FlowDec: Temporal Conditional Flow Decorruptor for Robust Continuous Vision-Language Navigation

Yufei Zhang[1] and Changhao Chen[1⋆]

The Hong Kong University of Science and Technology (Guangzhou), China
yzhang957@connect.hkust-gz.edu.cn changhaochen@hkust-gz.edu.cn

Abstract. Vision-and-Language Navigation in Continuous Environments (VLN-CE) requires agents to follow natural-language instructions in unseen scenes. While Large Models (LMs) have advanced VLN-CE, their performance remains severely degraded by real-world visual corruptions, a critical yet underexplored domain constraint. We introduce Temporal Conditional Flow Decorruptor (FlowDec), a novel image restoration framework tailored for LM-based VLN-CE. FlowDec integrates a hybrid temporal conditioning strategy to align the generative flow path with historical context and employs action-centroid guided filtering to dynamically assess and integrate outputs. Extensive experiments demonstrate that FlowDec outperforms state-of-the-art decorruption methods in both navigation accuracy and generation latency. Our approach establishes a robust, efficient paradigm for resilient embodied navigation in unpredictable real-world conditions.

Keywords: Vision Language Navigation · Conditional Flow Matching · Image Restoration

1 Introduction

Vision-and-Language Navigation (VLN) has emerged as a cornerstone problem in Embodied AI, requiring an agent to interpret natural language instructions and execute goal-directed navigation in a 3D environment [36]. Early work focused on discrete simulators with predefined navigation graphs [3, 8, 25, 37], where action spaces and connectivity were artificially constrained. While such settings enabled rapid algorithmic progress, they abstracted away many of the complexities inherent to real-world deployment. Recent advances have shifted toward VLN in continuous environments (VLN-CE) [24], where agents must generate physically executable trajectories, reason over long temporal horizons, and ground language in raw sensory streams. This paradigm—exemplified by datasets such as R2R-CE and RxR-CE—moves VLN closer to practical robotic systems operating in homes, offices, and outdoor spaces. In this setting, perceptual fidelity and action reliability are not merely performance metrics; they are safety-critical requirements.

⋆ Corresponding author.

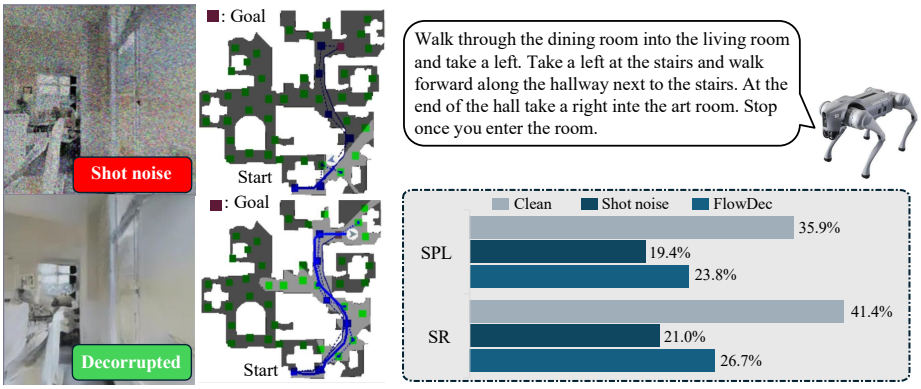

Fig. 1: LM-based VLN models suffer from corruption despite their strong generalization ability, while FlowDec can effectively alleviate the influence from corruption. The experiment is conducted on R2R-CE dataset using NaVid [48] as backbone.

The recent integration of Large Models (LMs), including Large Language Models and Vision-Language Models, has significantly advanced VLN-CE [10, 12, 30, 46–48]. By leveraging their strong reasoning and cross-modal alignment capabilities [4, 22, 54], LM-based agents such as NaVid [48] demonstrate competitive long-horizon planning and instruction grounding. These approaches suggest a promising route toward general-purpose embodied agents. However, their impressive reasoning abilities often mask a fundamental vulnerability: they implicitly assume clean and stable visual inputs. In practice, this assumption rarely holds. Real-world navigation inevitably exposes robots to diverse visual corruptions arising from both external and internal factors. Rapid ego-motion induces motion blur; low-light or over-exposure alters luminance statistics; rain, dust, or lens contamination introduce structured noise; and sensor imperfections generate stochastic artifacts. Such corruptions distort critical spatial cues—object boundaries, floor textures, depth discontinuities—that directly affect waypoint prediction and action selection. In safety-critical scenarios, even minor perceptual degradation can cascade into catastrophic failures, including collisions, disorientation, or inability to reach the goal.

Despite the growing realism of VLN-CE and the scaling of LM-based agents, robustness to visual corruptions has not been systematically studied in this setting. To our knowledge, this work is the first to explicitly investigate and address the problem of image decorruption for VLN in continuous environments. Our empirical analysis reveals the severity of this issue. As shown in Figure 1, when synthetic shot noise is applied to a representative LM-based VLN agent [48], its navigation success rate drops by nearly half, despite extensive pretraining. This degradation highlights a broader challenge: how can we endow VLN systems with robustness to unknown and diverse corruptions—without requiring prior knowledge of corruption types, without retraining the navigation backbone, and without sacrificing real-time performance?

Addressing this gap is non-trivial. A robust VLN-CE system must handle unknown and diverse corruptions without prior knowledge of their type or intensity, preserve temporally consistent observations across frames, maintain physically meaningful action predictions, and operate under strict real-time constraints. Existing image restoration approaches are not designed for this regime. Blind image denoising methods prioritize perceptual quality rather than downstream navigation consistency [7, 18, 45, 49]. Test-time adaptation (TTA) methods update model parameters during inference, introducing the risk of forgetting—especially problematic for large frozen LMs [32,33,42,43]. Diffusion-based TTA approaches mitigate parameter updates by refining inputs, yet they typically process frames independently and emphasize visual realism, overlooking temporal coherence and action-level consistency required for embodied navigation [14, 34, 41].

To address these limitations, we propose **FlowDec**, a Temporal Conditional **Flow Dec**orruptor framework designed specifically for robust embodied navigation. Unlike prior restoration methods developed for image enhancement or video generation, FlowDec is explicitly optimized for navigation-aware, real-time decorruption. Our method introduces two key modules: (1) a hybrid temporal conditioning strategy that dynamically integrates historical frames and condition types to align the generative flow path and enhance temporal consistency, and (2) Action centroids, which model differential feature distributions of atomic actions (*e.g.,* forward, turn left, turn right; predicted by LM-based VLN models [10, 48]) to assess output quality and guide adaptive integration of latent reconstructions. This design decouples robustness enhancement from the navigation backbone, making FlowDec broadly compatible with diverse LM-based VLN agents. Moreover, by leveraging flow matching instead of iterative diffusion sampling, FlowDec achieves substantially faster inference, meeting the stringent latency requirements of online robotic deployment. We evaluate FlowDec on R2R-CE and RxR-CE under six representative corruption types across more than 5,000 unseen trajectories. FlowDec improves relative navigation success by 25.33% and 9.38% on the two benchmarks, respectively, while achieving 3×–8× faster decorruption compared to representative diffusion-based TTA methods. We further validate its effectiveness in real-world robotic experiments, demonstrating consistent gains under practical sensory degradations. To summarize, our contributions are threefold:

1. We identify corruption robustness as a fundamental yet overlooked bottleneck for VLN in continuous environments and formulate navigation-aware decorruption as a real-time conditional flow-matching problem.
2. We propose FlowDec with hybrid temporal conditioning and action-centric guidance, enabling temporally consistent, physically meaningful restoration without modifying the navigation backbone.
3. Extensive experiments demonstrate that FlowDec achieves superior restoration speed and navigation success, establishing a new robustness paradigm for VLN under diverse corruptions.

2 Related Works

2.1 LM-based VLN Models

LLMs have advanced embodied AI by enabling strong reasoning and language grounding in robotic control [4, 22, 54] and discrete VLN planning [6, 35, 52, 53]. Recent VLN-CE works integrate LLMs for continuous navigation [10, 30, 47, 48]. End-to-end models like NaVid series fine-tune VLMs [11] to execute actions from RGB and language inputs [47,48], while NaVILA [10] adopts a hierarchical LLM planner + low-level controller [31] to improve navigation capability. LM-based VLN faces (i) limited spatial reasoning ability [44, 50], and (ii) sim-to-real gaps due to scarce real instruction and sensor data [2, 20, 21, 38]. Prior works improve grounding via chain-of-thought prompting [6,30,35,53] and additional data from video sources [10,27] or auxiliary tasks [47,48]. Robustness to visual corruptions remains largely unaddressed, despite strong impact in real deployment. Our work directly tackles this gap without retraining the VLN agent.

2.2 Image Decorruption

VLN robustness requires handling unpredictable visual corruptions that cannot be anticipated during training. Two primary paradigms exist: blind image denoising (BID) and test-time adaptation (TTA). BID removes unknown noise by learning to estimate corruption distributions [7, 18, 45, 49]. For example, SCUNet [49] synthesizes realistic degradations and uses pixel and adversarial losses for robust denoising. However, data-driven BID alone struggles with unseen real-world corruptions. TTA instead adapts to test-time inputs without source data [26, 42, 51]. Classical TTA methods focus on discrete actions [15, 33, 39] and are difficult to extend to VLN-CE, where pseudo-labels are ambiguous in continuous spaces and single-sample online adaptation risks catastrophic forgetting [14, 42, 43]. Diffusion-based TTA offers a promising alternative by correcting input images instead of modifying model parameters [14, 17, 34]. DDA [14] demonstrates improved robustness across noise types but requires many denoising steps and mainly targets noise rather than diverse corruptions [41]. Decorruptor [34] accelerates inference via latent denoising and augmentation but still processes frames independently. Yet VLN-CE requires real-time generation and temporal consistency across frames to maintain stable navigation. Our method addresses both challenges by improving efficiency and enforcing consistency across sequential observations, providing robust perception for online VLN-CE.

3 Temporal Conditional Flow Decorruptor

This section introduces FlowDec, a novel framework for mitigating image corruption in VLN-CE. FlowDec takes corrupted images and atomic action labels as input and outputs decorrupted images for the VLN agent. It is trained using

latent conditional flow matching (CFM) [13], augmented with a hybrid temporal conditioning strategy (Figure 2) to align the generative flow path with historical information. At inference, action centroids are used to assess inter-frame consistency and dynamically integrate outputs from multiple conditions (Figure 3). This mechanism ensures that generated images exhibit action-coherent transitions, enhancing spatial and temporal fidelity for robust downstream navigation.

3.1 Problem Formulation and Preliminaries

The objective of FlowDec is to reconstruct a decorrupted image $\mathbf{x}_{\mathrm{c\_decor}}$ from a corrupted input $\mathbf{x}_{\mathrm{c\_cor}}$ using CFM. Following prior CFM-based generation works [13, 16], FlowDec operates in the latent space using variable $\mathbf{z}$, obtained via a pretrained variational autoencoder (VAE) encoder $\mathcal{E}$ and reconstructed with decoder $\mathcal{D}$ [23].

Given the ground-truth observation image latent $\mathbf{z}_{\mathrm{c\_gt},1} = \mathcal{E}(\mathbf{x}_{\mathrm{c\_gt}}) \sim p_1$ and a random noise image latent $\mathbf{z}_0 \sim p_0$, the latent CFM network aims to estimate a coupling $\pi(p_0, p_1)$ that models the evolution between these distributions. This objective can be formulated as solving an ordinary differential equation (ODE):

$$\mathrm{d}\mathbf{z}_{\mathrm{c\_gt},t} = u(\mathbf{z}_{\mathrm{c\_gt},t}, t)\mathrm{d}t, \tag{1}$$

where $t \in [0, 1]$ denotes time, $u(\mathbf{z}_{\mathrm{c\_gt},t}, t) : [0, 1] \times \mathbb{R}^d \to \mathbb{R}^d$ is a time-varying velocity field, and d is the dimension of the latent feature space. By parameterizing the velocity field with a trainable neural network $v_\theta(\mathbf{z}_{\mathrm{c\_gt},t}, t)$, the estimation of θ reduces to a least-squares regression problem:

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{\mathbf{z}_{\mathrm{c\_gt},t},t}\left[\left\| v_\theta\left(\mathbf{z}_{\mathrm{c\_gt},t}, t\right) - u(\mathbf{z}_{\mathrm{c\_gt},t}, t)\right\|_2^2\right]. \tag{2}$$

To make the velocity field $u(\mathbf{z}_{\mathrm{c\_gt},t}, t)$ tractable, we introduce a conditioning latent variable $\mathbf{l}$ and define a marginal probability path $p_t(\mathbf{z}_{\mathrm{c\_gt},t} \mid \mathbf{l})$ that varies according to $\mathbf{l}$:

$$p_t(\mathbf{z}_{\mathrm{c\_gt},t}) = \int p_t(\mathbf{z}_{\mathrm{c\_gt},t} \mid \mathbf{l})q(\mathbf{l})\mathrm{d}\mathbf{l}, \tag{3}$$

where $q(\mathbf{l})$ is a distribution over the conditioning variable. Incorporating this into the flow matching framework, the CFM loss turns to:

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{\mathbf{z}_{\mathrm{c\_gt},t},\mathbf{l},t}\left[\left\| v_\theta\left(\mathbf{z}_{\mathrm{c\_gt},t}, t\right) - u(\mathbf{z}_{\mathrm{c\_gt},t}, t \mid \mathbf{l})\right\|_2^2\right]. \tag{4}$$

It has been shown [28, 40] that optimizing $\mathcal{L}_{\mathrm{FM}}(\theta)$ (Eq. 2) is equivalent to optimizing $\mathcal{L}_{\mathrm{CFM}}(\theta)$ (Eq. 4), *i.e.,* $\nabla_\theta \mathcal{L}_{\mathrm{FM}}(\theta) = \nabla_\theta \mathcal{L}_{\mathrm{CFM}}(\theta)$. Thus, for any conditioning variable $\mathbf{l}$, the optimization is valid provided the boundary conditions of $p_t(\mathbf{z}_{\mathrm{c\_gt},t} \mid \mathbf{l})$ are satisfied and the process remains continuous. Following [28, 29], we define the conditional probability path as a Gaussian distribution:

$$p_t(\mathbf{z}_{\mathrm{c\_gt},t} \mid \mathbf{l}) = \mathcal{N}(\mathbf{z}_{\mathrm{c\_gt},t} \mid t\mathbf{z}_{\mathrm{c\_gt},1} + (1-t)\mathbf{z}_0, \sigma^2), \tag{5}$$

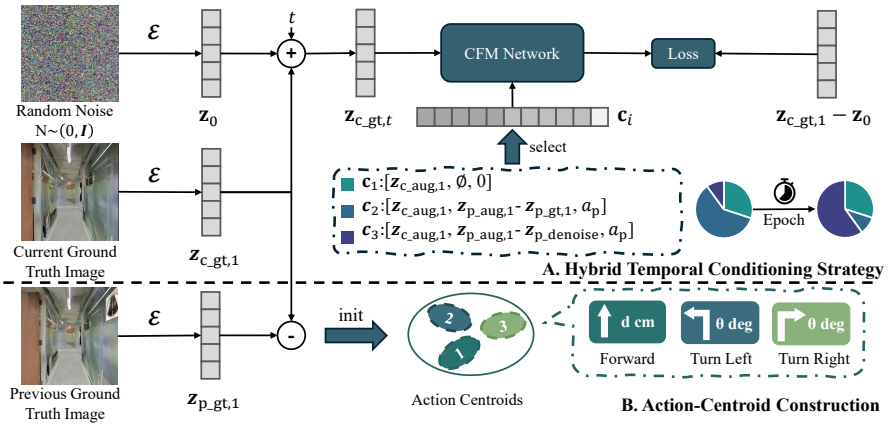

Fig. 2: Overview of training phase. **(A)** Model utilizes three types of conditions for training. The proportion of $\mathbf{c}_1$ remains constant, while $\mathbf{c}_3$ gradually replaces $\mathbf{c}_2$ as the number of epochs increases. **(B)** Ground-truth image pairs are used to construct action centroids, which capture expected latent differentials per atomic action.

where, as $\sigma \to 0$, the velocity field becomes constant:

$$u(\mathbf{z}_{\mathrm{c\_gt},t}, t \mid \mathbf{l}) = \mathbf{z}_{\mathrm{c\_gt},1} - \mathbf{z}_0, \tag{6}$$

and $\mathbf{z}_{\mathrm{c\_gt},t}$ reduces to a linear interpolation between $\mathbf{z}_{\mathrm{c\_gt},1}$ and $\mathbf{z}_0$:

$$\mathbf{z}_{\mathrm{c\_gt},t} = t\mathbf{z}_{\mathrm{c\_gt},1} + (1-t)\mathbf{z}_0. \tag{7}$$

3.2 Temporal Conditional Flow Matching Learning Strategy

Robust viewpoint recovery in sequential navigation demands temporal continuity and resilience to real-world corruptions. Standard flow matching on single images ignores inter-frame relationships and struggles with unseen degradations. While naive augmentation risks overfitting and fails to promote spatial understanding. Therefore simply apply these approaches are suitable for VLN-CE tasks. We propose a hybrid temporal conditioning strategy to align the generative flow path with historical context.

CFM Baseline. Building on the flow matching framework outlined in Section 3.1, the loss for CFM with a constant velocity field is expressed as:

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{\mathbf{z}_{\mathrm{c\_gt},t},t}\left[\left\|\mathbf{z}_{\mathrm{c\_gt},1} - \mathbf{z}_0 - v_\theta\left(\mathbf{z}_{\mathrm{c\_gt},t}, t\right)\right\|_2^2\right], \tag{8}$$

where $\mathbf{z}_{\mathrm{c\_gt},1} = \mathcal{E}(\mathbf{x}_{\mathrm{c\_gt}})$ is the latent representation of the ground-truth image, $\mathbf{z}_0$ is the random noise latent, and v_θ is the parameterized velocity field.

During inference, the corrupted image latent $\mathbf{z}_{\mathrm{c\_cor},1} = \mathcal{E}(\mathbf{x}_{\mathrm{c\_cor}})$ serves as conditional information $\mathbf{c}_i$, where i denotes different condition types. The CFM

loss is thus reformulated as:

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{\mathbf{z}_{\mathrm{c\_gt},t},t}\left[\left\|\mathbf{z}_{\mathrm{c\_gt,1}} - \mathbf{z}_0 - v_\theta\left(\mathbf{z}_{\mathrm{c\_gt},t}, \mathbf{c}_i, t\right)\right\|_2^2\right]. \tag{9}$$

Hybrid Temporal Conditions. Since corrupted images are inaccessible during training, we adopt the corruption modeling scheme from [34] to enhance the robustness of the FlowDec against unseen corruptions. This approach uses augmented images to simulate corruption, training the model to restore these images to their clean counterparts. We employ a hybrid augmentation strategy combining PIXMIX [19] and SimSiam [9], consistent with [34]. We introduce a hybrid temporal conditioning strategy through $\mathbf{c}_i$, which is defined as:

$$\mathbf{c}_i = \begin{cases} [\mathbf{z}_{\mathrm{c\_aug,1}}, \emptyset, 0] & , i = 1 \\ [\mathbf{z}_{\mathrm{c\_aug,1}}, \mathbf{z}_{\mathrm{p\_aug,1}} - \mathbf{z}_{\mathrm{p\_gt,1}}, a_p] & , i = 2 \\ [\mathbf{z}_{\mathrm{c\_aug,1}}, \mathbf{z}_{\mathrm{p\_aug,1}} - \mathbf{z}_{\mathrm{p\_denoise}}, a_p] & , i = 3 \end{cases}, \tag{10}$$

where $\mathbf{z}_{\mathrm{c\_aug,1}}$ is the latent of the current augmented image, $\mathbf{z}_{\mathrm{p\_aug,1}}$ and $\mathbf{z}_{\mathrm{p\_gt,1}}$ are the latents of the previous augmented and ground-truth images, respectively, $\mathbf{z}_{\mathrm{p\_denoise}}$ is the denoised latent from the previous frame, and a_p denotes the atomic action taken by the agent, as provided by LM-based navigation models [10,48]. The action a_p represents specific action with fixed movement amplitudes, capturing differential changes between consecutive image frames.

Condition Design and Training Schedule. Conditions in $\mathbf{c}_i$ correspond to different intensions. For $\mathbf{c}_1$, solely using current augmented image's latent features, represents the most commonly employed denoising method. While $\mathbf{c}_2$ and $\mathbf{c}_3$ incorporate information from the previous frame, enabling the model to leverage temporal context. Specifically, $\mathbf{c}_2$ approximates corruption features by computing the difference between the augmented and ground-truth latent of the previous frame, guiding the model toward effective image restoration. Since ground-truth images are unavailable during inference and the synthetic domain of generative models differs from that of real images [17,41], $\mathbf{c}_3$ uses the denoised latent from the previous frame to progressively bridge the domain gap between synthetic and real images.

During training, the three conditions are combined in a dynamic proportion. The weight of $\mathbf{c}_1$ remains constant, while $\mathbf{c}_2$ dominates in the early epochs to establish robust corruption modeling. As training stabilizes, the proportion of $\mathbf{c}_2$ is gradually reduced in favor of $\mathbf{c}_3$, allowing the model to adapt to its own domain gap.

3.3 Action-Centroid Guided Latent Filtering

Hybrid temporal conditioning enables dual inference paths using $\mathbf{c}_1$ (single-frame) and $\mathbf{c}_3$ (temporal). However, always using $\mathbf{c}_1$ discards valuable prior-frame context, while over-relying on $\mathbf{c}_3$ risks error accumulation during iterative inference—a critical failure mode in denoising. Our key insight: inter-frame consistency is tightly coupled with atomic actions. We thus propose action-centroid

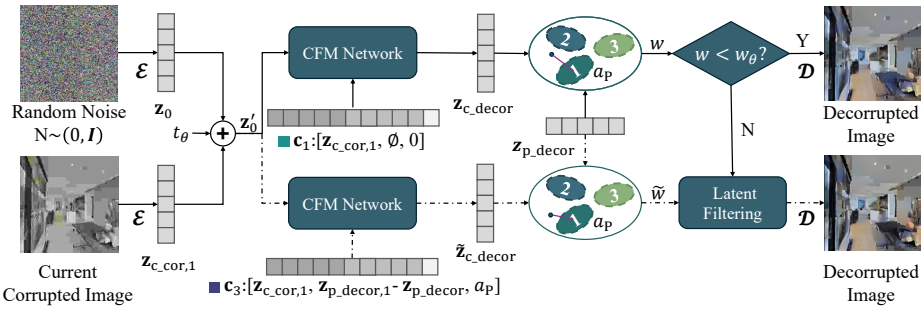

Fig. 3: Overview of the inference phase. The model primarily generates using condition $\mathbf{c}_1$. The auxiliary condition $\mathbf{c}_3$ is invoked only when the distance to the corresponding action centroid exceeds a threshold w_θ.

guided corruption filtering, which dynamically evaluates and fuses outputs from $\mathbf{c}_1$ and $\mathbf{c}_3$ based on action-specific latent distributions. This mechanism selectively leverages temporal cues while preserving single-frame stability, ensuring robust and consistent decorruption in sequential navigation (see Figure 3).

Action-Centroid Construction. To exploit navigation dynamics, we treat atomic actions as consistency priors. During training, we compute differential latents $\mathbf{z}_{\mathrm{p\_gt},1} - \mathbf{z}_{\mathrm{c\_gt},1}$ between consecutive ground-truth frames and record per-action mean and variance (Figure 2). This defines the action centroid:

$$A_{\mathrm{gt},a} = \mathcal{N}(\mu_{\mathrm{gt},a}, \sigma_{\mathrm{gt},a}), \tag{11}$$

where a denotes the type of atomic action. This centroid captures the expected latent differences associated with specific actions, facilitating consistent image reconstruction in continuous navigation tasks.

In sequential image generation, temporal conditioning improves model optimization and domain adaptation but can accumulate errors during iterative inference. To maintain stable and reliable outputs, we need a mechanism to selectively leverage temporal cues without propagating errors. We propose action-guided corruption filtering to balance temporal guidance and single-frame stability for robust latent reconstruction.

Model Inference and Latent Filtering. Although temporal conditions ($\mathbf{c}_2$, $\mathbf{c}_3$) provide valuable historical context, they risk accumulating generation errors during inference—particularly in iterative denoising during inference period, which is more lethal for iterative generative models. Therefore, we designate $\mathbf{c}_1$ as the primary condition and $\mathbf{c}_3$ as the auxiliary condition. Note that Reducing temporal conditioning reliance at inference does not diminish its training value. Since all conditioning objectives are aligned, flow path trained using temporal conditions also benefits $\mathbf{c}_1$ generation consistency (see validation in Section 4.3).

The inference procedure is illustrated in Figure 3. The model primarily generates decorrupted latent $\mathbf{z}_{\mathrm{c\_decor}}$ using condition $\mathbf{c}_1$. For the first frame, $\mathbf{z}_{\mathrm{c\_decor}}$ is output directly. For subsequent frames, the quality of the generated latent is

verified using action centroids. Specifically, we compute the mean Mahalanobis distance w between the differential features of the current and previous decorrupted latent ($\mathbf{z}_{\mathrm{p\_decor}} - \mathbf{z}_{\mathrm{c\_decor}}$) and the corresponding action centroid distribution, assuming all the dimensions are independent:

$$w = \sqrt{\sum_{k=1}^{d} \frac{\left(\mathbf{z}_{\mathrm{p\_decor},k} - \mathbf{z}_{\mathrm{c\_decor},k} - \mu_{\mathrm{gt},a_p,k}\right)^2}{d\sigma^2_{\mathrm{gt},a_p,k}}}, \tag{12}$$

where d is the size of latent dimension, $\mu_{\mathrm{gt},a_p,k}$ and $\sigma^2_{\mathrm{gt},a_p,k}$ are the mean and variance of the action centroid for action a_p. If w is below a predefined threshold w_θ, the output is $\mathbf{z}_{\mathrm{c\_decor}}$ from $\mathbf{c}_1$. Otherwise, model needs to decorrupt for the second round and calculate weight $\tilde{w}$ using $\mathbf{c}_3$. If $\tilde{w} < w$, the output result is represented as a weighted combination of $\mathbf{z}_{\mathrm{c\_decor}}$ and $\tilde{\mathbf{z}}_{\mathrm{c\_decor}}$:

$$\mathcal{E}(\mathbf{x}_{\mathrm{c\_decor}}) = \begin{cases} \frac{\tilde{w}}{\tilde{w}+w}\mathbf{z}_{\mathrm{c\_decor}} + \frac{w}{\tilde{w}+w}\tilde{\mathbf{z}}_{\mathrm{c\_decor}} & , \tilde{w} < w \\ \mathbf{z}_{\mathrm{c\_decor}} & , otherwise \end{cases}. \tag{13}$$

This design leverages the inherent stability of images generated with $\mathbf{c}_1$, invoking $\mathbf{c}_3$ only when current frame significantly deviates from previous frame, therefore balances inter-frame consistency with computational efficiency.

Typically, generation begins from random noise $\mathbf{z}_0$ and follows a complete integration path to $\mathbf{z}_1$. While we modify the starting point to further enhance inference efficiency:

$$\mathbf{z}'_0 = t_\theta \mathbf{z}_0 + (1 - t_\theta)\mathbf{z}_{\mathrm{c\_cor},1}. \tag{14}$$

Experiment in section 4.3 shows that This modification not only accelerates generation but also reduces randomness in the output, improving performance in different corruption scenarios.

4 Experiments

In this section, we mainly investigate three questions: **Q1**: How do different corruptions affect LM-based VLN models? **Q2**: Why does FlowDec outperform other SOTA decorruption models? **Q3**: Are all components of FlowDec necessary for its performance?

4.1 Implementation Details, Experiment Settings and Baselines

Implementation Details. In FlowDec, the conditional flow matching (CFM) module adopts a U-Net architecture following [13]. The VAE encoder and decoder are implemented based on the original designs in [13,45]. We train FlowDec for 20 epochs with a learning rate of 1×10^{-5}. The weighting coefficient w_θ is set to 0.25, and the time threshold t_θ is set to 0.95. FlowDec is a model-agnostic visual module that operates independently of the underlying navigation architecture. For evaluation, we adopt NaVid [48] as the baseline, as it is widely adopted

Table 1: Performance of NaVid under different corruption types on R2R-CE and RxR-CE datasets. All metrics are reported in %.

Dataset	Metric	Clean	Gauss.	Shot	Impul.	Defoc.	Motion.	Snow	Fog	Lightout	Pixel.	JPEG.	Cont.	Bright.
R2R-CE	SR	41.40	21.42	20.99	24.63	33.06	26.64	21.26	17.29	29.80	29.82	27.95	20.55	34.48
	SPL	35.85	18.93	19.42	22.56	29.83	23.70	18.43	14.33	26.65	25.93	25.10	18.43	31.65
RxR-CE	SR	45.65	31.33	30.12	34.64	40.56	42.27	30.40	29.15	39.36	39.14	34.24	29.25	40.53
	SPL	37.23	25.70	23.98	27.77	33.51	34.10	24.21	22.93	32.25	31.82	29.53	23.49	33.00

and representative among LM-based VLN frameworks [10,12,46,47]. All training and experiments are conducted on a single NVIDIA RTX 4090 GPU.

To ensure robustness under high intra-action variance across diverse geometric environments, we aggregate statistics from 2,440,814 image pairs collected from over 70 distinct scenes derived from the training splits of R2R-CE and RxR-CE. This large-scale aggregation enables the estimation of empirically stable action centroids with reliable covariance statistics. Additional implementation details are provided in the supplementary material.

Experiment Settings. Simulations are conducted using the Habitat platform [38] on the R2R-CE [24] and RxR-CE [25] datasets, comprising 1,839 and 3,669 episodes respectively, across over 50 scenes. Following [34,43], we construct 12 corruption types to assess the robustness of the backbone. Performance is measured using Success Rate (SR) and Success weighted by Path Length (SPL), while SPL is the primary metric due to its balance of accuracy and efficiency [1].

Real-world experiments are performed on the Unitree GO2 robot under four scenes, with 20 trials per setting (see Figure 5). An episode is successful if the agent stops within 3 m (simulation) or 0.5 m (real-world) of the goal. No real-world data is used for further training or fine-tuning.

Baselines. We compare FlowDec against following methods:

- SCUNet [49]: a representative BID approach.
- Dec-DPM and Dec-CM [34]: SOTA diffusion-based TTA using diffusion probabilistic models (5 NFEs) and consistency models (2 NFEs).

For fair comparison, Dec-DPM and Dec-CM are trained using the same data and augmentation strategies as FlowDec (PIXMIX [19] + SimSiam [9]). NFE settings are chosen to balance performance and inference speed.

4.2 Robustness Analysis of VLN Models

Table 1 reveals significant performance degradation under various corruptions. Among them, luminance variance (brightness and light-out) and image blur (motion blur and defocus blur) have minimal impact, while fog and snow cause severe drops. This disparity stems from domain mismatch in the training data: lighting variations and motion blur are well-represented in the source domain, enabling effective generalization. In contrast, weather effects (*e.g.,* Fog, Snow) and structured noise (*e.g.,* Gaussian) lie outside the training domain, despite their real-world prevalence, resulting in performance decline. Collecting real-world data

Table 2: Performance of different methods under corruption on R2R-CE and RxR-CE. All metrics are reported in %. The best results are highlighted in **bold**.

Method	Metric	R2R-CE							RxR-CE						
		Gauss.	Shot	Snow	Fog	JPEG.	Cont.	Avg.	Gauss.	Shot	Snow	Fog	JPEG.	Cont.	Avg.
NaVid	SR	21.42	20.99	21.26	17.29	27.95	20.55	21.58	31.33	30.12	30.40	29.15	34.24	29.25	30.75
	SPL	18.93	19.42	18.43	14.33	25.10	18.43	19.11	25.70	23.42	24.21	22.93	**29.53**	23.49	24.88
Dec-DPM	SR	22.62	22.84	20.71	24.90	22.73	29.85	23.94	24.48	25.43	19.27	25.94	24.97	29.95	25.01
	SPL	18.53	18.64	16.13	20.16	18.43	24.84	19.46	17.46	18.28	13.19	19.09	17.86	22.70	18.10
Dec-CM	SR	**28.53**	26.35	**23.27**	15.71	22.19	29.69	24.29	29.74	26.82	**31.34**	22.68	31.51	32.60	29.12
	SPL	**24.22**	22.62	**20.69**	13.46	20.86	27.24	21.52	23.13	20.88	**24.30**	16.98	25.85	26.23	22.90
SCUNet	SR	25.77	24.25	6.87	12.32	18.81	0.05	14.68	31.21	30.25	17.88	20.69	30.88	6.73	22.94
	SPL	23.26	21.50	6.47	12.11	17.83	0.04	13.54	24.18	23.38	14.82	15.93	26.22	6.66	18.53
FlowDec	SR	26.54	**26.70**	21.40	**26.87**	**29.85**	**30.90**	**27.04**	**32.33**	**31.35**	30.54	**35.93**	**35.76**	**35.88**	**33.63**
	SPL	23.53	**23.76**	18.38	**22.26**	**26.48**	**27.32**	**23.62**	**26.41**	**23.81**	24.17	**26.14**	29.20	**28.14**	**26.31**

Table 3: Performance of different methods using warp error(↓) on R2R-CE and RxR-CE. The best results are highlighted in **bold**.

Method	R2R-CE							RxR-CE						
	Gauss.	Shot	Snow	Fog	JPEG.	Cont.	Avg.	Gauss.	Shot	Snow	Fog	JPEG.	Cont.	Avg.
NaVid	0.47	0.52	0.12	0.28	**0.09**	**0.03**	0.25	0.47	0.51	0.14	0.28	**0.11**	**0.04**	0.26
Dec-DPM	0.30	0.29	0.34	0.33	0.36	0.27	0.31	0.32	0.31	0.35	0.34	0.37	0.29	0.33
Dec-CM	0.15	0.15	0.19	0.34	0.13	0.15	0.18	0.16	0.17	0.22	0.34	0.16	0.16	0.20
SCUNet	0.12	0.15	**0.11**	0.27	**0.09**	**0.03**	0.13	0.14	0.17	**0.13**	0.28	**0.11**	**0.04**	0.14
FlowDec	**0.10**	**0.11**	0.12	**0.15**	0.10	0.10	**0.11**	**0.12**	**0.13**	0.14	**0.18**	0.12	0.12	**0.13**

spanning such out-of-domain corruptions at scale is infeasible, highlighting the critical need for dedicated decorruption modules.

4.3 Performance Evaluation in R2R-CE and RxR-CE Navigation

In this section, we primarily focus on results for the 6 most destructive corruptions in Table 1, as these corruptions clearly reveal failure modes and limitations of baseline methods. Minor-impact corruptions are omitted for clarity.

As evidenced in Table 2, FlowDec consistently surpasses all baselines in aggregate performance, yielding relative SR gains of 25.33% and 9.37% over the vanilla model on R2R-CE and RxR-CE, respectively. The BID-based method demonstrates utility solely under Gaussian and shot noise; while in other corruptions its performance either degrades or collapses entirely. This situation is also confirmed in Figure 4(A) where denoising outcomes are illustrated across diverse corruptions. TTA approaches like Dec-DPM and Dec-CM on the other side, exhibit satisfactory denoising capabilities while their performance is generally inferior to FlowDec, and this gap becomes wider in RxR-CE.

Considering that Dec-DPM and Dec-CM are distilled from Instruct-Pix2Pix [5], a diffusion model trained on 45 million image-text pairs, while FlowDec solely utilizes R2R-CE and RxR-CE training splits for training. We attribute FlowDec's

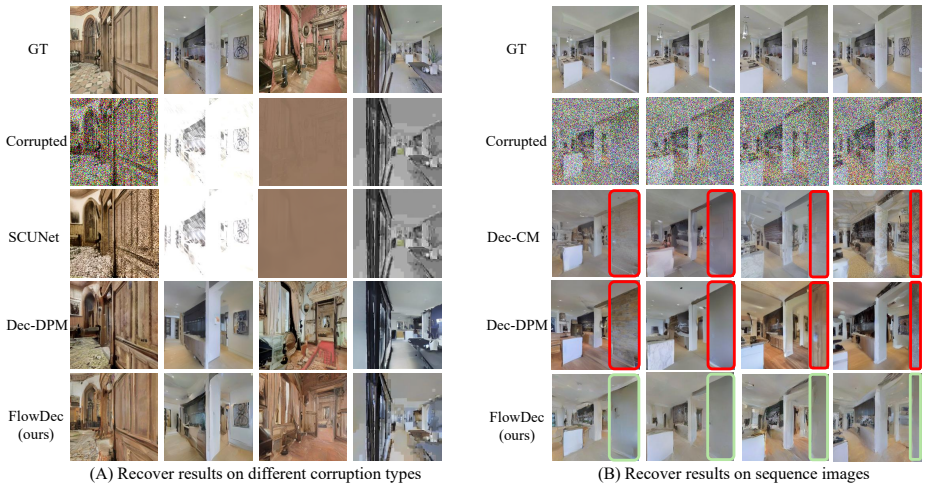

Fig. 4: Illustration of recovered images. (A) Recovery performance of different models under Gaussian noise, Snow, Contrast, and JPEG compression. (B) Recovery performance of different models for the same trajectory under Shot noise. Note in particular that for the same wall (marked within the frame), only our method yields results with high visual consistency.

Table 4: Comparison of model inference time.

Method	Time (ms)
NaVid	370
Dec-DPM	370+1411
Dec-CM	370+588
FlowDec	370+173

Table 5: Comparison on pristine conditions. All metrics are reported in %.

Method	Metrics	R2R-CE	RxR-CE
NaVid	SR	41.40	45.65
FlowDec	SR	40.89	45.41
NaVid	SPL	35.85	37.23
FlowDec	SPL	35.96	37.03

Table 6: Performance of model on real-world experiments. Each task is tested for 20 times.

Method	Indoor		Outdoor	
	Task1	Task2	Task3	Task4
NaVid	0.20	0.05	0.10	0.35
FlowDec	0.35	0.20	0.30	0.40

superior performance to inter-frame consistency. Figure 4(B) further elucidates this advantage by denoising consecutive frames along identical trajectories. While Dec-DPM and Dec-CM produce visually plausible per-frame outputs, they exhibit pronounced temporal inconsistencies (*e.g.,* walls highlighted in red frames). In contrast, FlowDec leverages hybrid temporal conditioning and action centroid guided filtering to ensure coherent generation, thereby facilitating more reliable scene interpretation by the downstream LM and enhancing navigation fidelity. Given that contemporary LM-driven VLN agents infer ego-localization and command compliance from image histories, inter-frame consistency emerges as a more decisive factor than isolated reconstruction quality.

To further evaluate FlowDec's temporal consistency, we randomly sample 100 test trajectories from R2R-CE and RxR-CE and measure performance using warp error. Unlike semantic-oriented metrics (*e.g.,* LPIPS), warp error directly quantifies motion consistency between consecutive frames and is less sensitive to corruption-induced semantic distortions.

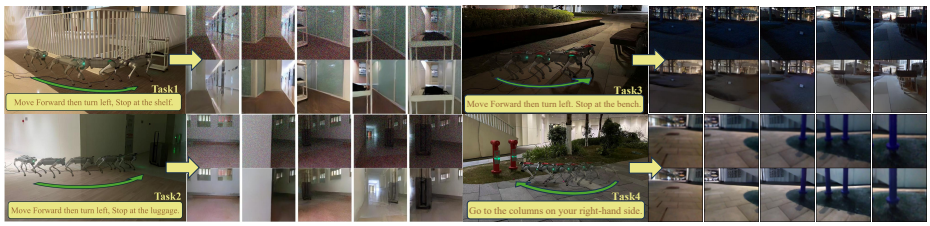

Fig. 5: Illustration of experiments in real-world environment. The two rows of images on the right side represent the original image and the decorrupted image respectively.

As shown in Table 3, FlowDec achieves the lowest average warp error across datasets and corruption types, with particularly significant improvements under motion-disruptive degradations such as Gaussian and shot noise. For corruptions that minimally affect apparent motion (*e.g.,* contrast shifts), semantic fidelity becomes relatively more important. Notably, we observe that substantial increases in warp error consistently lead to pronounced navigation degradation—even when semantic restoration appears adequate (*e.g.,* Dec-DPM under JPEG compression and Dec-CM under fog).

Inference latency remains a pivotal consideration for online VLN deployment. As reported in Table 4, FlowDec achieves single-step decorruption in merely 169 ms—delivering 3× to 8× speedup over Dec-DPM and Dec-CM. With episodes typically comprising ∼70 steps, the cumulative overhead of diffusion baselines renders them infeasible for real-time operation. FlowDec, uniquely, reconciles stringent robustness requirements with low-latency constraints, establishing it as a practical solution for embodied navigation in adverse visual conditions.

We further assess FlowDec's impact under pristine conditions. Results in Table 5 confirm that its integration incurs no statistically significant performance degradation on clean inputs. The minor degradation in clean unseen scenes is attributed to the domain gap between synthetic training corruptions and clean images [17, 41]. However, the trade-off is small compared to the statistically significant gains under corruption, which is the primary target scenario.

4.4 Deployment on Legged Robots.

To validate the effectiveness of our approach in real-world settings, we deploy our system on a legged robot and design four representative navigation tasks spanning both indoor and outdoor environments. The indoor tasks are conducted under synthetic Gaussian noise to simulate sensor degradation, while the outdoor tasks involve more severe luminance variations and motion blur caused by dynamic lighting and ego-motion. As reported in Table 6, FlowDec consistently outperforms the original baseline across all four tasks. Notably, the performance improvements are more substantial in the more challenging scenarios with pronounced luminance fluctuations (*e.g.,* Task 2 and Task 3), where perceptual degradation significantly impacts navigation reliability. These results

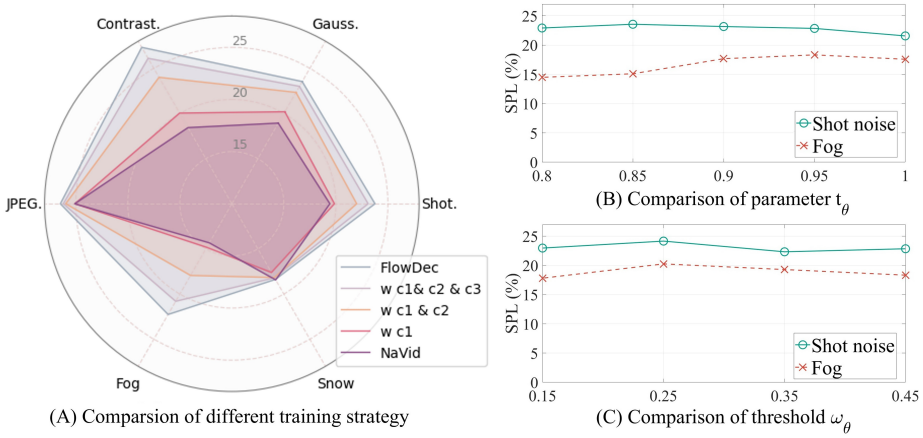

Fig. 6: Summary of ablation study results. All results are measured by SPL.

demonstrate that FlowDec generalizes effectively to previously unseen real-world conditions and robustly adapts to common environmental and sensor-induced disturbances.

4.5 Ablation Studies

Training strategy. The primary goal of incorporating temporal conditions ($\mathbf{c}_2$, $\mathbf{c}_3$) during training is to align the model's flow path with the primary condition ($\mathbf{c}_1$). Since all conditioning objectives are mutually consistent, auxiliary conditions enhance optimization of the main flow, improving $\mathbf{c}_1$ robustness even without temporal input at inference. To validate this, we evaluate models trained under different strategies on R2R-CE, using only $\mathbf{c}_1$ (without filtering) during inference. SPL results (Figure 6(A)) show universal gains after introducing the Temporal Condition Training Strategy. These improvements stem from enhanced spatial consistency: aligned flow paths produce coherent object representations across frames, even in single-frame inference. The term 'w $\mathbf{c}_1$&$\mathbf{c}_2$&$\mathbf{c}_3$' refers to the FlowDec training strategy without action-centroid guided latent filtering.

Various starting points. To accelerate inference, we modify the flow matching starting point via t_θ, initializing from a linear interpolation between random noise and the corrupted latent. While this reduces denoising capacity, it enhances consistency with the input image. Figure 6(B) confirms this trade-off: on shot noise and fog, SPL is lower when starting from pure noise ($t_\theta = 1$) than from an intermediate state. A small proportion of corrupted input ($t_\theta = 0.95$) significantly improves performance, whereas a larger proportion diminishes the ability of the model to decorruption, leading to a gradual decline in effectiveness.

Filtering threshold. The filtering threshold w_θ controls the contribution of the auxiliary condition $\mathbf{c}_3$ based on Mahalanobis distance to the action centroid, enabling adaptive use of historical context for spatial consistency. Across

corruptions, this distance ranges from 0.15 to 0.45. Figure 6(C) evaluates shot noise and fog measured by SPL: moderate integration ($w_\theta \geq 0.25$) optimally balances temporal guidance and primary condition fidelity, maximizing SPL. While excessive reliance on $\mathbf{c}_3$ ($w_\theta < 0.25$) degrades performance as well as increases unnecessary inference times.

5 Conclusion

VLN plays a fundamental role in embodied AI and continues to attract growing attention. In this work, we address the vulnerability of existing LM-based VLN models to visual corruptions by proposing the FlowDec, a novel module designed to enhance robustness. Through its hybrid temporal conditioning strategy and action-centroid guided filtering, FlowDec effectively mitigates the impact of unseen corruptions while preserving temporal consistency across generated images. This results in substantial performance improvements in both simulated and real-world environments. Moreover, the rapid inference speed of FlowDec supports real-time navigation, making it highly practical for online deployment. Looking ahead, we plan to extend this framework by developing denoising models robust to dynamically varying noise patterns, further advancing the resilience and applicability of VLN systems in more complex real-world scenarios.

Acknowledgments

This work was supported by National Natural Science Foundation of China (NFSC) under the Grant Number 62573370 and Key Area Project of Education Department of Guangdong Province (No. 2025ZDZX3051).

References

1. Anderson, P., Chang, A., Chaplot, D.S., Dosovitskiy, A., Gupta, S., Koltun, V., Kosecka, J., Malik, J., Mottaghi, R., Savva, M., et al.: On evaluation of embodied navigation agents. arXiv preprint arXiv:1807.06757 (2018)
2. Anderson, P., Shrivastava, A., Truong, J., Majumdar, A., Parikh, D., Batra, D., Lee, S.: Sim-to-real transfer for vision-and-language navigation. In: Conference on Robot Learning. pp. 671–681. PMLR (2021)
3. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., Van Den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3674–3683 (2018)
4. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al.: pi_0 : A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164 (2024)
5. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)

6. Chen, J., Lin, B., Xu, R., Chai, Z., Liang, X., Wong, K.Y.K.: Mapgpt: Map-guided prompting with adaptive path planning for vision-and-language navigation. arXiv preprint arXiv:2401.07314 (2024)
7. Chen, J., Chen, J., Chao, H., Yang, M.: Image blind denoising with generative adversarial network based noise modeling. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3155–3164 (2018)
8. Chen, S., Guhur, P.L., Tapaswi, M., Schmid, C., Laptev, I.: Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16537–16547 (2022)
9. Chen, X., He, K.: Exploring simple siamese representation learning. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2021), http://dx.doi.org/10.1109/cvpr46437.2021.01549
10. Cheng, A.C., Ji, Y., Yang, Z., Gongye, Z., Zou, X., Kautz, J., Bıyık, E., Yin, H., Liu, S., Wang, X.: Navila: Legged robot vision-language-action model for navigation. arXiv preprint arXiv:2412.04453 (2024)
11. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See https://vicuna.lmsys.org **2**(3), 6 (2023)
12. Contributors, I.: InternNav: InternRobotics' open platform for building generalized navigation foundation models (2025)
13. Dao, Q., Phung, H., Nguyen, B., Tran, A.: Flow matching in latent space. arXiv preprint arXiv:2307.08698 (2023)
14. Gao, J., Zhang, J., Liu, X., Darrell, T., Shelhamer, E., Wang, D.: Back to the source: Diffusion-driven adaptation to test-time corruption. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11786–11796 (2023)
15. Gao, J., Yao, X., Xu, C.: Fast-slow test-time adaptation for online vision-and-language navigation. arXiv preprint arXiv:2311.13209 (2023)
16. Gode, S., Nayak, A., Burgard, W.: Flownav: Learning efficient navigation policies via conditional flow matching. In: 2nd CoRL Workshop on Learning Effective Abstractions for Planning (2024)
17. Guo, J., Zhao, J., Du, C., Wang, Y., Ge, C., Ni, Z., Song, S., Shi, H., Huang, G.: Everything to the synthetic: Diffusion-driven test-time adaptation via synthetic-domain alignment. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 30503–30513 (2025)
18. Guo, S., Yan, Z., Zhang, K., Zuo, W., Zhang, L.: Toward convolutional blind denoising of real photographs. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1712–1722 (2019)
19. Hendrycks, D., Zou, A., Mazeika, M., Tang, L., Li, B., Song, D., Steinhardt, J.: Pixmix: Dreamlike pictures comprehensively improve safety measures. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2022), http://dx.doi.org/10.1109/cvpr52688.2022.01628
20. Hong, H., Qiao, Y., Wang, S., Liu, J., Wu, Q.: General scene adaptation for vision-and-language navigation. arXiv preprint arXiv:2501.17403 (2025)
21. Kamath, A., Anderson, P., Wang, S., Koh, J.Y., Ku, A., Waters, A., Yang, Y., Baldridge, J., Parekh, Z.: A new path: Scaling vision-and-language navigation with synthetic instructions and imitation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10813–10823 (2023)

22. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 (2024)
23. Kingma, D., Welling, M.: Auto-encoding variational bayes. arXiv: Machine Learning,arXiv: Machine Learning (Dec 2013)
24. Krantz, J., Wijmans, E., Majumdar, A., Batra, D., Lee, S.: Beyond the nav-graph: Vision-and-language navigation in continuous environments. In: European Conference on Computer Vision. pp. 104–120. Springer (2020)
25. Ku, A., Anderson, P., Patel, R., Ie, E., Baldridge, J.: Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. arXiv preprint arXiv:2010.07954 (2020)
26. Liang, J., He, R., Tan, T.: A comprehensive survey on test-time adaptation under distribution shifts. International Journal of Computer Vision **133**(1), 31–64 (2025)
27. Lin, K., Chen, P., Huang, D., Li, T.H., Tan, M., Gan, C.: Learning vision-and-language navigation from youtube videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8317–8326 (2023)
28. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
29. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
30. Long, Y., Cai, W., Wang, H., Zhan, G., Dong, H.: Instructnav: Zero-shot system for generic instruction navigation in unexplored environment. arXiv preprint arXiv:2406.04882 (2024)
31. Miki, T., Lee, J., Hwangbo, J., Wellhausen, L., Koltun, V., Hutter, M.: Learning robust perceptive locomotion for quadrupedal robots in the wild. Science robotics **7**(62), eabk2822 (2022)
32. Niu, S., Wu, J., Zhang, Y., Chen, Y., Zheng, S., Zhao, P., Tan, M.: Efficient test-time model adaptation without forgetting. In: International conference on machine learning. pp. 16888–16905. PMLR (2022)
33. Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., Tan, M.: Towards stable test-time adaptation in dynamic wild world. arXiv preprint arXiv:2302.12400 (2023)
34. Oh, Y., Lee, J., Choi, J., Jung, D., Hwang, U., Yoon, S.: Efficient diffusion-driven corruption editor for test-time adaptation. In: European Conference on Computer Vision. pp. 184–201. Springer (2024)
35. Pan, B., Panda, R., Jin, S., Feris, R., Oliva, A., Isola, P., Kim, Y.: Langnav: Language as a perceptual representation for navigation. arXiv preprint arXiv:2310.07889 (2023)
36. Park, S.M., Kim, Y.G.: Visual language navigation: a survey and open challenges. Artificial Intelligence Review p. 365–427 (Jan 2023), http://dx.doi.org/10.1007/s10462-022-10174-9
37. Qi, Y., Wu, Q., Anderson, P., Wang, X., Wang, W.Y., Shen, C., Hengel, A.v.d.: Reverie: Remote embodied visual referring expression in real indoor environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9982–9991 (2020)
38. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., et al.: Habitat: A platform for embodied ai research. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9339–9347 (2019)

39. Tan, M., Chen, P., Zhi, H., Mai, J., Rosman, B., Ji, D., Zeng, R.: Source-free elastic model adaptation for vision-and-language navigation. IEEE Transactions on Multimedia (2025)
40. Tong, A., Fatras, K., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Wolf, G., Bengio, Y.: Improving and generalizing flow-based generative models with minibatch optimal transport. arXiv preprint arXiv:2302.00482 (2023)
41. Tsai, Y.Y., Chen, F.C., Chen, A.Y., Yang, J., Su, C.C., Sun, M., Kuo, C.H.: Gda: Generalized diffusion for robust test-time adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23242–23251 (2024)
42. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726 (2020)
43. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7201–7211 (2022)
44. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10632–10643 (2025)
45. Yue, Z., Yong, H., Zhao, Q., Meng, D., Zhang, L.: Variational denoising network: Toward blind noise modeling and removal. Advances in neural information processing systems **32** (2019)
46. Zeng, S., Qi, D., Chang, X., Xiong, F., Xie, S., Wu, X., Liang, S., Xu, M., Wei, X.: Janusvln: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. arXiv preprint arXiv:2509.22548 (2025)
47. Zhang, J., Wang, K., Wang, S., Li, M., Liu, H., Wei, S., Wang, Z., Zhang, Z., Wang, H.: Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. arXiv preprint arXiv:2412.06224 (2024)
48. Zhang, J., Wang, K., Xu, R., Zhou, G., Hong, Y., Fang, X., Wu, Q., Zhang, Z., Wang, H.: Navid: Video-based vlm plans the next step for vision-and-language navigation. arXiv preprint arXiv:2402.15852 (2024)
49. Zhang, K., Li, Y., Liang, J., Cao, J., Zhang, Y., Tang, H., Fan, D.P., Timofte, R., Gool, L.V.: Practical blind image denoising via swin-conv-unet and data synthesis. Machine Intelligence Research **20**(6), 822–836 (2023)
50. Zhang, Y., Ma, Z., Li, J., Qiao, Y., Wang, Z., Chai, J., Wu, Q., Bansal, M., Kordjamshidi, P.: Vision-and-language navigation today and tomorrow: A survey in the era of foundation models. arXiv preprint arXiv:2407.07035 (2024)
51. Zhang, Y., Xu, Y., Wei, H., Lin, Z., Zou, X., Chen, C., Zhuang, H.: Analytic continual test-time adaptation for multi-modality corruption. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 1929–1937 (2025)
52. Zhou, G., Hong, Y., Wang, Z., Wang, X.E., Wu, Q.: Navgpt-2: Unleashing navigational reasoning capability for large vision-language models. In: European Conference on Computer Vision. pp. 260–278. Springer (2024)
53. Zhou, G., Hong, Y., Wu, Q.: Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7641–7649 (2024)
54. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: Conference on Robot Learning. pp. 2165–2183. PMLR (2023)