# FIRENAV: DECISION-ORIENTED RECONSTRUCTION EVIDENCE ROUTING FOR VISION-LANGUAGE NAVIGATION IN INDOOR FIRE ENVIRONMENTS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

In real-world indoor fire environments, visual degradation severely impairs instruction grounding, waypoint prediction, and long-horizon exploration in vision-language navigation (VLN). Although image restoration offers a natural means to mitigate such visual degradation, existing methods primarily optimize reconstruction fidelity, which does not necessarily translate into better navigation decisions and may even corrupt task-critical cues when applied indiscriminately. In this paper, we present FireNav, a decision-oriented reconstruction evidence routing framework that aligns visual restoration with navigation utility in VLN. To advance this task, we first contribute FireVLN-Bench, a paired indoor fire navigation benchmark with clean references preserving task semantics, safety supervision, and matched settings for continuous navigation and robot execution with simulated dynamics. Our FireNav retains native visual tokens and treats latent reconstruction features as auxiliary evidence, where a lightweight router leverages visual features, task context, and prior execution states to determine whether such evidence should be admitted and how strongly it should contribute to the visual tokens. To learn when reconstruction is beneficial, we further introduce decision utility calibration, which compares action prediction losses with and without candidate evidence under the same state, thereby supervising evidence selection according to its contribution to waypoint, heading, and arrival prediction. Experiments demonstrate that FireNav consistently outperforms state-of-the-art baselines in both navigation success and path efficiency across evaluation settings.

## 1 INTRODUCTION

Vision-language navigation (VLN) requires embodied agents to ground natural language instructions in visual observations and translate them into navigation actions (Zeng et al., 2026; Guo et al., 2026b). Recent methods have substantially advanced spatial reasoning, state modeling, and action planning, enabling increasingly capable navigation in complex indoor environments (Xue et al., 2026; Wei et al., 2026). However, most existing VLN methods assume clear or nominal visual observations, leaving robustness under adverse visual conditions less explored.

Adverse visual conditions in embodied navigation can propagate beyond perception to reasoning, decision making, and action execution (Li et al., 2026a; Wang et al., 2026c). Such errors may induce incorrect waypoints or stopping decisions and further affect subsequent observations (Wang et al., 2025; He et al., 2026). This issue is particularly critical in indoor fire environments, where smoke and poor visibility may obscure doorways, people, signs, or obstacle boundaries. Therefore, visual reconstruction for VLN should not only improve visibility, but also preserve task-relevant hazards, targets, and structural cues required for reliable navigation.

A natural solution is to directly apply existing image restoration methods to degraded observations before navigation. Recent approaches can handle heterogeneous degradations through task-aware prompting, adaptive routing, and degradation priors (Dong et al., 2026; Lin et al., 2026; Chen et al., 2025; Wang et al., 2026a). However, better reconstruction quality does not necessarily lead to better navigation decisions. Restoring every observation may introduce unnecessary computation, enhancing irrelevant regions may not benefit the next action, and unconditional fusion of reconstructed fea-



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tures may even disturb reliable native representations. This motivates a key question: *when should reconstruction influence VLN, and how strongly should it modify the native visual representation?*

In this paper, we investigate robust VLN in indoor fire environments. Specifically, we first introduce **FireVLN-Bench**, a paired benchmark that combines diverse indoor scenes, hazard layouts, and evolving visual degradations with clean references that preserve fire- and task-relevant semantics. Note that paired observations share the same camera poses and scene states, while matched episodes support continuous navigation and robot execution under simulated dynamics. Our benchmark provides shortest and risk-aware trajectories, enabling evaluation of both navigation efficiency and safety under consistent execution settings.

Building on FireVLN-Bench, we present **FireNav**, a robust framework for decision-oriented reconstruction evidence routing. Our method retains native visual tokens and treats latent reconstruction features as auxiliary evidence. Cross-attention extracts reconstruction evidence within each view, while a lightweight router uses visual features, task context, and prior execution states to determine where and how strongly such evidence should influence navigation. Different from existing methods that supervise decisions through action quality or task progress (Li et al., 2026b; Wang et al., 2026b; Guo et al., 2026b; Liang et al., 2026), FireNav explicitly measures the incremental utility of reconstruction for the current decision. At the same navigation state, predictions with and without candidate reconstruction evidence are compared through waypoint, heading, and arrival losses to supervise evidence routing. For efficient evidence integration, we adopt bounded residual fusion with zero-initialized evidence contribution and regularize the policy on clean observations, while incremental feature caching avoids redundant visual encoding. Experiments show that FireNav improves navigation success and path efficiency over reconstructed-RGB and capacity-matched adapter baselines, while preserving performance on clean observations with limited inference overhead.

Our main contributions are summarized as follows.

- We construct a high-quality VLN benchmark, FireVLN-Bench, for indoor fire navigation with task-preserving clean references, matched degraded observations, and consistent continuous and physical navigation settings.
- We propose FireNav, a decision-oriented reconstruction evidence routing framework that selectively injects reconstruction features into a frozen navigator while preserving its native visual interface for robust indoor fire VLN.
- We develop decision utility calibration and bounded residual fusion to supervise evidence admission based on its impact on action prediction and to constrain the resulting changes to native visual representations.
- Extensive experiments demonstrate that FireNav consistently improves navigation success and path efficiency under indoor fire conditions, establishing an effective bridge between visual reconstruction and robust VLN.

## 2 Related Work

**Vision-Language Navigation.** VLN maps visual observations conditioned on language to navigation actions and now spans discrete graph navigation, continuous control, open vocabulary search, persistent exploration, and vision-language-action policies. Recently, OmniNav (Xue et al., 2026) unifies instruction goal, object goal, and point goal navigation with frontier exploration through a fast policy for continuous waypoints and a slow planner with memory. DualVLN (Wei et al., 2026) adopts a dual system design that decouples VLM grounding from frequent trajectory execution using explicit pixel goals and implicit latent goals. However, most mainstream VLN methods assume clear or nominal visual observations, leaving robustness under adverse visual conditions less explored. AlldayWalker (Wang et al., 2026c) addresses this problem by adapting a navigator across low light, overexposed, and scattering environments while mitigating forgetting across scenarios. Such adaptation highlights the need for robust visual representations but does not study how reconstruction features can support closed loop navigation without replacing the pretrained visual pathway. In this paper, we explore robust VLN under degradation in indoor fire environments, where compound visual degradations jointly challenge perception and decision making.

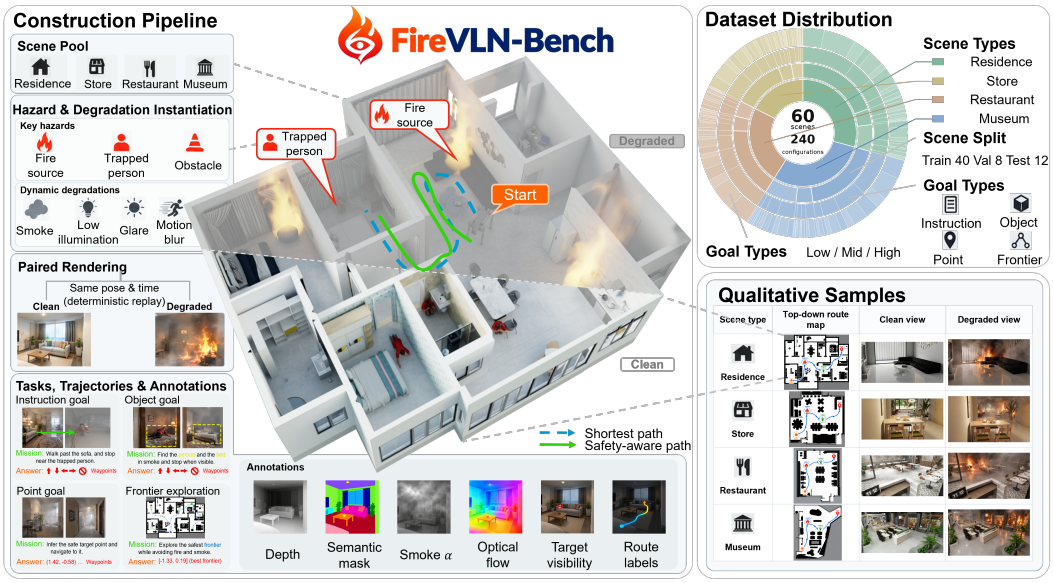



Figure 1: Overview of FireVLN-Bench. The figure shows construction pipeline indoor fire scene dataset statistics and qualitative samples.

**Visual Reconstruction.** Visual reconstruction aims to recover visibility, structure, and appearance from degraded observations, with most existing methods optimizing image fidelity or perceptual quality. This objective also extends to fire scenes obscured by smoke. Huang et al. (Huang et al., 2021) propose DesmokeGAN, which uses attention-guided adversarial learning for image desmoking and smoke detection, while SmokeBench (Jin et al., 2025) provides aligned real-world smoke/clean surveillance pairs for early-stage fire desmoking. Given the shared scattering and contrast attenuation effects of smoke and haze, recent dehazing methods also offer relevant reconstruction priors. IPC-Dehaze (Fu et al., 2025) iteratively refines discrete visual codes for real-world dehazing, while DCL (Fan et al., 2025) couples atmospheric scattering with temporal brightness consistency via shared depth estimation. Despite improving visibility and perception, these methods largely treat reconstructed images as final outputs. In contrast, embodied VLN requires decision-oriented reconstruction. Thus, we use reconstruction features as optional evidence and evaluate their utility in closed-loop navigation through navigation supervision and task-preservation constraints.

**Evidence Routing.** Moving from perceptual restoration to decision-oriented reconstruction raises a further question of which recovered features should influence the task representation. UniRestore (Chen et al., 2025) extends restoration toward downstream tasks through task aware feature adaptation and fusion. TaskTok (Lee et al., 2026) selectively refines restoration tokens relevant to the task rather than treating the full latent sequence uniformly. This selective use of latent evidence also appears in embodied decision making. DualVLN (Wei et al., 2026) transfers compact latent goals from a slow VLM planner to a fast trajectory policy. $x^2$-Fusion (Guo et al., 2026a) adjusts heterogeneous feature contributions according to their reliability under degradation. In indoor fire scenes, reconstruction cues may recover obscured evidence but also distort semantics critical to the task. FireNav therefore introduces decision-guided reconstruction evidence routing to use reconstruction evidence only when it is reliable and beneficial to navigation decisions.

## 3 Proposed Benchmark: FireVLN-Bench

**Overview.** To study robust VLN in indoor fire environments, we construct FireVLN-Bench, a paired benchmark that evaluates the effects of visual degradation and reconstruction on perception and navigation in a closed loop. In Isaac Sim, matched settings cover diverse scenes, hazard layouts, and temporal degradations for continuous navigation and robot execution with simulated dynamics. These settings support controlled evaluation of robustness, navigation, safety, and performance recovery through reconstruction. Figure 1 summarizes the benchmark construction pipeline, dataset distribution, and qualitative examples.

**Indoor Fire Environment Modeling.** We denote an environment configuration by $\mathcal{E}$, comprising scene geometry, materials, collision assets, and a fixed layout of fire sources, trapped persons, and task obstacles. For each camera view, degraded and clean observations at time $t$ are rendered as

$$I_t^d = \mathcal{R}(\mathcal{E}, p_t, \mathcal{D}_t), \\ I_t^c = \mathcal{R}(\mathcal{E}, p_t, \mathcal{D}_t^c), \tag{1}$$

where $\mathcal{R}$ denotes rendering and image formation, $p_t$ is the camera pose, and $\mathcal{D}_t$ specifies smoke, illumination, glare, sensor noise, and motion blur. The reference state $\mathcal{D}_t^c$ suppresses degradations while retaining fire cues and task semantics. Both observations share $\mathcal{E}$, $p_t$, and the fire and object states. This pairing applies to recorded frames and controlled trajectory replay.

**Benchmark Construction.** FireVLN-Bench is built from indoor scenes curated from SpatialVerse InteriorAgent and SAGE-3D InteriorGS USDZ (SpatialVerse Research Team, Manycore Tech Inc., 2025; Miao et al., 2026), including residences, convenience stores, restaurants, and museums. We retain only scenes with valid collision geometry and tasks that are feasible under clean observations. InteriorAgent provides editable mesh scenes, while SAGE-3D combines 3D Gaussian appearance with aligned collision assets.

To simulate fire conditions, we combine each scene with multiple hazard layouts to form configurations $\mathcal{E}$. Seeded schedules control fire intensity and the evolution of $\mathcal{D}_t$. Using the renderer $\mathcal{R}$ in Eq. equation 1, we generate paired observations $(I_t^d, I_t^c)$ at each recorded pose $p_t$. The clean rendering state $\mathcal{D}_t^c$ produces the reference $I_t^c$ for task preservation. We provide diagnostic images with visual degradation disabled and annotations including smoke alpha, depth, semantic masks, optical flow, and target visibility. Each configuration is checked for deterministic replay, connectivity, collision consistency, and pair alignment. Episodes include a shortest trajectory and an alternative that accounts for collision risk, fire proximity, dense smoke, blocked routes, and excessive turning.

**Task Formulation and Splits.** Following continuous and physically embodied VLN settings (Krantz et al., 2020; Wang et al., 2025), FireVLN-Bench contains 60 unique indoor scenes with four hazard layouts each, yielding 240 configurations $\mathcal{E}$. It provides two matched tracks whose canonical episodes share the environment configuration, task, initial state, and seeded degradation schedule. FireVLN-Bench-CE evaluates continuous navigation without teleportation between navigation graph nodes, prior knowledge of the environment topology, or oracle control of local motion. FireVLN-Bench-PE executes the same tasks using simulated robot dynamics and a locomotion controller, introducing contact dynamics, actuation error, camera disturbance, falls, and deadlocks.

The benchmark covers instruction following, object search, point goal navigation, and frontier exploration. For FireNav, the input image set $\mathcal{I}_t$ comprises selected history and current views from the degraded observation stream $I_t^d$. Each sample contains up to 20 history frames with their recorded poses $p_t$, current views from the left, front, and right, and five future local waypoints with heading and arrival labels. Corresponding references $I_t^c$ support reconstruction evaluation and comparisons under clean observations. Following protocols for generalization across environments (Ku et al., 2020), the scenes are split into 40 training scenes, 8 unseen scenes for validation, and 12 unseen scenes for testing, corresponding to 160, 32, and 48 hazard configurations. Validation on seen scenes uses separate episodes from the training environments. Splits are stratified by scene category and asset source. All configurations $\mathcal{E}$ derived from the same scene family remain in the same scene partition. Clean and degraded pairs, CE and PE counterparts, and nearby trajectories remain within the same episode split. Episodes involving extended navigation additionally provide rotational views, candidate frontiers, memory proxies, subgoals, and labels for safety and information gain.

# 4 Proposed Method: FireNav

## 4.1 Overall Framework

As shown in Figure 2, FireNav augments a pretrained navigator with reconstruction evidence while retaining its native visual pathway. We instantiate the framework on OmniNav (Xue et al., 2026) and preserve its visual encoder, vision-language backbone, waypoint interface, and planning hierarchy. The navigator and reconstruction encoder remain frozen, while the evidence interface and router are learned, limiting adaptation to auxiliary representations (Wei et al., 2026; Huang et al., 2026). At

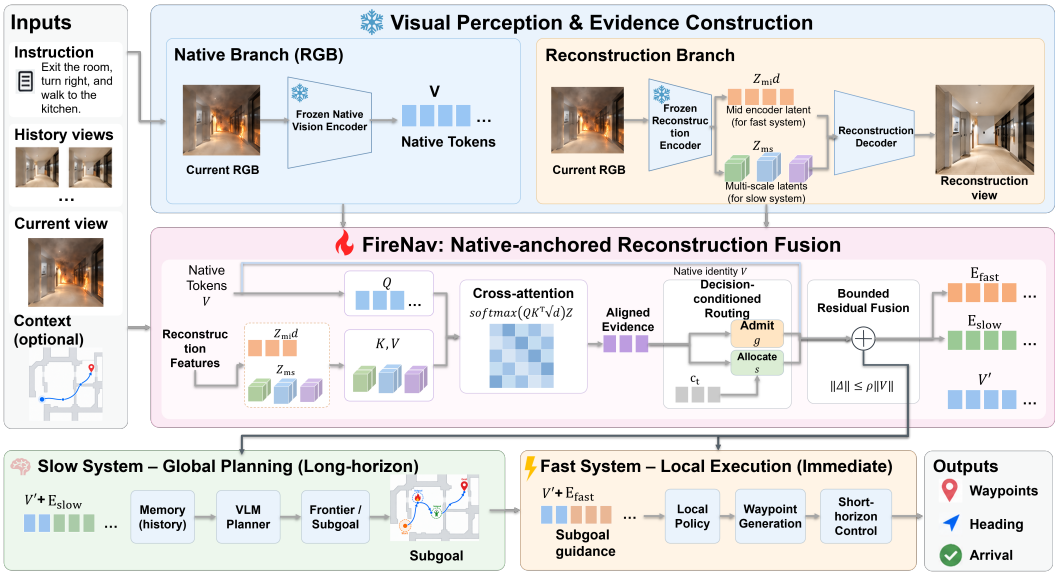



Figure 2: Overview of FireNav. A frozen native branch provides visual tokens, while a frozen reconstruction branch exposes mid-level features for fast execution and multiscale features for slow planning. FireNav aligns reconstruction features with native tokens, routes them according to the current decision state, and injects bounded residual evidence while preserving the native visual identity. The slow system uses multiscale evidence for memory-based subgoal planning, whereas the fast system uses mid-level evidence for local waypoint prediction.

step $t$, the agent receives a task condition $\ell_t$ and current and historical images $\mathcal{I}_t = \{I_{t,i}\}$ with pose annotations. Intermediate encoder features support fast execution, and multiscale features support slow planning. The slow system selects subgoals from memory and candidate frontiers. The fast system predicts $H = 5$ planar waypoints with headings and arrival probabilities, denoted by $\widehat{Y}_t$.

**Native Token Interface.** Reconstruction features provide an alternative to using restored RGB as the navigation input (Chen et al., 2025; Lee et al., 2026). For image $I_{t,i}$, the native encoder produces $V_{t,i} \in \mathbb{R}^{N_i \times d_v}$. An alignment module projects and pools the reconstruction features into tokens $Z^b_{t,i} \in \mathbb{R}^{M^b_i \times d_a}$ for branch $b \in \{\text{fast}, \text{slow}\}$. Native tokens retrieve evidence from the same image through cross attention (Jiang et al., 2026; Chen et al., 2025),

$$Q^b_{t,i} = V_{t,i} W^b_Q, \qquad C^b_{t,i} = \mathrm{MHA}\left(Q^b_{t,i}, Z^b_{t,i}, Z^b_{t,i}\right). \tag{2}$$

The evidence length $M^b_i$ need not equal $N_i$. The readout occupies the original visual token slots, leaving text, coordinate tokens, and sequence length unchanged. Appendix A.1 details feature selection and token construction.

### 4.2 Decision-Guided Evidence Routing

Adaptive routing allocates visual information according to input demands and reliability (Wang et al., 2026d; Guo et al., 2026a). FireNav additionally learns whether that information improves the navigation prediction. Decision context guides admission and spatial allocation, utility calibration supervises evidence selection, and residual control limits the representation change.

**Decision Conditioning.** Navigation depends on task progress and previous actions as well as current observations (Wang et al., 2026b; Li et al., 2026b). We condition image admission on visual statistics $q_{t,i}$ and context $c_t$, which includes the task, an available subgoal, the previous waypoint prediction, and measured execution progress. Previous waypoints are expressed in the current body frame, and unavailable context is masked. The context excludes the current prediction, clean references, and future outcomes. For each branch, the gates are

$$\begin{aligned} g^b_{t,i} &= \sigma\left(f^b_g([\bar{Q}^b_{t,i}, \bar{C}^b_{t,i}, q_{t,i}, c_t])\right), \\ s^b_{t,i,j} &= \sigma\left(f^b_s([Q^b_{t,i,j}, C^b_{t,i,j}])\right), \end{aligned} \tag{3}$$

where $\sigma$ denotes the sigmoid function, bars indicate token averages, brackets denote concatenation, and $f_g^b, f_s^b$ are lightweight MLPs. The scalar $g_{t,i}^b$ sets the contribution of an image, and $s_{t,i,j}^b$ distributes it across native tokens.

**Decision Utility Calibration.** Decision supervision can use action quality beyond direct imitation (Li et al., 2026b; Wang et al., 2026b). We supervise admission through the change in action prediction caused by candidate reconstruction evidence. A candidate intervention adds spatially weighted reconstruction features to the native tokens through bounded residual fusion, defined below in Eqs. equation 5–equation 7. The evidence transform and token allocation are first learned with admission fixed open, then frozen. On a disjoint subset of training episodes, we compare the native prediction $\widehat{Y}_t^0$ with $\widehat{Y}_t^{+i}$, obtained by admitting the frozen candidate from one view. Other observations, task inputs, and model states remain unchanged. With the branch index omitted for the calibrated prediction, the utility is

$$u_{t,i} = \mathrm{sg}\Big[\mathcal{L}_{\mathrm{act}}(\widehat{Y}_t^0, Y_t^\star) - \mathcal{L}_{\mathrm{act}}(\widehat{Y}_t^{+i}, Y_t^\star)\Big], \tag{4}$$

where sg stops gradients, $Y_t^\star$ denotes the expert target, and $\mathcal{L}_{\mathrm{act}}$ combines position, heading, and arrival errors. Utilities above $\delta > 0$ label beneficial evidence, values below $-\delta$ label harmful evidence, and the neutral interval is excluded. Binary cross entropy trains the corresponding admission gate. The candidate comparison is independent of the learned gate, so its current admission decision cannot determine its own utility label. These comparisons run only offline and measure isolated view interventions, not additive contributions or guaranteed episode success. The full action loss and calibration procedure appear in Appendix A.3.

**Bounded Residual Evidence.** Utility calibration supervises evidence admission, but gate values alone do not bound the resulting representation change. Small gates can still produce large updates when projected features have large magnitudes. To limit changes to native representations during auxiliary fusion (Jiang et al., 2026; Huang et al., 2026), we constrain the actual residual,

$$\Delta_{t,i}^b = \gamma_b g_{t,i}^b \big(s_{t,i}^b \odot (C_{t,i}^b W_O^b)\big), \tag{5}$$

$$E_{t,i}^b = \Delta_{t,i}^b \min\left(1, \frac{\rho_b \|V_{t,i}\|_F}{\|\Delta_{t,i}^b\|_F + \epsilon}\right), \tag{6}$$

$$\widetilde{V}_{t,i}^b = V'_{t,i} + b_{t,i} E_{t,i}^b, \qquad V'_{t,i} \equiv V_{t,i}. \tag{7}$$

Here, $\gamma_b$ is a learned scale, $W_O^b$ is the output projection, $\rho_b$ bounds relative perturbation, $\epsilon > 0$ stabilizes normalization, and $b_{t,i} \in \{0, 1\}$ indicates valid evidence. Token weights are broadcast over channels. Initializing $\gamma_b$ to zero recovers the native representation at initialization. Invalid evidence bypasses fusion. The same bounded fusion is used to construct candidate predictions during calibration and to inject routed evidence during inference. The constraint bounds feature changes, not action safety. The intervention ratio and validity checks are given in Appendix A.2.

### 4.3 Optimization and Execution

**Preservation and Task Learning.** We train on clean and degraded episodes with task learning and representation preservation (Huang et al., 2026; Wei et al., 2026),

$$\mathcal{L} = \mathcal{L}_{\mathrm{nav}}^{(h)} + \lambda_{\mathrm{ret}} \mathcal{L}_{\mathrm{ret}} + \lambda_{\mathrm{int}} \mathcal{L}_{\mathrm{int}} + \lambda_u \mathcal{L}_u. \tag{8}$$

The direct head uses the action loss, and the flow head retains its flow matching objective. Retention penalizes changes on clean states where native predictions meet expert tolerances. The intervention loss penalizes residual magnitude. Candidate learning omits $\mathcal{L}_u$. Calibration fixes the candidate and optimizes admission with the full objective. Appendix A.4 details losses and parameter updates.

**Cached Execution.** Following navigation systems that reuse visual history (Xue et al., 2026; Wei et al., 2026), we cache frozen reconstruction features for their original frames and views. Routing is recomputed using the current context. The slow system selects a subgoal with multiscale evidence, and the fast system uses intermediate features for waypoint prediction. Missing, incompatible, or unfinished evidence follows the native path. Caching avoids repeated encoding, not the cost of processing new images. The inherited planning logic and locomotion controller remain unchanged. Cache keys and scheduling details are provided in Appendix A.5.

Table 1: Quantitative comparison on FireVLN-Bench unseen test split. Among methods with degraded observations, best and second-best results are highlighted in bold and underline, respectively.

| Method | Venue | Physical Environment | | | | Continuous Environment | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | NE↓ | OS↑ | SR↑ | SPL↑ | NE↓ | OS↑ | SR↑ | SPL↑ |
| Native VLN (clean) | - | 4.58 | 68.2 | 54.1 | 42.3 | 3.32 | 79.7 | 67.5 | 56.6 |
| Native VLN (degraded) | - | 7.34 | 43.0 | 26.8 | 17.3 | 5.94 | 56.3 | 41.8 | 29.9 |
| *Reconstructed RGB Frontends* | | | | | | | | | |
| AdaIR (Cui et al., 2025) | ICLR'25 | 7.10 | 45.1 | 29.4 | 19.7 | 5.82 | 57.2 | 42.7 | 30.5 |
| DCPT (Hu et al., 2025) | ICLR'25 | 6.68 | 48.8 | 34.1 | 25.1 | 5.33 | 62.0 | 47.5 | 36.6 |
| MoCE-IR-S (Zamfir et al., 2025) | CVPR'25 | <u>6.08</u> | 50.9 | 34.8 | <u>27.8</u> | 5.21 | 62.9 | <u>51.2</u> | 38.3 |
| DFPIR (Tian et al., 2025) | CVPR'25 | 6.65 | 49.5 | 32.9 | 23.7 | 5.46 | 60.9 | 46.2 | 34.9 |
| R2R-Compos (Wang et al., 2026a) | CVPR'26 | 6.52 | 50.5 | 34.5 | 24.8 | 5.32 | 61.9 | 48.2 | 36.6 |
| DATPRL-IR (Dong et al., 2026) | ICLR'26 | 6.68 | 49.0 | 33.5 | 23.7 | 5.44 | 60.8 | 47.2 | 35.6 |
| HOGformer (Wu et al., 2026) | AAAI'26 | 6.84 | 47.7 | 31.5 | 22.3 | 5.57 | 60.0 | 44.5 | 33.1 |
| BiLaLoRA (Zhang et al., 2026) | CVPR'26 | 6.34 | <u>54.2</u> | 35.6 | 25.2 | <u>4.86</u> | 63.1 | 48.8 | <u>40.1</u> |
| *Task Adaptation and Reconstruction Evidence* | | | | | | | | | |
| Raw RGB + Adapter | - | 6.35 | 52.4 | 35.3 | 25.2 | 5.04 | 64.6 | 49.8 | 38.6 |
| Restored RGB + Adapter | - | 6.16 | 53.9 | <u>37.3</u> | 27.1 | 4.99 | <u>66.2</u> | 50.8 | 39.5 |
| **FireNav** | **Ours** | **5.65** | **59.0** | **41.6** | **32.0** | **4.53** | **68.8** | **56.1** | **45.5** |

# 5 Experiments

## 5.1 Experimental Settings

**Benchmarks and Tasks.** Experiments use FireVLN-Bench's matched CE and PE tracks in Isaac Sim (Section 3). Instruction following on the unseen test split is the main task. Additional scene splits and goal tasks appear in Appendix E. Methods share task specifications, initial states, hazard layouts, degradation schedules, and episode manifests, but execute independent trajectories.

**Metrics.** We report Navigation Error (NE), Oracle Success (OS), Success Rate (SR), and Success weighted by Path Length (SPL) (Xue et al., 2026; Wei et al., 2026; Wang et al., 2025). NE is in meters, and OS, SR, and SPL are percentages. Action diagnostics report average and final displacement errors (ADE/FDE), heading error, and arrival F1. Clean observations are evaluated to assess performance retention. Physical failures and stopping errors appear in Appendix D.

**Baselines.** The RGB frontends are AdaIR (Cui et al., 2025), DCPT (Hu et al., 2025), MoCE-IR-S (Zamfir et al., 2025), DFPIR (Tian et al., 2025), R2R-Compos (Wang et al., 2026a), DATPRL-IR (Dong et al., 2026), HOGformer (Wu et al., 2026), and BiLaLoRA (Zhang et al., 2026). Each supplies reconstructed images to the same navigator. Raw RGB + Adapter and Restored RGB + Adapter add navigation supervision to the corresponding inputs. The latter uses the same reconstruction model as FireNav. FireNav uses the R2R-Compos encoder and learns only the evidence interface and router. Clean and degraded native inputs provide reference conditions.

**Implementation.** Comparisons share the navigation checkpoint, visual history, waypoint interface, controller, and STOP rule. Reconstruction weights remain frozen. The reader uses width 384 and six attention heads. The main comparisons use the direct waypoint head. Runtime covers auxiliary encoding and fusion and reports P50/P95 latency, update frequency, and peak memory.

Closed-loop methods share initial conditions and task definitions but execute independent trajectories. Paired clean/degraded observations are used only for same-state comparisons and are excluded from online policy inputs. This avoids reference leakage during closed-loop navigation.

## 5.2 Main Results

Table 1 lists 41.6 SR and 32.0 SPL for FireNav in PE, and 56.1 SR and 45.5 SPL in CE. Relative to the degraded native baseline, SR/SPL are higher by 14.8/14.7 points in PE and 14.3/15.6 points in

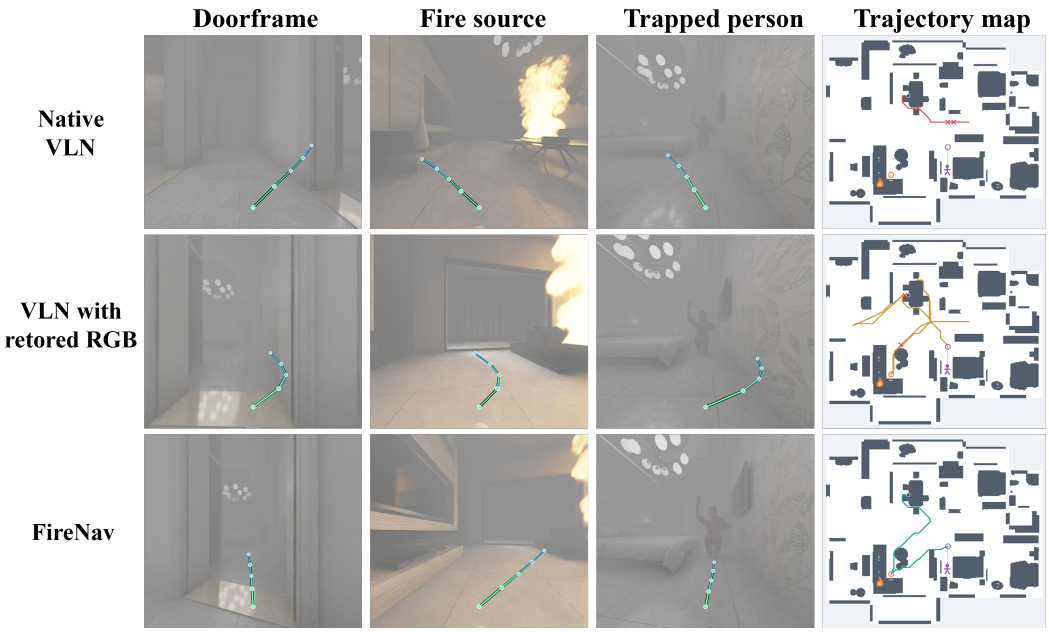



Figure 3: Qualitative comparison of waypoint predictions and trajectories. Each row shows a model configuration. The first three columns show local waypoint predictions at keyframes, and the last shows trajectories on occupancy maps. Red crosses mark task interruptions (`STOP`).

CE. FireNav has the best values in all eight columns among methods using degraded observations. Its SPL exceeds the strongest competing value by 4.2 points in PE and 5.4 points in CE. The clean reference is excluded from this ranking.

Relative to Restored RGB + Adapter, SR/SPL differences are 4.3/4.9 points in PE and 5.3/6.0 points in CE. NE decreases by 0.51 and 0.46 m. Both configurations share the reconstruction model but integrate its output differently, comparing restored RGB with routed reconstruction features. This isolates the benefit of feature-level evidence routing beyond reconstruction and task adaptation.

Across methods, PE remains harder than CE, where physical execution introduces contact dynamics, actuation error, and deadlocks. FireNav maintains its advantage under challenging fire conditions in both tracks, showing consistent gains beyond continuous navigation. This suggests that routed reconstruction evidence remains effective under additional physical execution disturbances.

Figure 3 complements the aggregate results with local predictions and trajectory views. The selected keyframes cover a doorway, a fire source, and a person under smoke occlusion. Waypoint overlays expose differences in local direction, while the maps show route choices and interruption locations. The keyframes depict different camera positions along the respective trajectories rather than a controlled comparison at an identical navigation state. The visualization further highlights how local prediction differences accumulate into distinct navigation outcomes.

### 5.3 Analysis of Decision-Guided Evidence Routing

**Component Analysis.** In Table 2, unconditional fusion has 6.0/6.9 points lower PE/CE SPL than the full configuration. Removing decision context lowers SPL by 3.0/3.4 points, and removing utility calibration lowers it by 4.0/4.9 points. Without residual control, SPL decreases by 2.3/3.6 points. Image gating alone loses 1.2/1.1 points. Removing clean retention raises degraded SPL by 1.0/0.8 points but lowers clean SR by 4.2/3.4 points. The full configuration balances degraded performance and clean retention rather than maximizing every column.

**Evidence Content.** Appendix C.2 distinguishes trained variants from input interventions on one checkpoint. Shuffled content has 13.0/15.1 points lower PE/CE SPL than matched evidence. Shuffling only the router context reduces SPL by 3.3/4.5 points. These comparisons examine dependence

Table 2: Ablation study of FireNav. Clean SR measures retention of the native behavior under clean observations. All variants use the same navigation backbone and evaluation protocol.

| Variant | Physical Environment | | | Continuous Environment | | |
|---|---|---|---|---|---|---|
| | SR↑ | SPL↑ | Clean SR↑ | SR↑ | SPL↑ | Clean SR↑ |
| **Full FireNav** | **41.6** | **32.0** | 52.8 | **56.1** | **45.5** | 67.1 |
| Always-on evidence | 36.1 | 26.0 | 50.4 | 50.5 | 38.6 | 65.6 |
| w/o decision context | 38.9 | 29.0 | 53.2 | 53.7 | 42.1 | 67.1 |
| w/o quality statistics | 39.8 | 30.0 | 53.6 | 54.4 | 43.1 | 66.8 |
| w/o utility calibration | 37.8 | 28.0 | 52.9 | 52.2 | 40.6 | 66.7 |
| w/o residual control | 39.9 | 29.7 | 51.6 | 53.2 | 41.9 | 66.2 |
| w/o clean retention | 42.8 | 33.0 | 48.6 | 57.2 | 46.3 | 63.7 |
| Image gate only | 40.8 | 30.8 | 53.8 | 55.4 | 44.4 | 67.8 |
| Random encoder + trained interface | 37.8 | 27.5 | 52.8 | 51.6 | 39.7 | 67.0 |

Table 3: Runtime comparison under the same execution setting. P50/P95 denote update latency.

| Method | Additional Cost | | End-to-End Latency | | System | |
|---|---|---|---|---|---|---|
| | Encoder/Restoration (ms)↓ | Fusion (ms)↓ | P50 (ms)↓ | P95 (ms)↓ | Hz↑ | Peak GB↓ |
| Native VLN | 0.0 | 0.0 | **280.1** | **359.2** | **3.53** | **18.06** |
| AdaIR | 209.9 | 2.0 | 491.9 | 597.4 | 2.02 | 20.26 |
| DCPT | 96.5 | 2.0 | 378.8 | 464.8 | 2.62 | 19.16 |
| MoCE-IR-S | 126.1 | 2.0 | 408.5 | 498.1 | 2.43 | 19.36 |
| DFPIR | 113.9 | 2.0 | 396.9 | 485.0 | 2.51 | 19.66 |
| R2R-Compos | 78.3 | 2.0 | 361.2 | 445.0 | 2.75 | 19.06 |
| DATPRL-IR | 134.8 | 2.0 | 417.3 | 508.5 | 2.38 | 19.86 |
| HOGformer | 270.4 | 2.0 | 551.3 | 668.4 | 1.80 | 20.56 |
| BiLaLoRA | 71.8 | 2.0 | 354.0 | 436.5 | 2.80 | 19.56 |
| Raw RGB + Adapter | 0.0 | 8.0 | 288.0 | 367.3 | 3.43 | 18.46 |
| Restored RGB + Adapter | 77.9 | 10.0 | 368.7 | 452.4 | 2.69 | 19.46 |
| **FireNav** | 17.9 | 12.0 | 310.1 | 389.4 | 3.19 | 19.16 |

on evidence content and decision context. The random encoder and adapter configurations provide separate controls for reconstruction pretraining and task adaptation.

### 5.4 Action Quality and Efficiency

**Action Prediction.** Table 7 in Appendix C.3 lists degraded ADE/FDE of 0.504/0.861 m for FireNav, compared with 0.555/0.926 m for Restored RGB + Adapter. Heading error is $15.2^\circ$ versus $17.0^\circ$, and arrival F1 is 0.667 versus 0.586. Clean ADE remains close to the native value, at 0.506 versus 0.504 m. These local diagnostics are distinct from episode success and STOP outcomes .

**Runtime.** Table 3 lists 17.9 ms for reconstruction feature encoding and 12.0 ms for fusion. FireNav's P95 latency is 389.4 ms, 8.4% above the native baseline and 63.0 ms below Restored RGB + Adapter. The update frequencies are 3.19, 3.53, and 2.69 Hz. Direct use of encoder features avoids RGB decoding on the fast path. Appendix F separates history length from feature reuse. These latency figures describe navigation updates rather than the frequency of the locomotion controller.

## 6 Conclusion

We have presented FireNav, a decision-oriented reconstruction evidence routing framework for robust VLN in indoor fire environments. To advance benchmarking research in this field, FireVLN-Bench is first introduced as a paired benchmark with clean references, safety supervision, and matched settings for continuous navigation and robot execution. Our method preserves native visual tokens while selectively incorporating reconstruction evidence through lightweight routing and decision utility calibration. Extensive experiments demonstrate consistent improvements over state-of-the-art baselines in navigation success and path efficiency across evaluation settings.

### AI USE STATEMENT

Generative AI tools were used to assist manuscript writing and editing. All final content was reviewed and approved by the authors.

### ETHICS STATEMENT

FireNav and FireVLN-Bench use simulated fire environments and contain no human-subject or identifiable data. Results do not establish rescue safety and require validation before deployment.

### REPRODUCIBILITY STATEMENT

Code, configurations, benchmark scripts, data splits, and tools will be released. The paper and appendix provide the model, training, and evaluation details required for reproduction.

## REFERENCES

I-Hsiang Chen, Wei-Ting Chen, Yu-Wei Liu, Yuan-Chun Chiang, Sy-Yen Kuo, and Ming-Hsuan Yang. Unirestore: Unified perceptual and task-oriented image restoration model using diffusion prior. In *CVPR*, 2025.

Yuning Cui, Syed Waqas Zamir, Salman Khan, Alois Knoll, Mubarak Shah, and Fahad Shahbaz Khan. AdaIR: Adaptive all-in-one image restoration via frequency mining and modulation. In *ICLR*, 2025.

Guanglu Dong, Chunlei Li, Chao Ren, Jingliang Hu, Yilei Shi, Xiao Xiang Zhu, and Lichao Mou. Learning domain-aware task prompt representations for multi-domain all-in-one image restoration. In *ICLR*, 2026.

Junkai Fan, Kun Wang, Zhiqiang Yan, Xiang Chen, Shangbing Gao, Jun Li, and Jian Yang. Depth-centric dehazing and depth-estimation from real-world hazy driving video. *AAAI*, 39(3):2852–2860, 2025. doi: 10.1609/aaai.v39i3.32291.

Jiayi Fu, Siyu Liu, Zikun Liu, Chun-Le Guo, Hyunhee Park, Ruiqi Wu, Guoqing Wang, and Chongyi Li. Iterative predictor-critic code decoding for real-world image dehazing. In *CVPR*, pp. 12700–12709, 2025.

Ruishan Guo, Ciyu Ruan, Haoyang Wang, Zihang Gong, Jingao Xu, and Xinlei Chen. $x^2$-Fusion: Cross-modality and cross-dimension flow estimation in event edge space. In *CVPR*, 2026a.

Wenxuan Guo, Xiuwei Xu, Yichen Liu, Xiangyu Li, Hang Yin, Huangxing Chen, Wenzhao Zheng, Jianjiang Feng, Jie Zhou, and Jiwen Lu. AwareVLN: Reasoning with self-awareness for vision-language navigation. In *CVPR*, pp. 4065–4075, 2026b.

Honglin He, Yukai Ma, Brad Squicciarini, Wayne Wu, and Bolei Zhou. From seeing to experiencing: Scaling navigation foundation models with reinforcement learning. In *ICLR*, 2026.

Jiakui Hu, Lujia Jin, Zhengjian Yao, and Yanye Lu. Universal image restoration pre-training via degradation classification. In *ICLR*, 2025.

Chengyue Huang, Mellon M. Zhang, Robert Azarcon, Glen Chou, and Zsolt Kira. MAPS: Preserving vision-language representations via module-wise proximity scheduling for better vision-language-action generalization. In *CVPR*, 2026.

Yufeng Huang, Xiang Chen, Lei Xu, and Kaiyuan Li. Single image desmoking via attentive generative adversarial network for smoke detection process. *Fire Technology*, 57(6):3021–3040, 2021.

Wen Jiang, Kangyao Huang, Li Wang, Wang Xu, Wei Fan, Jinyuan Liu, Shaoyu Liu, Hanfang Liang, Hongwei Duan, Bin Xu, and Xiangyang Ji. Spatialfly: Geometry-guided representation alignment for uav vision-and-language navigation in urban environments. *arXiv preprint arXiv:2603.21046*, 2026.

Wenzhuo Jin, Qianfeng Yang, Xianhao Wu, Hongming Chen, Pengpeng Li, and Xiang Chen. Smokebench: A real-world dataset for surveillance image desmoking in early-stage fire scenes. In *ACM MM*, pp. 12722–12728, 2025.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *ECCV*, pp. 104–120, 2020.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *EMNLP*, pp. 4392–4412, 2020.

Hongjae Lee, Sojung Kang, Jaeseong Yu, and Seung-Won Jung. Tasktok: Delving into task tokens for task-driven image restoration. In *ECCV*, 2026.

Chunyi Li, Jiahao Xiao, Jianbo Zhang, Farong Wen, Zicheng Zhang, Yuan Tian, Xiangyang Zhu, Xiaohong Liu, Zhengxue Cheng, Weisi Lin, and Guangtao Zhai. Image quality assessment for embodied AI. In *ICLR*, 2026a.

LinFeng Li, Jian Zhao, Yuan Xie, Xin Tan, and Xuelong Li. CompassNav: Steering from path imitation to decision understanding in navigation. In *ICLR*, 2026b.

Yuanchang Liang, Xiaobo Wang, Kai Wang, Shuo Wang, Xiaojiang Peng, Haoyu Chen, David Kim Huat Chua, and Prahlad Vadakkepat. Adaptive action chunking at inference-time for vision-language-action models. In *CVPR*, pp. 20802–20811, 2026.

Jingbo Lin, Zhilu Zhang, Wenbo Li, Renjing Pei, Hang Xu, Hongzhi Zhang, and Wangmeng Zuo. UniRestorer: Universal image restoration via adaptively estimating image degradation at proper granularity. In *ICLR*, 2026.

Bingchen Miao, Rong Wei, Zhiqi Ge, Xiaoquan Sun, Shiqi Gao, Jingzhe Zhu, Renhan Wang, Siliang Tang, Jun Xiao, Rui Tang, and Juncheng Li. Towards physically executable 3d gaussian for embodied navigation. In *ICLR*, 2026.

SpatialVerse Research Team, Manycore Tech Inc. Interioragent: Interactive usd interior scenes for isaac sim-based simulation, 2025. Hugging Face dataset.

Xiangpeng Tian, Xiangyu Liao, Xiao Liu, Meng Li, and Chao Ren. Degradation-aware feature perturbation for all-in-one image restoration. In *CVPR*, pp. 28165–28175, 2025.

Chenxu Wang, Kai Zhang, and Jian Yang. Retrieve-to-Restore: Efficient all-in-one image restoration with a retrieval-based degradation bank. In *CVPR*, pp. 1277–1287, 2026a.

Liuyi Wang, Xinyuan Xia, Hui Zhao, Hanqing Wang, Tai Wang, Yilun Chen, Chengju Liu, Qijun Chen, and Jiangmiao Pang. Rethinking the embodied gap in vision-and-language navigation: A holistic study of physical and visual disparities. In *ICCV*, pp. 9455–9465, 2025.

Shuo Wang, Yucheng Wang, Guoxin Lian, Yongcai Wang, Maiyue Chen, Kaihui Wang, Bo Zhang, Zhizhong Su, Yutian Zhou, Wanting Li, Deying Li, and Zhaoxin Fan. Progress-Think: Semantic progress reasoning for vision-language navigation. In *CVPR*, pp. 4076–4086, 2026b.

Xudong Wang, Gan Li, Zhiyu Liu, Yao Wang, Lianqing Liu, and Zhi Han. All-day multi-scenes lifelong vision-and-language navigation with tucker adaptation. In *ICLR*, 2026c.

Yixiao Wang, Mingxiao Huo, Zhixuan Liang, Yushi Du, Lingfeng Sun, Haotian Lin, Jinghuan Shang, Chensheng Peng, Mohit Bansal, Mingyu Ding, and Masayoshi Tomizuka. VER: Vision expert transformer for robot learning via foundation distillation and dynamic routing. In *ICLR*, 2026d.

Meng Wei, Chenyang Wan, Jiaqi Peng, Xiqian Yu, Yuqiang Yang, Delin Feng, Wenzhe Cai, Chenming Zhu, Tai Wang, Jiangmiao Pang, and Xihui Liu. Ground slow, move fast: A dual-system foundation model for generalizable vision-and-language navigation. In *ICLR*, 2026.

Jiawei Wu, Zhifei Yang, Zhe Wang, and Zhi Jin. Gradient as conditions: Rethinking HOG for all-in-one image restoration. In *AAAI*, volume 40, pp. 10682–10690, 2026. doi: 10.1609/aaai.v40i13.38042.

Xinda Xue, Junjun Hu, Minghua Luo, Shichao Xie, Jintao Chen, Zixun Xie, Quan Kuichen, Wei Guo, Mu Xu, Zedong Chu, and Zhengzhou Zhu. Omninav: A unified framework for prospective exploration and visual-language navigation. In *ICLR*, 2026.

Eduard Zamfir, Zongwei Wu, Nancy Mehta, Yuedong Tan, Danda Pani Paudel, Yulun Zhang, and Radu Timofte. Complexity experts are task-discriminative learners for any image restoration. In *CVPR*, pp. 12753–12763, 2025.

Shuang Zeng, Dekang Qi, Xinyuan Chang, Feng Xiong, Shichao Xie, Xiaolong Wu, Shiyi Liang, Mu Xu, and Xing Wei. JanusVLN: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. In *ICLR*, 2026.

Yan Zhang, Long Ma, Yuxin Feng, Zhe Huang, Fan Zhou, and Zhuo Su. Bilevel layer-positioning LoRA for real image dehazing. In *CVPR*, pp. 29969–29978, 2026.

# A Additional Method Details

## A.1 Evidence Sources and Token Construction

The reconstruction encoder provides an intermediate feature $R^{\mathrm{mid}}_{t,i}$ and multiscale features $\{R^l_{t,i}\}_{l\in\mathcal{L}}$. The evidence source for branch $b$ is

$$\mathcal{R}^b_{t,i} = \begin{cases} R^{\mathrm{mid}}_{t,i}, & b = \mathrm{fast}, \\ \{R^l_{t,i}\}_{l\in\mathcal{L}}, & b = \mathrm{slow}. \end{cases} \tag{9}$$

Scale projections and spatial pooling construct the evidence tokens,

$$Z^b_{t,i} = A^b_\phi\left(\mathcal{R}^b_{t,i}, q_{t,i}, \nu_i\right) \in \mathbb{R}^{M^b_i \times d_a}. \tag{10}$$

Here $q_{t,i}$ contains brightness, contrast, saturation, and gradient statistics, and $\nu_i$ identifies the view. Cross attention returns $N_i$ tokens. The output projection restores the native width $d_v$ before residual addition. Encoder evidence does not require RGB decoding on the fast path.

## A.2 Residual Validity and Initialization

The validity flag $b_{t,i}$ differs from the branch index $b$. Frame identity, preprocessing, checkpoint consistency, and finite tensor values are checked before fusion. Invalid evidence returns the native tokens directly, rather than multiplying invalid values by zero. The notation $V'_{t,i} \equiv V_{t,i}$ in Figure 2 denotes the identity stream. Evidence is added once. Initializing $\gamma_b = 0$ gives the native representation. For $\rho_b \geq 0$ and $\epsilon > 0$,

$$r^b_{t,i} = \frac{\|b_{t,i} E^b_{t,i}\|_F}{\|V_{t,i}\|_F + \epsilon} \leq \rho_b. \tag{11}$$

The bound follows from Eq. equation 6 and applies to the whole image representation. It is not a bound on each token or on action error.

## A.3 Action Loss and Utility Calibration

For predicted positions $\hat{p}_j$, headings $\hat{\theta}_j$, and arrival probabilities $\hat{a}_j$, with expert targets marked by $\star$, the action loss is

$$\begin{aligned} \mathcal{L}_{\mathrm{act}}(\widehat{Y}, Y^\star) = \frac{1}{H}\sum_{j=1}^{H} \Big[ & \lambda_p \,\mathrm{SmoothL1}(\hat{p}_j - p^\star_j) \\ & + \lambda_\theta\big(1 - \cos(\hat{\theta}_j - \theta^\star_j)\big) \\ & + \lambda_a \,\mathrm{BCE}(\hat{a}_j, a^\star_j)\Big]. \end{aligned} \tag{12}$$

Coordinates follow the native decoder and normalization convention. Calibration compares the native prediction with a frozen candidate admitted from one view. All other observations, task

inputs, and model states are fixed. The candidate admission is open when constructing the utility label. It does not depend on the learned gate. Values within $[-\delta, \delta]$ are excluded,

$$\mathcal{L}_u = \mathbb{E}_{|u_{t,i}|>\delta} \operatorname{BCE}(g_{t,i}, \mathbf{1}[u_{t,i} > \delta]) . \tag{13}$$

The branch index is suppressed for the calibrated prediction. Calibration episodes belong to the training split and are excluded from candidate learning. The labels describe isolated view interventions. Predictions using multiple views are also supervised by the navigation objective.

### A.4 Loss Terms and Training Stages

The direct head uses Eq. equation 12. The Method also describes retaining the native objective for a flow head (Xue et al., 2026; Wei et al., 2026), which requires its own checkpoint and decoding convention. No flow-head results are included in this revision. Clean retention compares assisted outputs with detached native predictions only where the latter meet expert tolerances. It uses position, circular heading, and arrival probability differences. The intervention term is

$$\mathcal{L}_{\text{int}} = \mathbb{E}_t \left[ \frac{1}{|\mathcal{B}_t| \, |\mathcal{I}_t|} \sum_{b \in \mathcal{B}_t} \sum_i (r^b_{t,i})^2 \right] , \tag{14}$$

where $\mathcal{B}_t$ contains the active branches. Candidate learning fixes admission open and trains the evidence transform and token allocation without $\mathcal{L}_u$. Calibration freezes the candidate and trains admission using Eq. equation 8. Gradients pass through the frozen navigator without updating its weights.

### A.5 Cache Consistency and Scheduling

Cache keys contain frame identity, view, preprocessing transform, checkpoint, and precision. Only frozen reconstruction features are reused. Gates are recomputed from the current context, and historical features remain attached to their original observations. Missing, incompatible, or unfinished evidence uses the native path. A pending feature may later serve its matching historical frame. Reuse does not avoid encoding new observations or change the inherited planner and locomotion controller.

## B Additional Experimental Details

### B.1 Benchmark Allocation

Table 4: Data allocation in FireVLN-Bench. Validation seen uses held-out episodes from the training scenes and therefore does not introduce additional scene identities.

| Split | Scenes | Layouts | Routes/Layout/Task | Episodes/Task | Four-Task Episodes |
|---|---|---|---|---|---|
| Train | 40 | 160 | 100 | 16,000 | 64,000 |
| Validation Seen | 40 | 160 | 25 | 4,000 | 16,000 |
| Validation Unseen | 8 | 32 | 25 | 800 | 3,200 |
| Test Unseen | 12 | 48 | 25 | 1,200 | 4,800 |

Table 4 specifies 40 training, 8 unseen validation, and 12 unseen test scenes. Seen validation reuses training scenes but reserves separate episodes. The listed instruction following allocation is 1,200 test episodes per track. CE and PE are separate executions of canonical tasks, not additional scene identities.

### B.2 Evaluation Protocol

Closed loop comparisons share initial conditions and task definitions, not all subsequent observations. Each policy generates its own trajectory. Paired clean and degraded frames instead share a camera pose and scene state for controlled image or prediction comparisons. Action diagnostics and arrival classification are separate from episode metrics and accepted STOP events. Clean SR measures task performance, not equality of individual predictions. Collision, fall, and stuck events need not be mutually exclusive.

# C Extended Routing and Action Analysis

## C.1 Encoder and Routing Strategy

Table 5: Factorial comparison of reconstruction encoders and evidence routing strategies. Values are SPL (%). Each encoder uses the same evidence budget and optimization recipe.

| Encoder | Physical Environment | | | Continuous Environment | | |
|---|---|---|---|---|---|---|
| | Always-on | Quality Gate | FireNav | Always-on | Quality Gate | FireNav |
| AdaIR | 23.6 | 26.0 | **27.8** | 35.6 | 38.0 | **40.3** |
| DCPT | 27.0 | 28.6 | **32.5** | 39.2 | 41.2 | **46.0** |
| MoCE-IR-S | 26.4 | 29.4 | **32.3** | 39.1 | 42.1 | **45.7** |
| DFPIR | 25.4 | 28.6 | **30.8** | 37.9 | 41.2 | **44.2** |
| R2R-Compos | 26.0 | 29.2 | **32.0** | 38.6 | 41.9 | **45.5** |
| DATPRL-IR | 26.0 | 28.9 | **31.7** | 38.3 | 41.9 | **45.4** |
| HOGformer | 25.4 | 27.8 | **30.2** | 37.0 | 40.3 | **43.9** |
| BiLaLoRA | 26.7 | 28.5 | **31.3** | 38.2 | 41.4 | **44.9** |

Within each encoder in Table 5, FireNav routing has higher SPL than unconditional fusion or quality gating. R2R-Compos gives 32.0/45.5 PE/CE SPL, compared with 32.5/46.0 for DCPT. Encoder choice and routing strategy are distinct factors. This table alone does not explain why R2R-Compos is selected.

## C.2 Evidence Content Interventions

Table 6: Evidence intervention analysis with the same trained FireNav checkpoint. Evidence content or router context is modified without retraining the navigator.

| Intervention | Physical Environment | | | Continuous Environment | | |
|---|---|---|---|---|---|---|
| | SR↑ | SPL↑ | Clean SR↑ | SR↑ | SPL↑ | Clean SR↑ |
| **Matched evidence** | **41.6** | **32.0** | 52.8 | **56.1** | **45.5** | 67.1 |
| Exact bypass | 26.8 | 17.3 | 54.1 | 41.8 | 29.9 | 67.5 |
| Zero spatial features | 28.0 | 18.0 | 54.0 | 41.2 | 29.5 | 68.1 |
| Shuffled content | 29.3 | 19.0 | 53.6 | 42.4 | 30.4 | 66.9 |
| Incorrect view content | 27.3 | 17.4 | 53.4 | 40.5 | 29.2 | 67.3 |
| Incorrect time content | 26.9 | 17.1 | 53.3 | 40.0 | 28.6 | 67.2 |
| Shuffled router context only | 38.2 | 28.7 | 53.8 | 52.2 | 41.0 | 67.4 |
| Invalid metadata / bypass | 26.8 | 17.3 | 54.1 | 41.8 | 29.9 | 67.5 |

The exact and invalid-metadata bypass rows in Table 6 match the native reference. Shuffled content has 19.0/30.4 PE/CE SPL, compared with 32.0/45.5 for matched evidence. Incorrect views or times also reduce the listed scores. These are interventions on one checkpoint, not independently trained ablations. The injection point must distinguish a valid packet containing altered content from an invalid packet rejected by the bypass checks.

## C.3 Local Action Diagnostics

Table 7 separates local prediction errors from closed loop results. Relative to Restored RGB + Adapter, FireNav has $1.8^\circ$ lower degraded heading error and 0.081 higher arrival F1. Clean ADE is 0.506 m, compared with 0.504 m for the native model. An arrival classification score is not a rate of successful episode termination.

Table 7: Action prediction under degraded and clean observations. ADE/FDE are measured in meters, heading error in degrees, and arrival F1 as a fraction.

| Method | Degraded Observations | | | | Clean Observations | | |
|---|---|---|---|---|---|---|---|
| | ADE↓ | FDE↓ | Heading↓ | Arrival F1↑ | ADE↓ | Heading↓ | Arrival F1↑ |
| Native VLN | 0.666 | 1.101 | 21.0 | 0.474 | **0.504** | **17.0** | **0.670** |
| MoCE-IR-S | 0.595 | 0.991 | 18.8 | 0.546 | 0.516 | 17.5 | 0.657 |
| R2R-Compos | 0.615 | 1.021 | 19.5 | 0.523 | 0.522 | 17.8 | 0.644 |
| HOGformer | 0.615 | 1.036 | 19.8 | 0.509 | 0.511 | 17.3 | 0.663 |
| BiLaLoRA | 0.600 | 1.001 | 18.6 | 0.552 | 0.514 | 17.4 | 0.663 |
| Raw RGB + Adapter | 0.565 | 0.956 | 17.4 | 0.575 | 0.507 | 17.2 | 0.667 |
| Restored RGB + Adapter | 0.555 | 0.926 | 17.0 | 0.586 | 0.509 | 17.4 | 0.663 |
| **FireNav** | **0.504** | **0.861** | **15.2** | **0.667** | 0.506 | 17.1 | 0.669 |

Table 8: Relationship between image reconstruction quality, action prediction, and RGB reconstruction cost.

| Method | Image Quality | | Action ADE | | Cost |
|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | Degraded↓ | Clean↓ | RGB ms↓ |
| Raw degraded | 16.8 | 0.680 | 0.666 | **0.504** | 0.0 |
| AdaIR | 15.9 | 0.648 | 0.645 | 0.533 | 69.9 |
| DCPT | 17.8 | 0.717 | 0.603 | 0.518 | 31.8 |
| MoCE-IR-S | 18.2 | 0.728 | **0.595** | 0.516 | 42.0 |
| DFPIR | 17.6 | 0.711 | 0.608 | 0.520 | 38.1 |
| R2R-Compos | 16.2 | 0.658 | 0.615 | 0.522 | 26.1 |
| DATPRL-IR | 17.5 | 0.707 | 0.610 | 0.519 | 44.9 |
| HOGformer | 16.6 | 0.697 | 0.615 | 0.511 | 90.4 |
| BiLaLoRA | **18.6** | **0.745** | 0.600 | 0.514 | 24.0 |

### C.4 Image Quality and Action Utility

BiLaLoRA has the highest PSNR and SSIM among the RGB frontends in Table 8, while MoCE-IR-S has the lowest degraded ADE. R2R-Compos has lower ADE than raw input despite lower PSNR and SSIM. The listed rankings differ. They do not establish the correlation between image quality and navigation utility across individual states.

## D Physical Failures and Stopping

Among the degraded configurations in Table 9, FireNav has the lowest listed collision, stuck, and false STOP rates. Relative to Restored RGB + Adapter, PE collision and stuck rates are lower by 3.7 and 2.4 points. Its fall rate is 2.2%, compared with 2.1% for the adapter and 1.4% for DFPIR. These categories do not support an advantage on every failure mode or a claim about cumulative smoke exposure.

## E Performance Across Splits and Tasks

### E.1 Seen and Unseen Environments

FireNav exceeds Restored RGB + Adapter on each split in Table 10. Its PE/CE SPL is 41.7/52.5 on seen validation and 32.0/45.5 on unseen testing. The comparison uses different scene sets and is not a paired change in the same environments.

Table 9: Safety and terminal behavior on FireVLN-Bench. All values are percentages of test episodes.

| Method | Physical Environment | | | | Continuous Environment | |
|---|---|---|---|---|---|---|
| | Collision↓ | Fall↓ | Stuck↓ | False STOP↓ | Collision↓ | False STOP↓ |
| Native VLN (clean) | 10.3 | 1.9 | 6.3 | 3.2 | 3.5 | 3.3 |
| Native VLN (degraded) | 33.2 | 3.2 | 16.2 | 13.5 | 19.8 | 14.5 |
| AdaIR | 31.0 | 3.2 | 15.7 | 12.7 | 18.1 | 13.2 |
| DCPT | 23.6 | 2.2 | 13.7 | 10.0 | 13.2 | 10.2 |
| MoCE-IR-S | 22.5 | 2.1 | 12.8 | 9.0 | 11.8 | 10.1 |
| DFPIR | 24.8 | **1.4** | 13.8 | 9.8 | 13.8 | 10.8 |
| R2R-Compos | 24.8 | 2.7 | 13.7 | 9.8 | 13.8 | 10.5 |
| DATPRL-IR | 25.3 | 2.8 | 13.9 | 10.6 | 14.1 | 11.3 |
| HOGformer | 27.2 | 2.8 | 14.5 | 11.3 | 15.8 | 11.8 |
| BiLaLoRA | 23.2 | 2.7 | 12.8 | 9.3 | 12.9 | 9.5 |
| Raw RGB + Adapter | 21.8 | 2.1 | 12.3 | 7.7 | 11.0 | 9.2 |
| Restored RGB + Adapter | 20.6 | 2.1 | 11.7 | 7.2 | 9.9 | 7.8 |
| **FireNav** | **16.9** | 2.2 | **9.3** | **5.0** | **7.2** | **5.7** |

Table 10: Generalization across seen validation, unseen validation, and unseen test environments.

| Method / Split | Physical Environment | | Continuous Environment | |
|---|---|---|---|---|
| | SR↑ | SPL↑ | SR↑ | SPL↑ |
| Native VLN / Seen Val. | 36.4 | 26.2 | 45.8 | 35.0 |
| Native VLN / Unseen Val. | 33.5 | 22.6 | 40.0 | 29.5 |
| Native VLN / Unseen Test | 26.8 | 17.3 | 41.8 | 29.9 |
| BiLaLoRA / Seen Val. | 44.7 | 34.0 | 53.8 | 43.0 |
| BiLaLoRA / Unseen Val. | 40.6 | 29.6 | 48.0 | 37.4 |
| BiLaLoRA / Unseen Test | 35.6 | 25.2 | 48.8 | 40.1 |
| Restored RGB + Adapter / Seen Val. | 47.0 | 36.4 | 56.5 | 45.6 |
| Restored RGB + Adapter / Unseen Val. | 42.4 | 31.4 | 50.5 | 39.8 |
| Restored RGB + Adapter / Unseen Test | 37.3 | 27.1 | 50.8 | 39.5 |
| FireNav / Seen Val. | **51.8** | **41.7** | **63.0** | **52.5** |
| FireNav / Unseen Val. | **46.9** | **36.6** | **56.6** | **46.3** |
| FireNav / Unseen Test | **41.6** | **32.0** | **56.1** | **45.5** |

### E.2 Goal Conditioned Tasks

Table 11 reports performance by task. Relative to Restored RGB + Adapter, FireNav has 5.5/5.2 points higher PE/CE SPL on object search and 4.9/5.4 points higher on point goal navigation. Without holding out a task during training, these comparisons do not establish transfer to unseen tasks.

## F Visual History and Feature Reuse

Moving from 10 to 20 history frames in Table 12 adds 0.5/0.3 PE/CE SPL points. With 20 frames, cached and recomputed features have the same listed navigation scores, while P95 differs by 131.0 ms. Restricting evidence to current views loses 2.5/1.3 SPL points. Changing history length alters the available observations. Recomputing the same features instead tests reuse, subject to identical inputs and update scheduling.

Table 11: Generalization across instruction, object, and point goal tasks.

| Method / Task | Physical Environment | | Continuous Environment | |
|---|---|---|---|---|
| | SR↑ | SPL↑ | SR↑ | SPL↑ |
| Native VLN / Instruction | 26.8 | 17.3 | 41.8 | 29.9 |
| Native VLN / Object | 20.0 | 12.0 | 28.1 | 18.4 |
| Native VLN / Point | 48.3 | 35.8 | 63.0 | 50.7 |
| BiLaLoRA / Instruction | 35.6 | 25.2 | 48.8 | 40.1 |
| BiLaLoRA / Object | 25.3 | 16.1 | 34.3 | 23.5 |
| BiLaLoRA / Point | 56.4 | 42.8 | 70.1 | 56.9 |
| Restored RGB + Adapter / Instruction | 37.3 | 27.1 | 50.8 | 39.5 |
| Restored RGB + Adapter / Object | 27.5 | 17.8 | 36.8 | 24.9 |
| Restored RGB + Adapter / Point | 59.2 | 45.5 | 71.2 | 58.2 |
| FireNav / Instruction | **41.6** | **32.0** | **56.1** | **45.5** |
| FireNav / Object | **33.1** | **23.3** | **40.8** | **30.1** |
| FireNav / Point | **63.0** | **50.4** | **74.5** | **63.6** |

Table 12: Effect of visual history and feature caching.

| Configuration | Cache | | Runtime | | PE | | CE | |
|---|---|---|---|---|---|---|---|---|
| | Calls | Hit (%) | P50 | P95 | SR | SPL | SR | SPL |
| No history, 3 current views | 3 | 0.0 | 206.1 | 256.5 | 37.9 | 28.0 | 52.5 | 41.5 |
| 4 history + 3 current views | 3 | 57.1 | 242.1 | 302.2 | 40.8 | 30.7 | 54.8 | 43.9 |
| 10 history + 3 current views | 3 | 76.9 | 277.9 | 348.9 | 41.2 | 31.5 | 56.0 | 45.2 |
| 20 history + 3 current views | 3 | 87.0 | 310.1 | 389.4 | **41.6** | **32.0** | **56.1** | **45.5** |
| 20 history, front-view evidence only | 1 | 95.2 | 292.1 | 369.5 | 39.2 | 29.2 | 54.1 | 42.8 |
| 20 history, current evidence only | 3 | 0.0 | 306.5 | 384.8 | 39.6 | 29.5 | 55.2 | 44.2 |
| 20 history, recompute all 23 images | 23 | 0.0 | 430.3 | 520.4 | **41.6** | **32.0** | **56.1** | **45.5** |