PRIVACYNAV: CONTROLLING VISUAL REPRESENTATION DISCLOSURE IN VISION-LANGUAGE NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Navigation (VLN) systems rely on semantically rich visual representations for scene understanding and action generation. However, in modular, cross-device, or remote inference, these representations are often transmitted beyond the trusted perception boundary, where an honest-but-curious recipient could recover policy-defined sensitive information. To address this issue, we propose **PrivacyNav**, a state-adaptive visual disclosure framework that controls *which* visual evidence is disclosed and *how much* is disclosed at each navigation state, while keeping the underlying VLN policy frozen. PrivacyNav comprises two modules. *Privacy-Aware Evidence Estimator* predicts token-level disclosure safety together with state-level disclosure risk, yielding localized and contextual signals. *Adaptive Visual Disclosure Controller* ranks visual tokens by jointly considering disclosure safety and host-native navigation evidence, and maps state-level risk to a state-specific token retention ratio. We conduct extensive experiments across four heterogeneous VLN architectures, complemented by module ablations, operating-point analyses, and real-world deployment. Compared with disclosing unmodified representations, PrivacyNav reduces macro sensitive-concept recovery by 3.23–10.38 percentage points, with navigation success rate losses of 0.81–5.11 percentage points. These results demonstrate consistent reductions in sensitive semantic recoverability under the evaluated protocols, with host-dependent disclosure–utility trade-offs rather than a formal privacy guarantee.

1 INTRODUCTION

Vision-Language Navigation (VLN) enables embodied agents to follow natural-language instructions by grounding them in visual observations and navigation history Anderson et al. (2018); Krantz et al. (2020). Recent VLN systems increasingly leverage semantically rich representations produced by powerful vision encoders and vision-language models Wei et al. (2026); Zhang et al. (2025); Xue et al. (2026); Zhang et al. (2026), supporting stronger scene understanding and action generation. However, when VLN is deployed in modular, cross-device, or remote inference pipelines, intermediate visual representations may cross the trusted perception boundary while retaining recoverable environmental semantics, including sensitive information Xiu & Zhang (2025); Lu et al. (2022); Li et al. (2022). Existing VLN representations are optimized primarily for navigation and provide no explicit control over which semantics remain recoverable after crossing this boundary. This raises a representation-level disclosure problem: how can visual evidence be selectively disclosed while preserving the information required for navigation?

Controlling representation-level disclosure in VLN presents two critical challenges. First, **the overlap between navigation relevance and policy-defined sensitivity complicates token selection, as the same visual evidence can be useful for navigation yet sensitive to disclose**. For example, a bed may serve as a target landmark and also belong to a sensitive category, making its removal detrimental to navigation and its retention a potential source of sensitive disclosure (Figure 1(b)). This motivates jointly ranking visual tokens by navigation relevance and disclosure safety, prioritizing safer useful evidence while balancing the utility and sensitivity of dual-use evidence.

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.

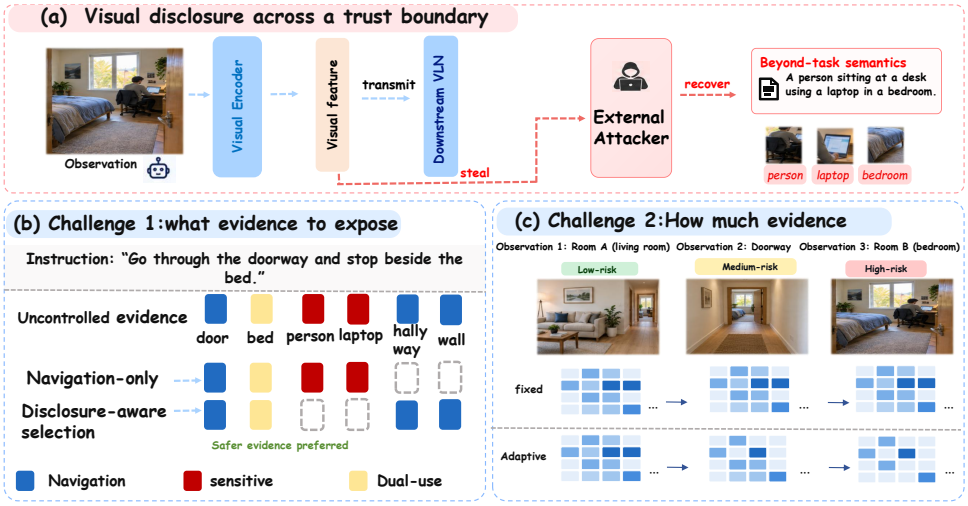

Figure 1: **Motivation of PrivacyNav.** Visual representations disclosed during VLN inference may expose policy-defined sensitive semantics that can overlap with navigation-relevant evidence. PrivacyNav controls *which* visual evidence is disclosed and *how much* is disclosed at each navigation state.

Second, **the variation in policy-defined sensitivity across navigation states complicates token budget allocation, as the same retention ratio can be adequate in one state yet excessive in another**. For instance, an ordinary corridor may contain little sensitive evidence, whereas a subsequent bedroom observation may reveal sensitive personal belongings, making a uniform retention ratio poorly suited to their different disclosure risks (Figure 1(c)). This motivates adapting the token retention ratio to the estimated risk of each navigation state, limiting visual disclosure in higher-risk states while retaining more navigation evidence in lower-risk states.

To address these challenges, we propose **PrivacyNav**, a state-adaptive visual disclosure framework for VLN that controls *which* visual evidence is disclosed and *how much* is disclosed at each navigation state. PrivacyNav consists of two modules: (1) a **Privacy-Aware Evidence Estimator**, which estimates token-level disclosure safety and state-level disclosure risk, providing localized and contextual signals for visual disclosure control; and (2) an **Adaptive Visual Disclosure Controller**, which uses the estimated state-level risk to adapt the token retention ratio and selects visual tokens by jointly considering disclosure safety and host-native navigation evidence. Together, these two modules construct a state-dependent disclosed representation without retraining the underlying VLN policy. We evaluate PrivacyNav across diverse VLN architectures and show that, under the evaluated recovery protocols, it consistently reduces macro sensitive-concept recovery by 3.23–10.38 percentage points, with navigation success rate losses of 0.81–5.11 percentage points. The StreamVLN ablation indicates that disclosure-aware evidence selection primarily accounts for the reduction in sensitive-semantic recoverability, whereas state-adaptive token retention improves navigation utility with a small increase in sensitive-concept recovery at the evaluated operating point.

Our main contributions are:

- We formulate **representation-level semantic disclosure in VLN** and evaluate both general and policy-defined sensitive-semantic recoverability from disclosed visual representations.
- We propose **PrivacyNav**, which combines privacy-aware disclosure estimation with adaptive visual disclosure control to determine *which* visual evidence is disclosed and *how much* is disclosed at each navigation state.
- We evaluate PrivacyNav across four heterogeneous VLN hosts, with module ablations, disclosure–utility analyses, in-domain semantic-recovery evaluation, and real-world deployment, demonstrating consistent reductions in sensitive-concept recovery under the evaluated protocols.

2 Related Work

Vision-Language Navigation. VLN agents follow natural-language instructions through visual grounding and sequential decision-making Anderson et al. (2018); Krantz et al. (2020). Prior work advances history modeling and global–local planning Chen et al. (2021; 2022), while recent systems employ powerful vision-language models for navigation Wei et al. (2026); Zhang et al. (2025); Xue et al. (2026); Zhang et al. (2026). PrivacyNav complements these advances by controlling semantic disclosure from visual representations crossing the trusted perception boundary, while keeping the navigation policy frozen.

Efficient visual representations. Token pruning and merging reduce visual computation by selecting informative tokens or combining redundant ones Rao et al. (2021); Liang et al. (2022); Bolya et al. (2023). Related methods accelerate vision-language models through attention-guided pruning and compression Chen et al. (2024); Yang et al. (2025). These approaches prioritize efficiency and task performance; token reduction alone does not determine which sensitive semantics remain recoverable. A token can be redundant for navigation yet still reveal sensitive scene content, while a navigation-critical token can itself be sensitive. This distinction motivates evaluating representation selection through both navigation utility and semantic recoverability. PrivacyNav explicitly incorporates disclosure safety alongside navigation evidence and adapts token retention to state-level risk.

Privacy and semantic disclosure. Prior work studies information recoverability through representation inversion Mahendran & Vedaldi (2015), reconstruction attacks and semantic recovery Li et al. (2022); Xiu & Zhang (2025), and private attribute inference Tömekçe et al. (2024). Defenses use dependence minimization, perturbation, or privacy-aware representation learning Vepakomma et al. (2020); Sun et al. (2021); Li et al. (2022; 2023); Athanasiou et al. (2026). Private multiparty perception also addresses privacy in navigation Lu et al. (2022). PrivacyNav studies state-adaptive visual evidence disclosure for frozen VLN hosts, accounting for the overlap between sensitive semantics and navigation requirements. Its evaluation measures empirical semantic recoverability under fixed protocols.

3 Method

3.1 Problem Formulation

At navigation step t, a VLN host produces a visual representation $V_t = \{v_{t,i}\}_{i=1}^{N_t}$ consisting of N_t visual tokens for downstream reasoning and action prediction. These host-native tokens carry the visual evidence whose disclosure PrivacyNav controls. We consider modular inference settings such as split-DNN, cross-device, and remote-inference pipelines, where the perception module executes inside a trusted domain while intermediate visual representations are transmitted to an external or less-trusted downstream component. The original RGB observation, the visual encoder, and the internal perception process remain inside the trusted domain. PrivacyNav controls visual evidence disclosure within this domain, before the resulting representation crosses the system boundary.

Threat model. We consider an honest-but-curious recipient or external observer that accesses the visual representation transmitted across the system boundary and attempts to recover environmental semantics. The adversary may know the host architecture and representation format, but cannot access the original RGB observation or modify the trusted perception module. We assess both general scene semantics and *policy-defined sensitive semantics* using the fixed recovery protocols in Sec. 4. We do not claim robustness against adaptive adversaries that retrain recovery models on PrivacyNav outputs, nor do we evaluate inference based on auxiliary side channels such as token counts or selection patterns. Our results measure empirical semantic recoverability under these protocols and do not constitute a formal privacy guarantee.

PrivacyNav constructs a disclosed representation $\widetilde{V}_t$ from V_t before it crosses the system boundary. Here, V_t denotes the host-specific visual token set subject to disclosure control, which may be a subset of the visual representations maintained by the host. Our objective is to reduce the recoverability of *policy-defined sensitive semantics* from the disclosed representation while retaining visual evidence needed for navigation. Sensitive semantics are a policy-defined subset of environmental

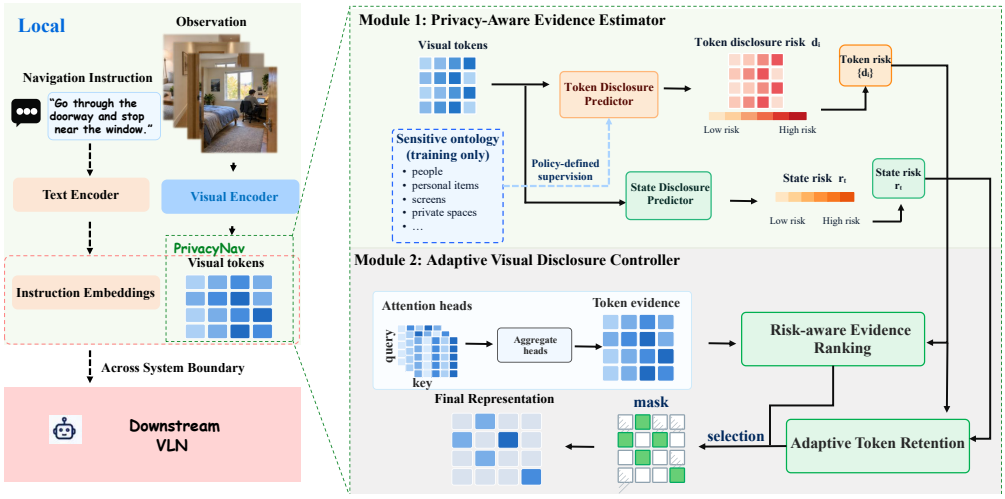

Figure 2: **Overview of PrivacyNav, illustrated on StreamVLN.** The **Privacy-Aware Evidence Estimator** provides token-level disclosure safety and state-level risk signals. The **Adaptive Visual Disclosure Controller** combines disclosure safety with host-native navigation evidence to select tokens and adapts their retention ratio to state-level risk. The resulting representation crosses the system boundary for downstream navigation.

semantics and may overlap with navigation-relevant information; they are not synonymous with task-irrelevant content. PrivacyNav therefore controls visual evidence disclosure without seeking to eliminate all recoverable semantics.

3.2 PrivacyNav Framework

PrivacyNav augments a frozen VLN host with two modules: a **Privacy-Aware Evidence Estimator** and an **Adaptive Visual Disclosure Controller** (Figure 2).

Given V_t, the estimator produces token-level disclosure safety scores $\{s^{\text{safe}}_{t,i}\}_{i=1}^{N_t}$ and a state-level risk score R_t. The controller uses R_t to determine the retention count K_t and combines disclosure safety with host-native navigation evidence $n_{t,i}$ to compute selection scores $q_{t,i}$. The selected indices I_t determine the disclosed representation $\widetilde{V}_t$:

$$\begin{aligned} R_t &\to K_t, & \left(s^{\text{safe}}_{t,i}, n_{t,i}\right) &\to q_{t,i}, \\ \left(\{q_{t,i}\}_{i=1}^{N_t}, K_t\right) &\to I_t, & \left(V_t, I_t\right) &\to \widetilde{V}_t. \end{aligned} \tag{1}$$

3.3 Privacy-Aware Evidence Estimator

Existing VLN representations are optimized for navigation and provide no explicit indication of policy-defined sensitive semantic exposure. We therefore use a dual-head estimator to assess visual evidence at token and state levels. Its outputs serve as sensitivity-based proxies for disclosure risk rather than guarantees of semantic non-recoverability.

Token-level Disclosure Estimation. For each token $v_{t,i}$, the estimator predicts a disclosure safety score $s^{\text{safe}}_{t,i} \in [0,1]$, where higher values indicate visual evidence estimated to be safer to disclose under the specified sensitive policy. The corresponding token-level risk score is

$$d_{t,i} = 1 - s^{\text{safe}}_{t,i}. \tag{2}$$

Because sensitive evidence may be spatially localized, averaging risk scores across all tokens can dilute the contribution of high-risk regions. We therefore aggregate the highest-risk fraction:

$$r^{\text{token}}_t = \text{MeanTop}_{\lceil \eta N_t \rceil}\left(\{d_{t,i}\}_{i=1}^{N_t}\right), \tag{3}$$

where $\mathrm{MeanTop}_k$ averages the k largest scores and $\eta = 0.1$ in our implementation. Risk-estimation inputs and token-selection rules are host-specific; implementation details are provided in Appendices A and B.

State-level Disclosure Estimation. A second head uses global scene context to predict a state-level sensitivity score r_t^{state}, supervised by scene-level pseudo-labels. Since the aggregated token risk and the state-head score may follow different distributions, we map them to a common percentile scale using empirical cumulative distribution functions (CDFs) estimated on a held-out calibration split:

$$\begin{aligned} p_t^{\text{token}} &= F_{\text{token}}(r_t^{\text{token}}), \\ p_t^{\text{state}} &= F_{\text{state}}(r_t^{\text{state}}), \\ R_t &= \omega p_t^{\text{token}} + (1-\omega) p_t^{\text{state}}, \end{aligned} \tag{4}$$

where $\omega \in [0, 1]$ balances the two signals. The resulting R_t is a normalized state-level disclosure risk score used by the controller; it is not a calibrated probability of semantic recovery.

The two signals capture complementary aspects of disclosure risk: token risk highlights localized sensitive evidence, while the state head captures broader scene context. For example, a sensitive object may appear in an otherwise ordinary room, while room context may indicate sensitivity beyond individual objects. Their fusion incorporates both cues into the retention decision.

Estimator Supervision. We train the estimator with policy-defined automatic weak supervision. Scene-level pseudo-labels supervise the state head, while detections from a fixed sensitive-object ontology are mapped to host visual tokens to provide localized supervision for the token head. The token-labeling protocol specifies how ambiguous tokens outside detected sensitive regions are handled, rather than assuming that all such tokens are verified safe negatives.

The estimator is optimized by

$$\mathcal{L}_{\text{est}} = \mathcal{L}_{\text{token}} + \beta \mathcal{L}_{\text{state}}, \tag{5}$$

where $\mathcal{L}_{\text{token}}$ and $\mathcal{L}_{\text{state}}$ denote the token-level and state-level supervised losses, respectively, and β controls their relative weight. The VLN backbone remains frozen, and neither navigation rewards nor semantic recovery metrics are used as training objectives. Pseudo-label generation, token-label assignment, loss definitions, network architecture, calibration procedures, and parameter settings are provided in Appendix A.

3.4 Adaptive Visual Disclosure Controller

The controller determines which visual evidence is disclosed to the downstream VLN policy and how much is disclosed through token selection.

State-Adaptive Token Retention. A fixed token retention ratio applies the same selection proportion across navigation states despite differences in disclosure risk. PrivacyNav maps an effective state-level risk score to a host-specific token retention ratio. For StreamVLN, a high-risk decision persists for two navigation steps; in hosts without temporal hold, the effective score equals the current risk score:

$$\rho_t = g_h(\overline{R}_t), \qquad K_t = \lceil \rho_t N_t \rceil, \tag{6}$$

where $\overline{R}_t$ is the effective risk after any temporal hold, ρ_t is the token retention ratio, and K_t is the number of selected visual tokens. The mapping g_h is non-increasing in $\overline{R}_t$: higher-risk states retain a smaller or equal fraction of tokens. Its host-specific configuration accommodates differences in sensitivity to visual token reduction across VLN architectures. Appendix D reports the release settings and selected-token ratios used in the evaluated configurations.

Disclosure-Aware Evidence Selection. The token retention ratio alone does not determine which visual evidence is retained. PrivacyNav therefore combines disclosure safety with host-native navigation evidence $n_{t,i}$, an attention-derived proxy for navigation relevance obtained from the frozen VLN host. This avoids an additional learned task-necessity head. The navigation evidence must be available within the trusted domain before the disclosed representation crosses the system boundary. Host-specific interfaces are summarized in Table 5 of Appendix B.

For token-wise rank fusion, the two signals are normalized and combined as follows:

$$\begin{aligned} \widehat{s}_{t,i} &= \mathrm{RankNorm}(s^{\mathrm{safe}}_{t,i}), \\ \widehat{n}_{t,i} &= \mathrm{RankNorm}(n_{t,i}), \\ q_{t,i} &= \alpha \widehat{s}_{t,i} + (1-\alpha)\widehat{n}_{t,i}, \end{aligned} \tag{7}$$

where higher input scores receive higher normalized ranks, and $\alpha \in [0, 1]$ balances disclosure safety and navigation evidence.

The ranking determines which tokens are preferred, while the retention count determines how many are selected. Including navigation evidence allows useful but sensitive landmarks to remain eligible for selection, reflecting the overlap between navigation relevance and policy-defined sensitivity.

For token-wise selection, the selected indices and disclosed representation are

$$\begin{aligned} I_t &= \mathrm{TopK}\left(\{q_{t,i}\}_{i=1}^{N_t}, K_t\right), \\ \widetilde{V}_t &= \mathcal{A}_h(V_t, I_t), \end{aligned} \tag{8}$$

where TopK returns the indices of the K_t highest-scoring tokens and $\mathcal{A}_h$ denotes the host-specific representation adapter. Selected tokens are restored to their original visual order before disclosure. The adapter constructs the output in a format compatible with the frozen host, using token pruning or structure-preserving operations. Here, the retention ratio describes the fraction of selected visual tokens and does not necessarily equal the reduction in output sequence length. The treatment of unselected tokens for each host is specified in Appendix B.

Thus, state-level risk determines *how many* visual tokens are retained, while disclosure safety and navigation evidence jointly determine *which* tokens are selected, without retraining the underlying VLN policy.

4 Experiments

4.1 Experimental Setup

VLN platforms and navigation. We evaluate PrivacyNav on StreamVLN, Uni-NaVid, OmniNav, and ActiveVLN Wei et al. (2026); Zhang et al. (2025); Xue et al. (2026); Zhang et al. (2026). The host policies remain frozen. We evaluate each method on the 1,839 episodes of R2R val-unseen and report success rate (SR) and success weighted by path length (SPL) Anderson et al. (2018).

Semantic recovery. We evaluate a frozen caption recovery model on the 5,000 images of COCO val2017 Lin et al. (2014) and report CIDEr Vedantam et al. (2015), SPICE Anderson et al. (2016), and METEOR Banerjee & Lavie (2005). We additionally audit recovered captions against eight policy-defined concepts: *person*, *bed*, *toilet*, *laptop*, *tv*, *cell phone*, *book*, and *keyboard*. Macro recovery averages concept recall over classes with positive ground truth; micro recovery pools their true positives and false negatives. Any-Sens. measures the fraction of positive images for which the recovered caption mentions at least one ground-truth sensitive concept. Lower scores indicate less semantic recoverability under the evaluated protocol, not a formal privacy guarantee. We also evaluate 973 navigation states using host-specific semantic queries, reporting true positive rate (TPR) on sensitive states and false positive rate (FPR) on ordinary states. Appendix C details both recovery protocols and Table 6 reports the in-domain host results.

Baselines. Raw releases the original visual representation. VisionZip compresses visual tokens without an explicit disclosure objective Yang et al. (2025). Gaussian adds noise to the visual representation. For each host, the same policy and method configuration are used for navigation and semantic-recovery evaluation. The per-host release and noise settings are listed in Table 7 of Appendix D.

4.2 Main Results

Table 1 compares navigation utility and semantic recoverability across four frozen hosts.

Sensitive recovery and navigation. On StreamVLN, PrivacyNav lowers macro recovery from 50.44% to 40.06% and micro recovery from 64.79% to 53.09%, while SR decreases from 0.5813

Table 1: Disclosure–utility comparison on four VLN hosts. Navigation is evaluated on R2R val-unseen and semantic recovery on COCO-5000. Macro and Micro denote sensitive-concept recovery; Any-Sens. denotes positive-image recovery of at least one ground-truth sensitive concept. Bold and underline mark the best and second-best result within each host and metric.

Host	Method	Navigation Utility		General Semantic Recoverability			Sensitive Semantic Recoverability		
		SR↑	SPL↑	CIDEr↓	SPICE↓	METEOR↓	Macro↓	Micro↓	Any-Sens.↓
StreamVLN	Raw	**0.5813**	**0.5164**	1.1148	0.2086	0.2747	50.44%	64.79%	74.98%
	VisionZip	0.5721	0.5101	1.1084	0.2076	0.2738	50.12%	64.87%	75.11%
	Gaussian	0.4535	0.4103	**0.9855**	0.1936	**0.2509**	46.92%	63.37%	73.81%
	PrivacyNav	0.5318	0.4713	1.0064	**0.1928**	0.2554	**40.06%**	**53.09%**	**62.33%**
Uni-NaVid	Raw	**0.5307**	**0.4876**	1.0325	0.1955	0.2647	51.45%	64.18%	74.13%
	VisionZip	0.5019	0.4550	1.0283	0.1945	0.2640	52.41%	63.70%	73.30%
	Gaussian	0.0729	0.0542	**0.9614**	**0.1845**	**0.2539**	49.05%	65.38%	76.38%
	PrivacyNav	0.4796	0.4384	0.9651	0.1866	0.2555	**46.17%**	**62.12%**	**73.11%**
OmniNav	Raw	0.6862	0.6544	1.0587	0.2003	0.2647	47.69%	63.19%	74.10%
	VisionZip	**0.6933**	**0.6593**	1.0358	0.1986	0.2612	47.62%	62.83%	73.62%
	Gaussian	0.6694	0.6360	**0.9440**	0.1885	**0.2488**	46.32%	62.76%	73.68%
	PrivacyNav	0.6781	0.6453	0.9616	**0.1882**	0.2510	**44.46%**	**59.83%**	**70.58%**
ActiveVLN	Raw	**0.4834**	**0.4399**	1.0542	0.2028	0.2657	49.35%	58.41%	69.09%
	VisionZip	0.4823	0.4353	1.0375	0.1998	0.2635	49.05%	62.94%	73.68%
	Gaussian	0.4530	0.4145	0.9743	0.1940	0.2626	48.91%	62.83%	73.27%
	PrivacyNav	0.4415	0.3949	**0.9555**	**0.1890**	**0.2571**	**40.36%**	**56.75%**	**66.74%**

Table 2: StreamVLN module ablation on 1,839 navigation episodes and COCO-5000. Retention is the average fraction of visual tokens selected. Macro and Micro denote sensitive-concept recovery. Bold and underline mark the best and second-best value in each metric.

Variant	Retention	SR↑	SPL↑	CIDEr↓	SPICE↓	Macro↓	Micro↓
PrivacyNav	75.27%	0.5318	0.4713	1.0064	0.1928	40.06%	53.09%
w/o disclosure-aware selection	75.31%	**0.5710**	**0.5072**	1.1108	0.2084	50.50%	65.02%
w/o state-adaptive retention	75.51%	0.5073	0.4553	**0.9934**	**0.1921**	**39.62%**	**52.15%**

to 0.5318. Macro recovery decreases by 3.23–8.99 percentage points on the other three hosts, with SR losses of 0.81–5.11 percentage points. Within each host, PrivacyNav has the lowest macro and micro recovery among the compared methods.

Comparison with release baselines. On StreamVLN, VisionZip changes macro recovery only from 50.44% to 50.12%, whereas PrivacyNav reaches 40.06%. Relative to Gaussian, PrivacyNav improves both SR (0.5318 vs. 0.4535) and macro recovery (40.06% vs. 46.92%), although Gaussian has the lower CIDEr (0.9855 vs. 1.0064). The navigation and macro-recovery advantages over Gaussian also hold on Uni-NaVid and OmniNav. On ActiveVLN, PrivacyNav obtains a lower macro recovery (40.36% vs. 48.91%) with a lower SR (0.4415 vs. 0.4530).

4.3 Analysis and Ablation

Module ablation. We evaluate three StreamVLN variants on all 1,839 navigation episodes and COCO-5000 at similar average retention ratios. The first ablation removes disclosure safety from token ranking while retaining the PrivacyNav release controller; the second uses a fixed token retention ratio while retaining disclosure-aware selection.

Table 2 shows that removing disclosure-aware selection raises macro recovery from 40.06% to 50.50% and SR from 0.5318 to 0.5710 at similar mean retention. Thus, disclosure-aware selection lowers sensitive recovery at a navigation cost. Relative to fixed retention, PrivacyNav raises SR from 0.5073 to 0.5318 and SPL from 0.4553 to 0.4713, gains of 2.45 and 1.60 percentage points. Macro and micro recovery also rise slightly, from 39.62% to 40.06% and from 52.15% to 53.09%, respectively. At this operating point, state adaptation improves navigation utility while allowing a small increase in sensitive-concept recovery. To probe risk alignment separately, a 200-episode study compares fixed, shuffled, and risk-adaptive budgets: their SR values are 0.545, 0.535, and 0.595, but the paired intervals cross zero (Appendix E).

Table 3: In-domain semantic recovery on 973 VLN states. TPR measures sensitive-state recovery and FPR false recovery on ordinary states. Lower values indicate less sensitive-state recovery or fewer false recoveries under the fixed host-specific query protocol. Bold marks the better result for each reported metric within a host.

	TPR↓		FPR↓	
Host	Raw	PrivacyNav	Raw	PrivacyNav
StreamVLN	70.0%	**60.3%**	15.2%	**12.5%**
Uni-NaVid	78.7%	**54.3%**	8.1%	**5.7%**
OmniNav	88.8%	**43.8%**	**23.1%**	23.8%
ActiveVLN	86.5%	**35.6%**	19.5%	**5.5%**

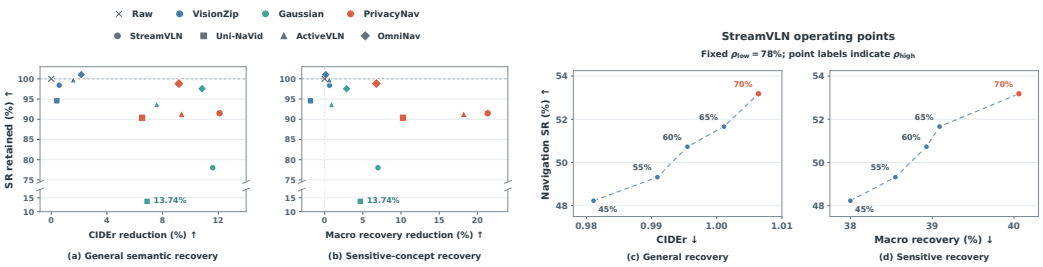

Figure 3: Disclosure–utility analysis. (a,b) Cross-host comparison of relative CIDEr and macro-recovery reductions against SR retention, all relative to Raw. The vertical axes omit 18–74.5%. (c,d) StreamVLN operating curves for CIDEr and macro recovery; only ρ_{high} changes, while $\rho_{\text{low}} = 78\%$. The 70% endpoint is the PrivacyNav result in Table 1.

In-domain semantic recovery. We evaluate 973 R2R val-unseen states with no scene overlap with the estimator training or calibration data. Sensitive-state recovery is measured by TPR and false recovery on ordinary states by FPR using host-specific query protocols.

Table 3 shows TPR decreases of 9.7, 24.4, 44.9, and 50.9 percentage points for StreamVLN, Uni-NaVid, OmniNav, and ActiveVLN. FPR decreases on three hosts but increases from 23.1% to 23.8% on OmniNav. These fixed-query results support within-host comparisons only; they do not by themselves establish selective suppression or robustness to a retrained attacker.

Disclosure–utility operating range. We vary only the high-risk retention ratio $\rho_{\text{high}} \in \{45\%, 55\%, 60\%, 65\%, 70\%\}$ on StreamVLN, holding the low-risk ratio at 78% and the other settings fixed. Across 1,839 episodes and COCO-5000, SR rises from 0.4823 to 0.5318, CIDEr from 0.9811 to 1.0064, and macro recovery from 38.00% to 40.06% (Figure 3(c,d)). Increasing the high-risk retention budget therefore improves navigation success while increasing recoverable general and sensitive semantics under these protocols. All five measured points are listed in Table 8.

4.4 Real-world Deployment

We deploy PrivacyNav on a SCOUT MINI with a RealSense D435i camera, Livox Mid-360 LiDAR, and Jetson AGX Orin. Figure 4 illustrates the navigation tasks; Appendix F details the platform and deployment. The sensitive audit uses seven observations with ten descriptions each; physical navigation evaluates 80 short- and 40 long-horizon episodes per method.

Table 4 shows fewer sensitive mentions with PrivacyNav than Raw (57.1% vs. 85.7%), at lower navigation success. Compared with Gaussian, PrivacyNav reduces sensitive mentions and improves success in both settings. Appendix F separately reports a pipeline replay on 122 real-camera observations.

Table 4: Real-world disclosure and navigation. Sensitive mentions are measured over 70 generated descriptions per method from seven source observations. Short- and long-horizon navigation results aggregate 80 and 40 physical episodes per method, respectively; parentheses give success rates.

Method	Sensitive Mentions↓	Short Horizon↑	Long Horizon↑
Raw	85.7%	**74/80 (92.5%)**	**24/40 (60.0%)**
VisionZip	85.7%	70/80 (87.5%)	20/40 (50.0%)
Gaussian	82.9%	55/80 (68.8%)	15/40 (37.5%)
PrivacyNav	**57.1%**	68/80 (85.0%)	18/40 (45.0%)

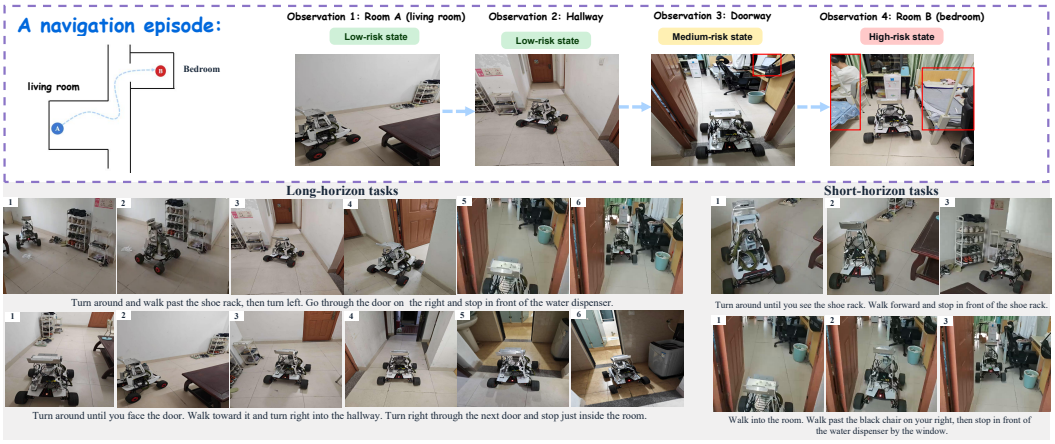

Figure 4: Real-world navigation with PrivacyNav. Top: Representative observations along a navigation episode illustrate changing disclosure risk; highlighted regions show policy-defined sensitive content. Bottom: Long-horizon (left) and short-horizon (right) navigation sequences with their instructions. Numbers indicate frame order.

5 Limitations

Our results measure recovery under fixed protocols, without formal privacy guarantees or validation against adaptive attackers. Effectiveness depends on ontology coverage, weak-supervision quality, and transfer across environments. Incomplete training labels and independently human-verified evaluation annotations are detailed in Appendices A and C. Inference from token counts and selection patterns remains unevaluated.

6 Conclusion

We presented PrivacyNav, a visual disclosure control framework that determines which visual evidence to disclose and how much to retain at each navigation state. Its Privacy-Aware Evidence Estimator and Adaptive Visual Disclosure Controller combine disclosure-aware selection with state-adaptive retention while keeping the underlying VLN policy frozen.

Experiments across four heterogeneous VLN hosts, complemented by module ablations and real-world deployment, show reduced sensitive-semantic recovery under the evaluated protocols, with host-dependent navigation costs. In the StreamVLN ablation, disclosure-aware selection primarily drives the recovery reduction, while adaptive retention improves navigation with a small increase in recovery at the evaluated operating point. The budget sweep further demonstrates an adjustable disclosure–utility trade-off without policy retraining. These findings support selective visual disclosure as an empirical approach to managing semantic exposure in embodied navigation.

AI USE STATEMENT

In this work, we used generative AI tools to assist with scientific writing and language editing, as well as literature retrieval and discovery. AI-assisted weak labeling and evaluation pre-annotation are described in Appendices A and C; all candidate annotations for the in-domain evaluation cohort were manually verified. We did not use generative AI tools to prove mathematical claims. All AI-assisted writing was reviewed and revised by the authors, and relevant literature was checked against the original sources. We take full responsibility for the final content of this work, including its text, claims, and results.

REFERENCES

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pp. 382–398, 2016.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Nikos Athanasiou, Ilya A. Petrov, Angela Yao, Shugao Ma, Eric Sauser, Edoardo Remelli, Shreyas Hampali, Johannes Schönberger, Fadime Sener, and Bugra Tekin. Trustclip: Learning private visual features via adversarial reconstruction. In *European Conference on Computer Vision*, 2026.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, 2005.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *International Conference on Learning Representations*, 2023.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, 2024.

Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. 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*, 2022.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the navgraph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, 2020.

Jingtao Li, Adnan Siraj Rakin, Xing Chen, Zhezhi He, Deliang Fan, and Chaitali Chakrabarti. Ressfl: A resistance transfer framework for defending model inversion attack in split federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10194–10202, 2022.

Ming Li, Xiangyu Xu, Hehe Fan, Pan Zhou, Jun Liu, Jia-Wei Liu, Jiahe Li, Jussi Keppo, Mike Zheng Shou, and Shuicheng Yan. Stprivacy: Spatio-temporal privacy-preserving action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5106–5115, 2023.

Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. EViT: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations*, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Hui Lu, Mia Chiquier, and Carl Vondrick. Private multiparty perception for navigation. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Jingwei Sun, Ang Li, Binghui Wang, Huanrui Yang, Hai Li, and Yiran Chen. Soteria: Provable defense against privacy leakage in federated learning from representation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9311–9319, 2021.

Batuhan Tömekçe, Mark Vero, Robin Staab, and Martin Vechev. Private attribute inference from images with vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4566–4575, 2015.

Praneeth Vepakomma, Abhishek Singh, Otkrist Gupta, and Ramesh Raskar. NoPeek: Information leakage reduction to share activations in distributed deep learning. *arXiv preprint arXiv:2008.09161*, 2020.

Meng Wei, Chenyang Wan, Xiqian Yu, Tai Wang, Yuqiang Yang, Xiaohan Mao, Chenming Zhu, Wenzhe Cai, Hanqing Wang, Yilun Chen, Xihui Liu, and Jiangmiao Pang. Streamvln: Streaming vision-and-language navigation via slowfast context modeling. In *IEEE International Conference on Robotics and Automation*, 2026.

Kedong Xiu and Sai Qian Zhang. Caprecover: A cross-modality feature inversion attack framework on vision language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3808–3816, 2025. doi: 10.1145/3746027.3755203.

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

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19792–19802, 2025. doi: 10.1109/CVPR52734.2025.01843.

Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. In *Proceedings of Robotics: Science and Systems*, 2025. doi: 10.15607/RSS.2025.XXI.013.

Zekai Zhang, Weiye Zhu, Hewei Pan, Xiangchen Wang, Rongtao Xu, Xing Sun, and Feng Zheng. Activevln: Towards active exploration via multi-turn rl in vision-and-language navigation. In *IEEE International Conference on Robotics and Automation*, 2026.

A Implementation Details

Disclosure policy and weak supervision. The Privacy-Aware Evidence Estimator is developed using 6,000 RGB observations from 200 R2R-train rollout episodes. A scene-disjoint partition assigns 4,890 observations from 163 episodes and 44 MP3D scenes to training and 1,110 observations from 37 episodes and 11 scenes to risk-CDF calibration. The model does not train on COCO captions or the in-domain evaluation cohort. We define private-room context as bedroom or bathroom. A frozen Qwen2.5-VL room query, greedily decoded into a closed room vocabulary, supplies automatic weak labels for the frame head. YOLO-Master proxy boxes at confidence 0.25 and 640-pixel input provide token targets for a declared sensitive-object policy: person, bed, toilet, bathtub, shower, sink, bathroom vanity, TV/screen, laptop, cell phone, book/document, keyboard, and monitor. Only seven box classes—bed, person, TV, toilet, sink, book, and cell phone—appear as accepted detections in this corpus. These automatically generated targets provide policy-defined weak supervision; they are not human-verified privacy annotations or ground-truth Matterport room polygons. Unknown room labels are masked from the frame-level supervision loss, while token-level supervision follows the masking rules described below. The weak labels used for estimator training are never reused as evaluation targets. In-domain evaluation annotations are constructed separately through automated pre-annotation followed by manual verification against database records.

Sensitive boxes are projected onto each host's actual visual grid. For StreamVLN, box occupancy is rasterized on the 27×27 source grid and bilinearly mapped to the 14×14 pooled grid. A pooled token with positive box support is a sensitive token. The `mil_ignore` rule treats a private-room label as a positive *frame bag*: box-covered tokens receive localized supervision, while non-box tokens in such a frame are excluded from token loss. A room-only positive contributes frame loss but no token loss. This prevents assigning sensitivity to every background token merely because the frame depicts a private room.

Estimator architecture and optimization. For the Qwen-domain representation, the token head is $\mathrm{LN}(D) \to \mathrm{Linear}(D, 256) \to \mathrm{GELU} \to \mathrm{Dropout}(0.1) \to \mathrm{Linear}(256, 2)$. An independent attention-pooling frame head predicts binary private-room context. The token keep target is one minus the token privacy target. Training combines masked weighted token BCE (sensitive weight 4.0, clean weight 0.5) and frame cross-entropy (weight 1.0), with unknown room labels excluded from the latter. We use balanced bedroom/bathroom/object/clean sampling, a learning rate of 5×10^{-5}, seed 42, and at most 1,000 steps. Host VLN backbones stay frozen. StreamVLN uses the same label semantics but trains a separate dual-head estimator on its 196×3584 post-vision tokens.

Risk calibration and budget. Let $s_{t,i}$ be the token keep probability and $d_{t,i} = 1 - s_{t,i}$ its disclosure-risk score. StreamVLN pools the largest $\lceil 0.10 N_t \rceil$ token risks to obtain a token-level state score. Token and frame-head scores are each converted to percentiles by their own empirical CDFs constructed without gradient updates on the calibration partition. For the reported PrivacyNav configuration on StreamVLN, the token/room percentile-fusion weights are 0.9/0.1, the risk-percentile trigger is 0.8, and a high-risk decision is held for two decision steps. The high/low-risk disclosure ratios are 0.70/0.78; the trajectory decay is zero. A fixed rank-fusion coefficient of 0.8 weights token disclosure safety, leaving 0.2 for host-native navigation evidence. The controller decides the number of selected tokens, while this fixed ranking decides which tokens receive that budget. These numerical constants describe PrivacyNav on StreamVLN. Across all four hosts, PrivacyNav uses the same two-module configuration: disclosure-aware evidence estimation and state-adaptive selective release, without additive feature noise. Host-specific interfaces and the reported release ratios are described below.

Rank normalization and tie handling. In StreamVLN and the Qwen fusion path, a score vector $x \in \mathbb{R}^n$ with $n > 1$ is normalized by stable ascending sorting:

$$\pi = \mathrm{argsort}_{\mathrm{ascending,stable}}(x), \qquad \hat{x}_{\pi_j} = \frac{j}{n-1}, \quad j = 0, \ldots, n-1.$$

Higher scores receive higher ranks. For equal input scores, larger original indices receive higher normalized ranks. Thus, a constant score vector receives index-dependent ranks, rather than identical normalized values. For $n = 1$, the normalized score is one. StreamVLN normalizes over its current token vector, whereas the Qwen path normalizes separately within each image segment.

Table 5: Host integration used by the reported runs. All four PrivacyNav integrations use the same two-module framework with host-specific selection and interface operations, without post-release Gaussian perturbation. Ratios refer to the visual token sets subject to selection, not necessarily the full downstream sequence. The ActiveVLN navigation ratio is the mean realized ratio over selection calls and includes current and historical visual tokens; risk is estimated from the current image only.

Host	Visual boundary	Navigation evidence	Selection/interface	Selected-token ratio
StreamVLN	Projected, pooled 196×3584 tokens	Frozen SigLIP final-block attention	Rank fusion; unselected positions filled with the selected-token mean	75.27% nav; 77.21% COCO
Uni-NaVid	EVA-domain navigation tokens; 16 groups $\times$ 4 tokens	Frozen host-native navigation evidence	Group ranking; exact-budget token pruning	81.08% nav; 84.13% COCO
OmniNav	Matched post-merger/pre-LLM boundary in the COCO audit	Host-native attention	Post-merger token pruning; retained tokens kept in original order	69.50% nav; 69.01% COCO
ActiveVLN	Qwen post-merger visual tokens	Host-native navigation evidence	Post-merger token pruning; retained tokens kept in original order	69.21% nav; 79.11% COCO

The fusion score is $q_i = 0.8\hat{s}_i + 0.2\hat{n}_i$. Selection uses `torch.topk` with `largest=True` and `sorted=False`, followed by sorting the selected indices into their original order. No explicit secondary tie-breaking rule is applied to equal fused scores at the selection boundary.

Uni-NaVid ranks 16 groups using the ascending sorting key $(\text{score}, -\text{group_id})$ and divides each resulting rank position by 15. Smaller group IDs therefore receive higher ranks when input scores are equal. Let T_g denote the host-native navigation-evidence score and I_g the estimator-derived disclosure-risk score for group g, with higher I_g indicating greater estimated sensitivity. The implementation computes $q_g = \hat{T}_g - 2\hat{I}_g$, favoring stronger navigation evidence and lower estimated disclosure risk. Equal group-selection scores favor smaller group IDs. Within each group, tokens are selected in ascending original index order until the budget is filled, allowing partial selection of the boundary group.

Weak-supervision coverage. Of the 6,000 room-query outputs in the estimator-development corpus, 1,177 contain an invalid repeated negative-number pattern and are assigned the room label `unknown` by the frozen parsing pipeline. These observations remain in their assigned data partitions. During training, unknown room labels are conservatively masked from the frame-level supervision loss rather than treated as negative targets; token-level supervision remains governed by the rules described above. This count characterizes incomplete room-level supervision coverage, not a measured label-error rate. This limitation concerns automatic weak supervision for estimator development and is separate from the independently constructed, human-verified annotations used for in-domain evaluation.

B Integration with VLN Hosts

The Adaptive Visual Disclosure Controller runs at a host-native post-vision/pre-reasoning boundary. Ranking for selection does not reorder the retained visual units; selected indices are restored to their native spatial order. Table 5 summarizes the host-specific visual boundaries, navigation evidence, and token transport.

StreamVLN. In the simulator benchmark, unselected positions are filled by the mean of selected tokens from the same frame before the frozen host continues. The downstream tensor therefore remains 196×3584; "selected K tokens" does not imply a length-K reasoning input. The real-

world split packet serializes only selected BF16 tokens and their indices before receiver-side mean-fill reconstruction.

Uni-NaVid. Uni-NaVid follows the same disclosure-estimation and adaptive-retention framework as the other hosts, with a host-specific group-based selection interface. PrivacyNav ranks 16 groups of four navigation tokens using navigation evidence and estimator-derived disclosure risk, allowing partial selection of the boundary group. Unselected tokens are physically removed, changing the controlled token tensor from $[1, 64, 4096]$ to $[1, K, 4096]$. Retained tokens preserve their original order. The adapter adds no explicit original spatial position IDs and makes no separate modification to the LLM position-encoding scheme. The host adjusts attention masks to the resulting sequence lengths and masks any required batch padding.

OmniNav and ActiveVLN. For ActiveVLN, state-risk estimation uses only the current image, while token selection operates on the current and historical visual tokens supplied in the same call. The Qwen fusion path physically removes unselected post-merger visual tokens, without contextual merging or mean filling. Retained tokens preserve their original order. Corresponding visual positions in `input_ids`, `attention_mask`, and `position_ids` are removed consistently. If position IDs are not already available, they are computed from the original input and visual grid before selection. Retained tokens are therefore not reassigned consecutive spatial coordinates. The original `image_grid_thw` is retained for interface compatibility, and the subsequent forward call uses `pixel_values=None` to avoid image re-encoding.

Execution and capture interfaces. The Uni-NaVid, OmniNav, and ActiveVLN simulator integrations use in-process tensor adaptation, rather than a shared cross-device packet protocol. COCO captures store released feature tensors and separately record token counts and retained indices for auditing. The presence of metadata in audit records does not imply that it is supplied to the recovery model. The selected-BF16-token packet and receiver-side mean-fill reconstruction are specific to the StreamVLN real-world split implementation. The three pruning paths above have no mean-fill reconstruction step.

C Semantic Recovery Protocols

General caption recovery. We use the frozen CapRecover/BLIP-2 evaluator on host-specific released visual representations from COCO-val2017. Evaluation uses 5,000 images, seed 42, ten beams, sampling enabled, top-$k = 10$, top-$p = 0.95$, temperature 1.0, maximum length 50, and COCOEvalCap with PTB tokenization. Attacker checkpoint and input adapter are fixed within each host-method comparison. CIDEr and SPICE summarize general recoverability of image-caption semantics; they are neither sensitive-concept classifiers nor privacy guarantees. The OmniNav Raw and defended conditions use the same post-merger/pre-LLM visual boundary.

Sensitive-concept scoring. The frozen COCO policy contains eight instance categories: *person, bed, toilet, laptop, TV, cell phone, book, keyboard*. Recovered captions are scored using a frozen synonym table and thresholded lexical matching. This scoring procedure determines which policy concepts are named in a recovered caption; it does not generate the reference labels used for evaluation. For each category c with ground-truth positives, $\mathrm{Recall}_c = TP_c/(TP_c + FN_c)$ and $\mathrm{Macro} = |\mathcal{C}_+|^{-1} \sum_{c \in \mathcal{C}_+} \mathrm{Recall}_c$, where $\mathcal{C}_+$ is the set of policy categories with ground-truth positives. Micro recall pools TP_c and FN_c across categories. Any-Sensitive Recovery is computed on images containing at least one ground-truth policy concept. An image is counted as recovered only if the recovered caption names a concept present in that image's ground-truth concept set. The same eight-concept policy and synonym matcher are used for all four hosts. Ordinary-image FPR is the fraction of ground-truth-negative images whose recovered captions name any policy concept. The COCO subset contains 3,154 positive images and 1,846 ordinary controls. For ActiveVLN, the reported navigation SR and SPL are 0.4415 and 0.3949; CIDEr, SPICE, and METEOR are 0.9555, 0.1890, and 0.2571. Macro/Micro/Any-Sensitive recovery is 40.36/56.75/66.74%, respectively. PrivacyNav uses selective release without additive feature noise.

In-domain evaluation annotation protocol. The in-domain cohort contains 973 VLN observations from ten val-unseen scenes: 267 policy-positive states (bedroom/bathroom) and 706 ordinary

Table 6: Complementary in-domain room-query evaluation on 973 states with machine-assisted, human-verified annotations: 267 policy-positive states and 706 ordinary controls. TPR and FPR compare Raw and PrivacyNav within each host.

Host	Method	Sensitive TPR ↓	Ordinary FPR ↓
StreamVLN	Raw	70.0%	15.2%
	PrivacyNav	60.3%	12.5%
Uni-NaVid	Raw	78.7%	8.1%
	PrivacyNav	54.3%	5.7%
OmniNav	Raw	88.8%	23.1%
	PrivacyNav	43.8%	23.8%
ActiveVLN	Raw	86.5%	19.5%
	PrivacyNav	35.6%	5.5%

controls. Evaluation annotations are constructed independently of the weak supervision used for estimator training. An automated DeepSeek–Codex pipeline first generates candidate annotations. Each candidate is then manually cross-checked against the corresponding database records, with inconsistent or ambiguous cases resolved through human correction and adjudication. This verification covers all 973 observations, including both policy-positive states and ordinary controls. The final evaluation targets are therefore *machine-assisted, human-verified annotations*, rather than raw model-generated pseudo-labels. The weak labels used to train the disclosure estimator are never reused as evaluation targets.

In-domain room query. Each host uses a fixed room-query prompt, decoding procedure, representation adapter, and answer mapping. TPR is the fraction of human-verified policy-positive states classified as sensitive; FPR is the fraction of human-verified ordinary controls classified as sensitive. Comparisons are paired *within* each host, while absolute rates across hosts reflect different attacker interfaces. The Uni-NaVid room-query evaluation applies its navigation-token selector to all 973 observations. Under the host-specific query protocol reported in the main paper, OmniNav sensitive-state TPR changes from 88.8% to 43.8%, while ordinary-state FPR changes from 23.1% to 23.8%. We report the lower TPR together with the slight FPR increase.

For the paired TPR difference (PrivacyNav minus Raw), the scene-clustered 95% intervals are $[-16.9, -3.4]$ percentage points for StreamVLN and $[-38.8, -11.0]$ percentage points for Uni-NaVid. These estimates apply to the 973-state cohort with human-verified evaluation annotations described above; they do not establish a host-independent attacker benchmark.

C.1 Paired COCO Recovery Examples

The following captions are paired Raw and PrivacyNav outputs for the same StreamVLN COCO-5000 images. They illustrate individual recoveries rather than the frequency of these outcomes.

COCO image 000000564336.

> **Raw:** A man sitting at a table in front of a laptop computer.
> **PrivacyNav:** A wooden table and chairs inside of a building.

COCO image 000000000139.

> **Raw:** A woman standing in a living room next to a tv.
> **PrivacyNav:** A living room filled with wooden furniture next to a window.

In these two examples, the PrivacyNav captions omit the person and device terms present in the Raw captions while retaining descriptions of furniture and room context.

D Baseline and Release Configurations

Table 7 reports the method identity used in the main navigation/caption table. Release ratios are realized over the indicated navigation trajectory or COCO image set when available; configured values

Table 7: Visual-token selection ratios associated with the main-table methods. Navigation and COCO ratios refer to different observation sets; "cfg." identifies a configured rather than measured value. Raw and Gaussian retain all visual tokens, although Gaussian perturbs their feature values. For ActiveVLN, the PrivacyNav navigation ratio is the mean realized ratio over 42,187 selection calls covering 1,839 episodes. Selection includes current and historical visual tokens, while risk is estimated from the current image only. Selected-token ratios do not necessarily equal downstream sequence-length or transmitted-byte reductions.

Host	Method	Navigation selected (%)	COCO selected (%)	Setting
StreamVLN	Raw	100	100	Original representation
	VisionZip	75.00	75.00	Token compression
	Gaussian	100	100	$\sigma = 1.4$
	PrivacyNav	75.27	77.21	Adaptive selection
Uni-NaVid	Raw	100	100	Original representation
	VisionZip	79.69	79.69	Token compression
	Gaussian	100	100	$\sigma = 1.0$
	PrivacyNav	81.08	84.13	Adaptive selection
OmniNav	Raw	100	100	Original representation
	VisionZip	69.01 (cfg.)	69.00	Token compression
	Gaussian	100	100	$\sigma = 1.1$
	PrivacyNav	69.50	69.01	Adaptive selection
ActiveVLN	Raw	100	100	Original representation
	VisionZip	69.21 (cfg.)	69.50	Token compression
	Gaussian	100	100	$\sigma = 0.5$
	PrivacyNav	69.21	79.11	Risk-adaptive selection; budget 30–100%

are marked explicitly. Because the controller responds to the input-risk distribution, its navigation and COCO means need not match. For the independent Gaussian baseline only, noise is applied to feature values while retaining all visual units; sigma denotes the recorded relative feature-noise setting. PrivacyNav itself does not add Gaussian noise on any of the four hosts, including ActiveVLN.

ActiveVLN selection-ratio accounting. For ActiveVLN, "cfg." denotes the configured VisionZip navigation ratio. The PrivacyNav navigation ratio is measured and averaged equally over $M = 42{,}187$ selection calls covering all 1,839 evaluation episodes:

$$\bar{\rho}_{\text{call}} = \frac{1}{M} \sum_{j=1}^{M} \frac{K_j}{N_j} \approx 69.21\%,$$

where N_j and K_j are the input and selected visual-token counts in call j. Each call includes current and historical visual tokens supplied to the selector, while state risk is estimated from the current image only. The token-weighted ratio is

$$\rho_{\text{token}} = \frac{\sum_j K_j}{\sum_j N_j} = \frac{49{,}327{,}754}{72{,}916{,}360} \approx 67.65\%.$$

These totals count token occurrences across calls, including repeated appearances of historical tokens, rather than unique image tokens. The COCO ratios are measured on the separate image evaluation set.

Gaussian is an independent baseline. On ActiveVLN, it obtains SR 0.4530, SPL 0.4145, CIDEr 0.9743, SPICE 0.1940, and METEOR 0.2626. The Uni-NaVid $\sigma = 1.0$ Gaussian run has three numerical failures handled by a fail-closed protocol; its SR is 0.0729.

The navigation selected-token fractions for VisionZip and PrivacyNav are close on StreamVLN (75.00% vs. 75.27%), Uni-NaVid (79.69% vs. 81.08%), and OmniNav (69.01% configured vs. 69.50%). Under the frozen StreamVLN main-table evaluation, VisionZip and PrivacyNav obtain

Table 8: Internally matched StreamVLN operating points; each row uses 1,839 episodes and 5,000 COCO images in the operating-curve capture. The 70% high-risk budget corresponds to the PrivacyNav configuration reported in Table 1.

High-risk budget	SR ↑	SPL ↑	CIDEr ↓	Macro ↓
45%	0.4823	0.4238	0.9811	38.00%
55%	0.4932	0.4351	0.9909	38.55%
60%	0.5073	0.4491	0.9955	38.93%
65%	0.5166	0.4588	1.0011	39.09%
70%	0.5318	0.4713	1.0064	40.06%

SR 0.5721 and 0.5318, CIDEr 1.1084 and 1.0064, and macro recovery 50.12% and 40.06%, respectively. Thus the nearby navigation-token budgets accompany different measured disclosure–utility outcomes. COCO selection fractions differ from navigation fractions and, on Uni-NaVid, from each other by 4.44 percentage points (VisionZip 79.69%, PrivacyNav 84.13%). The reported points do not establish an equal-budget cross-method Pareto frontier. The five-point curve in Table 8 varies only PrivacyNav's StreamVLN high-risk budget.

PrivacyNav configuration on StreamVLN. PrivacyNav on StreamVLN uses top-10% token-risk pooling, separate token and room CDF calibration, percentile-fusion weights of 0.9/0.1, an upper-risk trigger at 0.8, high/low release ratios of 0.70/0.78, a two-decision-step high-risk hold, and fixed safety/navigation rank-fusion weights of 0.8/0.2. Its main-table, ablation, and operating-point results use CIDEr 1.0064, with SPICE 0.1928 and macro sensitive-concept recovery 40.06%. StreamVLN caption recovery uses one process with sampling enabled.

E Additional Analyses

Full-scale module ablations. On StreamVLN, removing disclosure-aware selection retains the PrivacyNav dynamic budget but ranks tokens only by host navigation evidence. Its mean navigation release is 75.31%; across 1,839 episodes it reaches SR 0.5710 and SPL 0.5072, while COCO-5000 CIDEr is 1.1108 versus 1.0064 for PrivacyNav. Removing state-adaptive budgeting retains the disclosure-aware selector but fixes the budget at 148/196 tokens (75.51%). Across 1,839 episodes, this variant reaches SR 0.5073 and SPL 0.4553, compared with 0.5318 and 0.4713 for the full PrivacyNav method. The corresponding gains from state-adaptive budgeting are 2.45 and 1.60 percentage points. The fixed-budget variant has CIDEr 0.9934, SPICE 0.1921, macro recovery 39.62%, and micro recovery 52.15%; the full PrivacyNav method has 1.0064, 0.1928, 40.06%, and 53.09%, respectively. Thus, at this operating point, state adaptation improves navigation utility while allowing slightly greater semantic recovery.

State-risk alignment. A separate 200-episode controlled analysis holds the selector fixed and nearly matches navigation release (Fixed 75.51%, Shuffled-Dynamic 75.26%, Risk-Dynamic 74.99%). SR point estimates are 0.545, 0.535, and 0.595. The paired 95% intervals for Risk-Dynamic minus Fixed and Risk-Dynamic minus Shuffled are $[-1.0, +11.0]$ and $[-0.5, +12.5]$ points, respectively. COCO-500 CIDEr in this independent-image protocol is 1.0378, 1.0235, and 1.0342, respectively; Risk-Dynamic is not the lowest-leakage point in this particular comparison. The experiment supports a risk-alignment trend in navigation point estimates but no significant superiority claim.

Budget operating points. With the estimator, CDFs, selector, backbone, risk partition, and low-risk budget 0.78 fixed, we change only the StreamVLN high-risk budget. Table 8 reports all five evaluated full-scale operating points.

The monotone point estimates show a controllable disclosure–utility operating range. We refer to this as an empirical operating curve, not a theoretically optimized Pareto frontier. We make no claim of robustness to a defense-aware retrained attacker, which was not run.

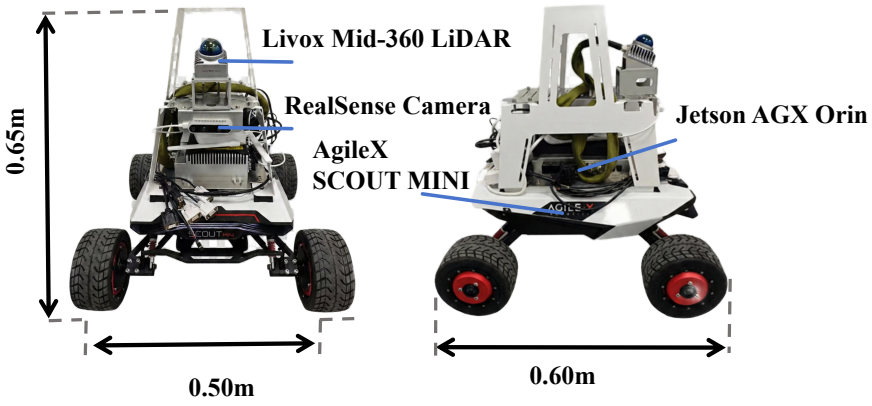

Figure 5: Physical robot platform used for real-world deployment. The SCOUT MINI carries a RealSense D435i camera, a Livox Mid-360 LiDAR, and a Jetson AGX Orin.

F Real-world Deployment

The hardware stack comprises a SCOUT MINI mobile base, an Intel RealSense D435i RGB-D camera, a Livox Mid-360 LiDAR, an NVIDIA Jetson AGX Orin (32 GB), and ROS 2 Humble. The real-world StreamVLN checkpoint's discrete STOP, 0.25-m forward, and 15° left/right actions are translated into sequential `/cmd_vel` commands with odometry feedback. RGB is the VLN model's visual input; the camera also provides depth, and LiDAR is mounted on the robot. The real-world checkpoint differs from the R2R benchmark checkpoint. Figure 5 shows the deployed platform.

As reported in Table 4, physical navigation evaluates 80 short-horizon and 40 long-horizon episodes per method. Raw completes 74/80 and 24/40 episodes; VisionZip completes 70/80 and 20/40; Gaussian completes 55/80 and 15/40; and PrivacyNav completes 68/80 and 18/40. We report these episode-level outcomes separately from the caption audit and offline replay.

The sensitive-description audit uses seven distinct observation sources, ten repeated descriptions per source, and four release methods, for 70 descriptions per method. The observed sensitive-mention rates are 85.7%, 85.7%, 82.9%, and 57.1% for Raw, VisionZip, Gaussian, and PrivacyNav, respectively. Repeated descriptions of one source are not independent scenes; these lexical mentions should not be interpreted as ground-truth-adjudicated attack TPR.

Separately, PrivacyNav was replayed offline on 122 real camera observations from 23 logged sessions. All 122 representations were processed; current-frame risk triggered in 59 observations, and the hold continued the high-risk budget for 23 additional observations. Mean realized token retention was 72.92%; mean visual encoding and selection-plus-packet times were 142.81 and 18.00 ms. This replay checks the representation pipeline and is distinct from the physical navigation success counts in Table 4.

F.1 Paired Real-world Recovery Examples

The following pairs use the same anonymized robot observation under Raw and PrivacyNav release.

Observation sample_005.

> **Raw:** A young man using a laptop computer in a living room.
> **PrivacyNav:** A young man standing in a living room.

Observation sample_006 (failure case).

> **Raw:** A man standing in a bathroom next to a toilet.
> **PrivacyNav:** A man standing in a bathroom next to a toilet.

In sample_005, the laptop is absent from the PrivacyNav caption, but the person and room remain described. In sample_006, the sensitive content remains recoverable under both releases.