[Table 3](#) is intentionally limited to adaptation protocol and interpretation boundaries. The adapted references are not treated as complete re-implementations of YUME, Matrix-Game, NWM, or ANWM in their native settings. The primary controlled comparison in this manuscript is instead between the prediction-only WorldModel, descriptor-level ablations, and Full FIM, because these variants share the same predictive interface, action space, episode conditions, and evaluation protocol.

6.4 FIM Ablation Study

The ablation study isolates whether performance comes from a single descriptor or from their joint use. [Table 4](#) is populated from the supplied ablation logs with N = 150 episodes per method.

[Table 4](#) shows that single descriptors can reduce immediate collisions, but this reduction is often accompanied by route-trap outcomes rather than successful completion. Because SR, CR, RTR, and TO are mutually exclusive, each method sums to 100% across these terminal categories. Full FIM provides the most balanced result in this ablation, with SR = 71.3%, CR = 0.0%, RTR = 0.0%, TO = 28.7%, and the largest mean clearance among the non-oracle ablations. The remaining failures are timeouts, indicating conservative behavior rather than unsafe collision behavior.

6.5 Cross-Environment Generalization and Robustness Validation

The simulation study was conducted in NVIDIA Isaac Sim using a suite of structured indoor navigation environments built from primitive 3D assets. Each scenario follows a common start-goal navigation task, but the source of ambiguity differs across environments. Several layouts contain an explicit trade-off

between a shorter but potentially hazardous route and a longer safer alternative, whereas others introduce narrow passages, delayed occlusions, moving obstacles, or disturbance-induced low-clearance conditions. The validation suite includes six environments: a Dynamic Forklift Warehouse with shelf-lined aisles and delayed vehicle intrusion; an Occluded Blind-Corner Arena where tall occluders hide cross-traffic until late in the approach; a Wind-Disturbance Industrial Corridor with narrow lanes, static equipment, and transient lateral gusts; a Drone Racing/Gate Navigation Arena with tight gate-like passages; a Search-and-Rescue Semantic Arena containing debris clusters and route-trap geometry; and a Multi-Level Parking Garage with pillars, narrow bends, and delayed vehicle intrusion. Across the suite, difficulty is controlled by varying hazard onset, disturbance timing, or route-blockage timing, thereby increasing future ambiguity and forcing the UAV to trade off immediate progress against anticipated risk.

Table 5 reports Full-FIM performance across these six validation environments. Figs. 5–7 complement the table by showing representative hard-setting trajectories, temporal risk-anticipation traces, and environment-wise easy-setting aggregate metrics. Because the prediction-only baseline is already near saturation in several environments, the cross-environment results are interpreted primarily as evidence of robustness and safe generalization, rather than as a claim that Full FIM uniformly outperforms the baseline in every setting.

Table 5: Cross-environment validation of Full FIM over six environments (N = 150 episodes per environment).

Environment	Success Rate	Collision Rate	Route-Trap Rate	Min. Clearance (m)	Mean Steps	Main Failure Mode
Dynamic Forklift Warehouse	0.94	0.00	0.03	0.17	75.95	Timeout
Occluded Blind-Corner Arena	1.00	0.00	0.00	0.15	69.00	None
Wind-Disturbance Industrial Corridor	0.94	0.00	0.06	0.16	75.20	Route trap
Drone Racing/Gate Navigation Arena	1.00	0.00	0.00	0.27	72.86	None
Search-and-Rescue Semantic Arena	1.00	0.00	0.00	0.13	68.05	None
Multi-Level Parking Garage	1.00	0.00	0.00	0.23	71.17	None
Overall Weighted Mean	0.98	0.00	0.01	0.18	72.04	Timeout/route trap residual

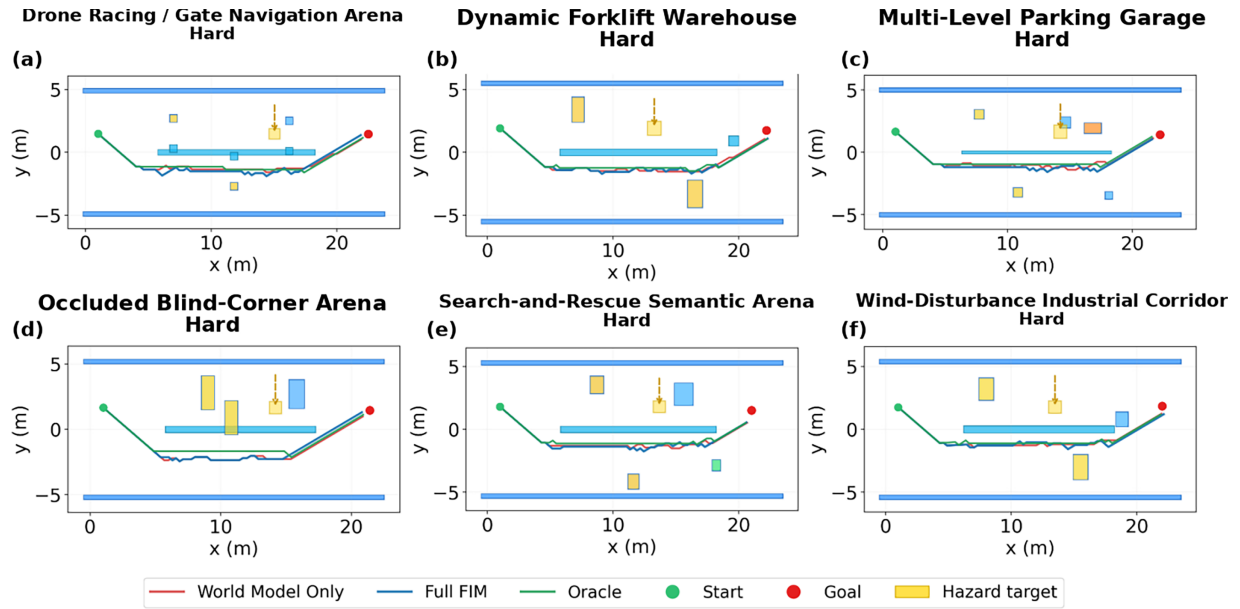


Figure 5: Cross-environment UAV trajectory overlays in the hard setting. Panels show (a) racing, (b) warehouse, (c) parking, (d) blind corner, (e) search-and-rescue, and (f) wind-disturbance. Red, blue, and green curves denote WorldModel-only, Full FIM, and Oracle, respectively.

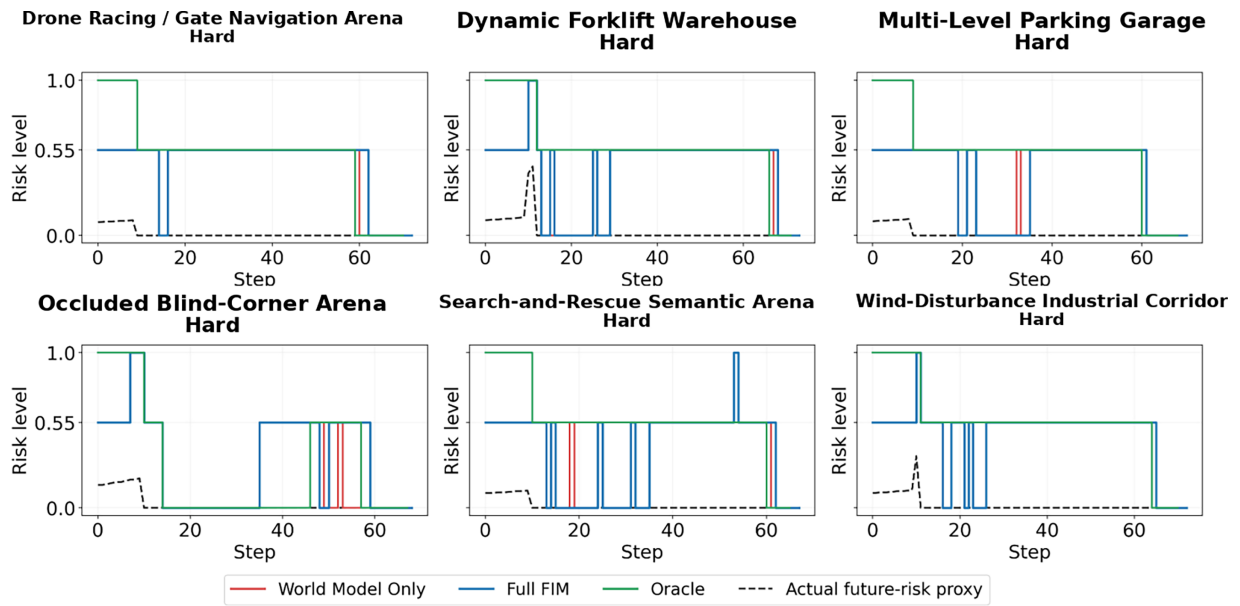


Figure 6: Representative risk-anticipation traces in the hard setting across the six validation environments. The dashed black curve denotes the actual future-risk proxy, and the colored curves show the risk level predicted or selected by each controller.

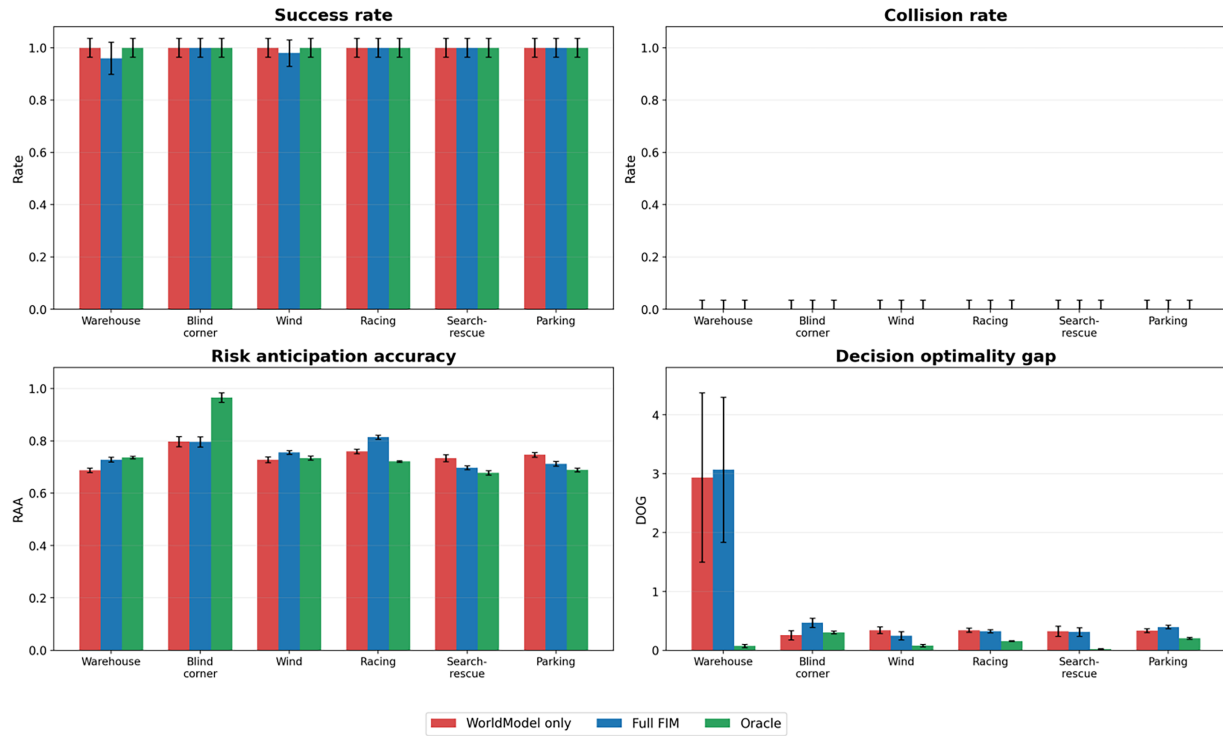


Figure 7: Quantitative comparison across the six validation environments in the easy setting. The plots report success rate, collision rate, risk anticipation accuracy (RAA), and decision optimality gap (DOG). Because the easy setting is near saturation, the figure is interpreted descriptively rather than as evidence of uniform method dominance.

Table 5 supports the robustness claim: across 900 Full-FIM episodes, the method reaches $SR = 0.980$, $CR = 0.000$, and route-trap rate = 0.014. The matched prediction-only WorldModel baseline is also near saturated in this cross-environment suite ($SR = 0.999$, $CR = 0.000$), which means this suite is not the primary evidence for success-rate superiority. In paired comparison, both methods succeeded in 881 episodes, FIM succeeds when WorldModel fails in 1 episode, and WorldModel succeeds when FIM fails in 18 episodes. The cross-environment result is therefore reported as evidence of safe generalization and interpretability transfer, not as a claim that Full FIM uniformly surpasses the baseline in every environment.

Together, Figs. 5–7 show that the cross-environment validation is not limited to aggregate success values: Full FIM remains collision-free across the reported suite, the qualitative risk traces remain interpretable, and the easy-setting aggregate plots confirm that the strongest cross-environment claim is robustness rather than universal numerical dominance over an already strong prediction-only baseline.

6.6 Explanation Evaluation

The explanation module is evaluated separately from the controller because the LLM does not control the UAV. Grounding accuracy measures whether the explanation mentions the selected action, rejected action, dominant risk cause, and relevant FIM descriptor. Consistency measures whether repeated prompts over the same structured trace yield the same rationale. Hallucination rate measures the fraction of explanations that mention unsupported causes. Latency is reported separately for rule-based and LLM-based explanations. Representative trace-grounded cases are reported in Table 6.

Table 6: Representative grounded explanations generated from structured FIM traces.

Scenario	Selected Action	Rejected Action	Dominant FIM Reason	Generated Explanation	Grounded
Delayed upper-lane hazard	RIGHT	FORWARD	Route-commitment risk + shrinking TTC	The UAV turns right because the forward route will soon be blocked by the delayed obstacle, and remaining on the upper lane would increase collision risk.	Yes
Narrow passage with low clearance	LEFT	FORWARD	Minimum clearance below threshold	The UAV shifts left because the forward corridor does not preserve sufficient clearance, and the left alternative maintains a safer margin.	Yes
Safe open segment	FORWARD	LEFT	Higher progress with no active risk event	The UAV continues forward because no imminent collision or low-clearance event is predicted, and this action gives the best progress toward the goal.	Yes
Wind-disturbance corridor	Lateral correction	FORWARD	Wind-drift risk + route-trap warning	The UAV selects a lateral correction because the wind-disturbance corridor increases route-commitment risk; the direct forward option is rejected to avoid drifting into a low-clearance region.	Yes

Table 6 provides a qualitative trace-grounding check: an explanation is marked ‘Yes’ only when its selected action, rejected alternative, and dominant FIM cause agree with the structured controller trace. No aggregate grounding statistic is reported because the available annotation manifest does not support an auditable calculation. Human-rated explanation quality remains future work.

6.7 Runtime and Computational Efficiency

Because the framework is intended for onboard deployment, runtime is reported separately from navigation accuracy. The supplied navigation CSV logs include an end-to-end prototype control-loop range but do not isolate all per-component timings. Table 7 therefore reports only measured or explicitly available timing status and avoids unsupported real-time deployment claims.

Table 7: Runtime status and computational reporting for the current prototype.

Component	Reported Latency	Frequency/Trigger	Hardware Context
Perception + latent encoding	Included in end-to-end timing	Every control cycle	Intel i7-10750H + RTX 3070 Laptop GPU
Rollout + FIM descriptor computation	Included in end-to-end timing	Every control cycle	Intel i7-10750H + RTX 3070 Laptop GPU
Safety filtering + action selection	Included in end-to-end timing	Every control cycle	Intel i7-10750H + RTX 3070 Laptop GPU
End-to-end synchronous decision pipeline	50–60 ms	Every control cycle	Prototype implementation
Rule-based explanation	Not separately isolated in current logs	When explanation is requested	Deterministic fallback; no control authority
LLM-based explanation	Not measured in submitted logs	High-risk or ambiguous states only	Asynchronous; no control authority

The current logs report the end-to-end synchronous decision latency rather than isolated per-component profiling. The measured prototype decision pipeline runs in approximately 50–60 ms per control cycle on an Intel i7-10750H CPU with an NVIDIA RTX 3070 Laptop GPU. This timing includes perception, latent encoding, rollout evaluation, FIM descriptor computation, safety filtering, and action selection. The LLM explanation path is excluded from the synchronous control loop because it is triggered only for high-risk or ambiguous states and has no authority over action selection.

Scalability depends primarily on the number of candidate actions, the rollout horizon, and the cost of applying perception and risk heads to each predicted state. In the current discrete implementation, rollout and FIM descriptor computation scale approximately linearly with the product of the action-set size and the prediction horizon. Larger environments mainly increase cost when they require more candidate routes, larger local planning crops, or more hazards to be represented in the local state. Denser obstacle fields can increase perception and risk-scoring load and can make the safety filter more conservative, which may raise timeout rates. Longer horizons improve delayed-hazard anticipation but also increase rollout cost and model-drift sensitivity. Practical deployment should therefore combine bounded local horizons, hierarchical route selection, descriptor caching, and batched GPU rollout evaluation. The reported 50–60 ms timing should be interpreted as a prototype result for the diagnostic local-planning settings, not as a guarantee for much larger maps, dense obstacle distributions, continuous action spaces, or substantially longer prediction horizons.

Table 7 is interpreted as a runtime reporting status rather than a deployment benchmark. A dedicated timing benchmark with per-component instrumentation is required before making real-time deployment claims beyond the prototype end-to-end control-loop range reported here.

7 Discussion and Limitations

The completed results support a more nuanced conclusion. In the descriptor ablation, Full FIM clearly provides the most balanced safety-performance trade-off by combining multiple descriptors rather than relying on TTC, clearance, risk trend, or route penalty alone. In the larger cross-environment validation, Full FIM achieves high robustness and zero collisions, but the prediction-only WorldModel baseline is also near

saturated. Therefore, the cross-environment table is interpreted as evidence that FIM generalizes safely, not as evidence of uniform superiority over the baseline in every environment.

FIM improves interpretability because each action can be traced to explicit descriptors such as time-to-collision, clearance, risk trend, predicted collision, and route-commitment state. This also supports explanation grounding: the language model explains the controller trace rather than inventing a post-hoc rationale.

Several limitations remain. First, the completed evidence is simulation-based and does not prove real-world deployment readiness. Real-world validity of descriptor logic and world-model predictions remains to be demonstrated. Second, the simulator uses simplified discrete control, proxy collision geometry, and realized closest-approach clearance statistics that should not be confused with the controller soft planning margin. Third, the cross-environment baseline is strong enough that success-rate superiority is not established in that suite. Fourth, the adapted foundation-model comparison does not evaluate those methods in their native interfaces. Fifth, LLM latency, hallucination risk, and human-perceived explanation quality still require dedicated timing and user-evaluation studies.

Future work will add hardware-in-the-loop validation, indoor UAV experiments with motion-capture or fiducial ground truth, domain randomization for texture and disturbance, continuous-action control, multi-UAV coordination, adaptive descriptor learning, formal explanation audits, human-centered evaluation of explanation quality, and a dedicated per-component runtime benchmark.

8 Conclusion

This paper introduced a predictive and explainable UAV navigation framework that separates future prediction, structured risk interpretation, safety-constrained decision-making, and explanation generation. The central premise is that predicting future states is not sufficient for safety-critical UAV navigation unless those predictions are transformed into structured, decision-relevant risk information. The proposed Future Interpretation Module (FIM) provides this intermediate layer by converting action-conditioned world-model rollouts into interpretable descriptors, including time-to-collision, minimum vertical clearance, risk trend, predicted collision, and route-commitment risk. These descriptors are used jointly for action selection, safety filtering, and grounded explanation, thereby improving the traceability of UAV decisions.

The experimental results show that descriptor-level integration is more effective than relying on individual risk descriptors in isolation. In the ablation study, Full FIM achieved the most balanced safety-performance trade-off, reaching 71.3% success while eliminating collisions and route traps. Across the six-environment validation suite, Full FIM achieved 98.0% success and 0.0% collisions over 900 episodes. The cross-environment trajectory overlays, temporal risk traces, and aggregate quantitative plots further indicate that the same interpretation mechanism remains applicable across warehouse, blind-corner, wind-disturbance, racing, search-and-rescue, and parking-garage scenarios.

The scope of these conclusions is intentionally bounded. Comparisons with adapted YUME, Matrix-Game, NWM, and ANWM are used only as diagnostic references under a unified UAV action interface, not as complete re-evaluations of those methods in their native settings. The strongest supported conclusion is that structured future-risk interpretation improves diagnostic delayed-risk reasoning, explanation grounding, and safety-constrained traceability. The broad cross-environment results support robust collision-free execution within the evaluated simulation suite, but they do not establish universal superiority over strong prediction-only baselines.

Future work should move from simulation toward staged real-world validation. The next steps include hardware-in-the-loop experiments with onboard computing and flight-control interfaces, indoor UAV trials

with continuous velocity and yaw commands, and extension from single-UAV navigation to multi-UAV coordination under shared risk and communication constraints. Human-centered studies are also needed to evaluate whether FIM-grounded explanations improve operator understanding, trust calibration, and decision support during safety-critical navigation. Together, these directions will help assess whether the proposed prediction–interpretation–decision–explanation pipeline can support reliable deployment beyond the current simulated benchmarks.

Acknowledgement: Not applicable.

Funding Statement: Ongoing Research Funding program, (ORF-2026-698), King Saud University, Riyadh, Saudi Arabia and ISEN Toulon Méditerranée, Toulon, France.

Author Contributions: Conceptualization, Rihem Farkh; methodology, Rihem Farkh; investigation, Rihem Farkh; writing—original draft, Rihem Farkh; writing—review and editing, Ghislain Oudinet, Alaeddine Moussa and Yasser Fouad; supervision, Yasser Fouad. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The diagnostic, ablation, and cross-environment simulation logs used in the reported tables are available from the corresponding author upon reasonable request. The anonymized CSV logs and environment-configuration scripts are being prepared for public archival.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hafner D, Lillicrap T, Fischer I, Villegas R, Ha D, Lee H, et al. Learning latent dynamics for planning from pixels. In: Proceedings of the 36th International Conference on Machine Learning; 2019 Jun 9–15; Long Beach, CA, USA. p. 2555–65.
2. Hafner D, Lillicrap T, Ba J, Norouzi M. Dream to control: learning behaviors by latent imagination. arXiv:1912.01603. 2020.
3. Kaufmann E, Loquercio A, Ranftl R, Dosovitskiy A, Koltun V, Scaramuzza D. Deep drone racing: learning agile flight in dynamic environments. *IEEE Trans Robot*. 2020;36(1):1–14. doi:10.1109/TRO.2019.2942989.
4. Kahn G, Villaflor A, Ding B, Abbeel P, Levine S. Self-supervised deep reinforcement learning with generalized computation graphs for robot navigation. In: Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA); 2018 May 21–25; Brisbane, QLD, Australia. p. 1–8.
5. Ames AD, Xu X, Grizzle JW, Tabuada P. Control barrier function based quadratic programs for safety critical systems. *IEEE Trans Autom Control*. 2017;62(8):3861–76. doi:10.1109/TAC.2016.2638961.
6. Fisac JF, Akametalu AK, Zeilinger MN, Kaynama S, Gillula J, Tomlin CJ. A general safety framework for learning-based control in uncertain robotic systems. *IEEE Trans Autom Control*. 2019;64(7):2737–52. doi:10.1109/TAC.2018.2876389.
7. Schwarting W, Alonso-Mora J, Rus D. Planning and decision-making for autonomous vehicles. *Annu Rev Control Robot Auton Syst*. 2018;1(1):187–210. doi:10.1146/annurev-control-060117-105157.
8. García J, Fernández F. A comprehensive survey on safe reinforcement learning. *J Mach Learn Res*. 2015;16(42):1437–80. doi:10.5555/2789272.2886795.
9. Miller T. Explanation in artificial intelligence: insights from the social sciences. *Artif Intell*. 2019;267(2):1–38. doi:10.1016/j.artint.2018.07.007.
10. Anjomshoe S, Najjar A, Calvaresi D, Främling K. Explainable agents and robots: results from a systematic literature review. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems; 2019 May 13–17; Montreal, QC, Canada. p. 1078–88. doi:10.65109/kczb5817.

11. Huang W, Abbeel P, Pathak D, Mordatch I. Language models as zero-shot planners: extracting actionable knowledge for embodied agents. In: Proceedings of the 39th International Conference on Machine Learning; 2022 Jul 17–23; Baltimore, MD, USA. p. 9118–47.
12. Ichter B, Brohan A, Chebotar Y, Finn C, Hausman K, Herzog A, et al. Do as I can, not as I say: grounding language in robotic affordances. In: Proceedings of the 6th Conference on Robot Learning; 2023 Nov 6–9; Atlanta, GA, USA. p. 287–318.
13. Yin S, Xiang Z. ADE-MIFS: real-time cooperative navigation for multi-USVs in dynamic maritime environments. *Ocean Eng.* 2026;352(3):124487. doi:10.1016/j.oceaneng.2026.124487.
14. Yin S, Xiang Z. Multi-objective collaborative path planning for heterogeneous autonomous underwater vehicles in cluttered environments. *Swarm Evol Comput.* 2026;100:102251. doi:10.1016/j.swevo.2025.102251.
15. Abdusalomov A, Zohirov K, Ochilova S, Oramov J, Ruziyev Z, Rustamova M, et al. Generative AI for efficient and secure authentication in UAV-enabled smart city transportation systems. *Comput Mater Contin.* 2026;88(2):46. doi:10.32604/cmc.2026.081292.
16. Shah S, Dey D, Lovett C, Kapoor A. AirSim: high-fidelity visual and physical simulation for autonomous vehicles. In: Field and service robotics. Cham, Switzerland: Springer International Publishing; 2018. p. 621–35. doi:10.1007/978-3-319-67361-5_40.
17. Mao X, Lin S, Li Z, Li C, Peng W, He T, et al. Yume: an interactive world generation model. arXiv:2507.17744. 2025.
18. Zhang Y, Peng C, Wang B, Wang P, Zhu Q, Kang F, et al. Matrix-game: interactive world foundation model. arXiv:2506.18701. 2025.
19. Bar A, Zhou G, Tran D, Darrell T, LeCun Y. Navigation world models. arXiv:2412.03572. 2024.
20. Zhang W, Tang P, Zeng X, Man F, Yu S, Dai Z, et al. Aerial world model for long-horizon visual generation and navigation in 3D space. arXiv:2512.21887. 2025.