SEMANTIC COVERAGE AND TEMPORAL SELECTION FOR VISUAL TOKEN PRUNING IN VISION-LANGUAGE NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language navigation (VLN) enables embodied agents to follow natural-language instructions in real-world environments. However, dense visual tokens incur substantial computation and inference latency, hindering deployment on resource-constrained robots. Existing VLN token-pruning methods neither explicitly suppress redundancy across historical frames nor exploit instruction semantics during selection. We propose a training-free spatio-temporal visual-token compressor that captures intra-frame semantic diversity and inter-frame novelty for efficient long-horizon inference. For the current frame, low-rank feature grouping identifies complementary semantic groups, while adaptive within-group selection uses feature-space non-maximum suppression to remove redundant tokens without sacrificing semantic coverage. For historical frames, we adopt the same strategy as current frame to construct an expanded candidate pool. Then, we assess each token using current-view relevance, instruction alignment, and visual saliency. Moreover, to reduce temporal redundancy, semi-relaxed temporal alignment uses optimal transport to measure inter-frame novelty against preceding-frame group prototypes. Task-aware token selection uses the combined scores and maximal marginal relevance to iteratively select the final token set. Experiments on standard VLN benchmarks show that our method consistently outperforms competing pruning baselines across compression ratios, achieving a better navigation performance and efficiency trade-off. Real-world quadruped robot experiments further validate its effectiveness for low-latency embodied navigation.

1 INTRODUCTION

Vision-language navigation (VLN) requires an embodied agent to follow natural-language instructions while moving through a changing visual environment. R2R (Anderson et al., 2018) and RxR (Ku et al., 2020) establish challenging instruction-following benchmarks in photo-realistic indoor scenes, while recent vision-language action (VLA) extends this capability to continuous control and legged robots (Cheng et al., 2025; Wei et al., 2026). In VLN, a landmark mentioned in an instruction may no longer be visible when the agent reaches the junction where it must turn. Thus, at each decision step, the agent often needs the current camera view and earlier observations. Retaining visual history preserves such evidence, but each additional frame can contribute hundreds of visual tokens. Repeatedly processing a growing history increases computation and action latency, making long-horizon reasoning difficult to reconcile with low-latency onboard navigation.

Existing methods handle visual history in several ways. A straightforward choice is to retain a bounded set of complete observations or uniformly sample frames along the trajectory (Cheng et al., 2025; Guo et al., 2026; Wang et al., 2026b). However, every retained frame still contributes many visual tokens, even including repeated views of the same landmark. In addition, geometry-aware methods project observations into an implicit 3D structure or bird's-eye-view memory and merge spatially overlapping evidence (Wang et al., 2023; Wei et al., 2026; Yang et al., 2026). Such aggregation reduces repeated spatial content, but relies on prior knowledge (e.g., depth) and may collapse observations of different times into the same spatial representation. Another different strategy allocates memory according to recency. For instance, Uni-Navid (Zhang et al., 2025a) retains more visual detail in recent observations and compresses older ones. DecoVLN (Xin et al., 2026) instead

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.

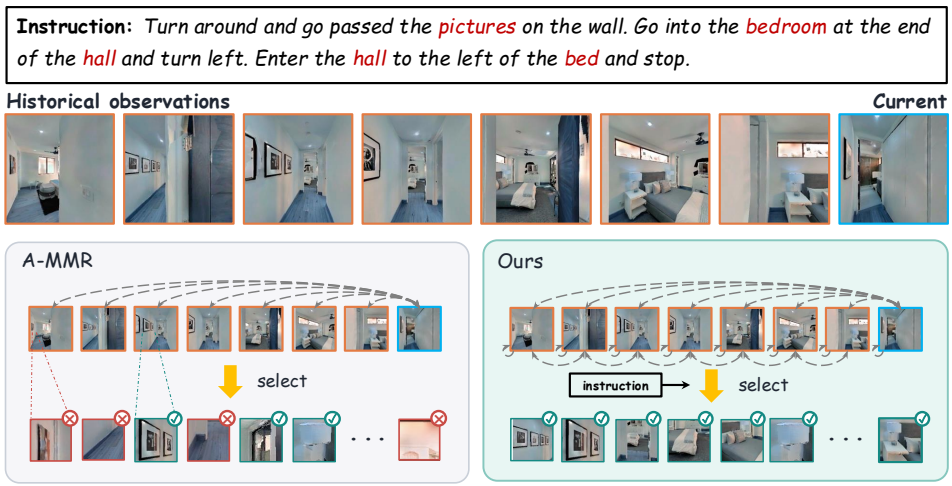


Figure 1: **Motivation**. Existing training-free VLN pruning method (Wang et al., 2026a) may miss instruction-relevant landmark tokens while retaining repeated historical content. Our method preserves landmark evidence and suppresses redundant history through intra-frame semantic coverage and inter-frame temporal selection.

refines a frame-level memory using instruction relevance, visual diversity, and temporal coverage. Although recency is a useful heuristic method, under a limited token budget, a distant but decisive landmark may receive less capacity than a recent and visually repetitive frame.

Recently, content-aware token pruning provides a more adaptive way to use the available token budget. STEP-Nav (Lu et al., 2026) removes irrelevant spatial content and temporally similar features, but restores performance through distortion-aware fine-tuning. Furthermore, Wang et al. introduce a training-free VLN pruning method that uses selected current-frame tokens as queries to reweight historical tokens (Wang et al., 2026a). Using the current view as the anchor naturally fits VLN, where the next action depends on the current observation and instruction, while history provides supporting evidence. Nevertheless, three limitations remain, as shown in Fig. 1. First, its importance measure dominated by [CLS] token may give small landmarks less weight than large and salient regions. Thus, a landmark that has left the current view may be ignored by current-view similarity. Second, scoring historical frames against the current view does not capture duplication between neighboring historical frames. For example, similar corridor patches can possess multiple tokens even when they convey nearly the same information. Third, the instruction is not explicitly used in historical token scoring, so visual relevance to the current view may not reflect relevance to the navigation goal.

These limitations call for two complementary selections. ***First***, the current observation should cover distinct semantic groups rather than devote most of its token budget to a few highly scored and mutually similar patches. ***Second***, historical observations should retain evidence useful for the instruction and current decision, while discarding content already represented in adjacent history.

We address both requirements with a training-free visual-token compressor that combines intra-frame semantic coverage with inter-frame temporal selection. It operates entirely on visual features and requires no retraining or modifying the navigation policy. For the current frame, Low-Rank Feature Grouping divides visual tokens into complementary semantic feature groups. Adaptive Within-Group Selection removes similar tokens within each group through feature-space non-maximum suppression, while Proportional Budget Allocation fairly distributes the K output slots across groups. This suppression prevents a few dominant regions from consuming most of the budget. For each historical frame, the same procedure forms an expanded pool of candidate tokens . We score each candidate token by its relevance to the current view and instruction, together with its visual saliency. These scores estimate the availability of a token for navigation goal, but cannot determine whether it repeats an earlier observation. To capture this temporal relation, Group Prototype Construction first represents each semantic feature group in the preceding historical frame with a single prototype. Semi-Relaxed Temporal Alignment then compares every candidate token in the current historical frame with these prototypes and assigns it a novelty score. Balanced OT (Cuturi,

2013) requires fixed matching amounts on both sides. This is too restrictive for VLN because objects can appear or disappear as the agent moves, so a later frame need not preserve the group distribution of the preceding frame. By contrast, semi-relaxed OT (Fukunaga & Kasai, 2022) constrains only the candidate side. Every candidate is still evaluated, but the new frame is not forced to preserve the preceding frame's group proportions, allowing the alignment to adapt as visual content changes. Tokens that closely match a prototype are treated as repeated content, whereas unmatched tokens receive higher novelty scores. Task-Aware Token Selection combines this temporal novelty with the three utility cues and selects the final K tokens while avoiding similar candidates.

Our contributions are threefold: 1) We present a training-free visual-token pruning framework for VLN that preserves diverse evidence from the current view while retaining instruction-relevant and non-redundant information from visual history. 2) We combine low-rank feature grouping, adaptive within-group selection, and proportional budget allocation to improve intra-frame semantic coverage, and introduce prototype-based semi-relaxed OT to measure temporal novelty across adjacent historical frames. 3) We evaluate navigation performance and efficiency across compression ratios on standard VLN benchmarks and deploy the method on a Unitree Go2 quadruped robot.

2 Related Work

Vision-Language Navigation. Recent vision-language action (VLA) extends vision-language navigation (VLN) from benchmark instruction following to continuous and closed-loop control. Building on R2R (Anderson et al., 2018) and RxR (Ku et al., 2020), recent methods combine streaming perception, long-horizon reasoning, and embodied action generation (Cheng et al., 2025; Wei et al., 2026; Guo et al., 2026; Wang et al., 2026b; Xin et al., 2026; Li et al., 2026). They select actions from the current observation while using visual history to recover out-of-view landmarks. Visual history is commonly represented through spatial or geometry-aware maps, compact memories, experience retrieval, or progressive compression (Wang et al., 2023; Yang et al., 2026; Zeng et al., 2026; Song et al., 2025; Xu et al., 2026; Zhang et al., 2025a). However, token-based policies must still process overlapping visual patches retained across observations, causing the input length and inference cost to grow with the trajectory. Therefore, VLN requires a compact history that removes redundancy without losing instruction-relevant evidence.

Visual Token Pruning. Visual-token pruning reduces inference cost in vision-language models by retaining representative image patches under a fixed budget (Chen et al., 2024; Zhang et al., 2025c; Alvar et al., 2025; Zhang et al., 2025b; Fang et al., 2026); video methods additionally remove cross-frame repetition (Yang et al., 2025). VLN imposes a stricter requirement: historical tokens must support both the instruction and the current decision. STEP-Nav (Lu et al., 2026) prunes spatiotemporal redundancy with distortion-aware fine-tuning, while Wang et al. (Wang et al., 2026a) provide training-free, current-view-guided pruning. However, current-view scoring may miss out-of-view landmarks and cannot directly identify repetition between historical frames. Our training-free method addresses both limitations through semantic coverage, instruction-aware scoring and adjacent-frame comparison.

3 Method

3.1 Overview

In vision-language navigation (VLN), an agent follows a textual instruction by repeatedly observing its surroundings and predicting the next action. A frozen vision encoder and projector convert each observed image into N visual patch tokens:$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top \in \mathbb{R}^{N\times d}$, where d is the feature dimension. At decision step t, $\mathbf{X}_t$ denotes the current frame, while $\mathbf{X}_h$, $1 \le h < t$, denotes a historical frame. The VLN policy processes the instruction embeddings $\mathcal{L} = \{\mathbf{e}_m\}_{m=1}^{L} \subset \mathbb{R}^d$, the current frame $\mathbf{X}_t$, and the historical frames $\{\mathbf{X}_h\}_{h=1}^{t-1}$ to predict the next action.

Our compressor is inserted after visual projection and before the navigation policy, as shown in Fig. 2. It reduces each frame from N to K original visual tokens without training the model. Low-Rank Feature Grouping first divides the tokens into groups with different visual patterns. Adaptive Within-Group Selection removes similar tokens inside each group, and Proportional Budget Allocation distributes the available slots across groups so that distinct scene content is preserved. For

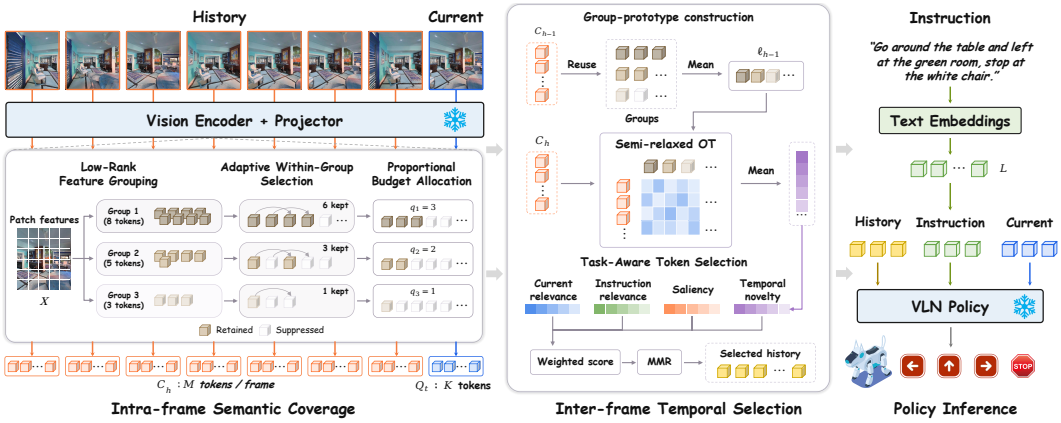


Figure 2: Overview of the proposed training-free visual-token pruning framework. Intra-frame semantic coverage groups complementary patch features, suppresses within-group redundancy, and allocates the token budget across groups. Inter-frame temporal selection aligns historical candidates with preceding-frame group prototypes and combines temporal novelty with current-view, instruction, and saliency cues before policy inference.

the current frame $\mathbf{X}_t$, these steps directly produce $\mathcal{Q}_t$ with K tokens. For each historical frame $\mathbf{X}_h$, $h < t$, they instead produce a larger candidate pool $\mathcal{C}_h$ of $M = \min(N, 2K)$ tokens, allowing less prominent but useful landmarks to survive. Group Prototype Construction summarizes the feature groups in the preceding historical frame, and Semi-Relaxed Temporal Alignment compares each candidate with these prototypes to distinguish new content from repeated observations. Task-Aware Token Selection then combines temporal novelty with relevance to the current view and instruction, as well as visual saliency, and retains the final K non-redundant tokens. Before policy inference, tokens keep their original patch order and frames remain in chronological order.

3.2 Intra-frame Semantic Coverage

Existing methods often rank all patch tokens against a single global representation (Wang et al., 2026a). This can focus the budget on large and salient regions and miss small instruction-relevant landmarks. However, VLN generally relies on multiple visual cues from different parts of a scene. Thus, we group tokens by shared semantic feature patterns rather than predefined object categories, remove redundancy only within each group, and distribute the budget across groups. This allows similar patches to compete without suppressing complementary content. For an arbitrary frame $\mathbf{X}$, let B denote its output token budget. We set $B = K$ for the current frame $\mathbf{X}_t$ and $B = M$ for each historical frame; the only difference between different frames is the output size.

Low-Rank Feature Grouping. Visual patches with similar content tend to activate similar feature dimensions, while patches from different regions vary along different dominant directions. A low-rank decomposition reveals these directions: patches with similar responses can be placed in the same group (Jolliffe & Cadima, 2016). This gives an insightful grouping of the visual tokens without additional signals. Before decomposition, we apply the element-wise sigmoid $\sigma(\cdot)$ to limit extreme activations and subtract the mean over N patches to highlight differences between patches:

$$\widehat{\mathbf{X}} = [\sigma(\mathbf{x}_i) - \frac{1}{N}\sum_{j=1}^{N}\sigma(\mathbf{x}_j)]_{i=1}^{N}. \tag{1}$$

We then apply truncated singular value decomposition (Eckart & Young, 1936) to retain R dominant directions:

$$\widehat{\mathbf{X}} \approx \mathbf{P}\mathbf{\Lambda}\mathbf{D}^{\top}. \tag{2}$$

Here, the R columns of $\mathbf{D} \in \mathbb{R}^{d\times R}$ represent latent feature directions, each describing a dominant pattern of variation among the patch features. The diagonal entries of $\mathbf{\Lambda} \in \mathbb{R}^{R\times R}$ indicate the overall strengths of these directions in the frame. Each row of $\mathbf{P} \in \mathbb{R}^{N\times R}$ records one patch's

responses to all R directions; specifically, P_{ig} is the signed response of i-th patch to direction g. Because the sign of each singular direction is arbitrary, we assign each patch to the direction with its largest absolute response:

$$g(i) = \arg \max_{1 \le g \le R} |P_{i,g}|. \tag{3}$$

Grouping patches by this assignment partitions the original N tokens into at most R non-empty groups, $G = \{G_1, \ldots, G_R\}$. Tokens within the same group share a dominant semantic pattern, whereas different groups retain complementary patterns from the frame.

Adaptive Within-Group Selection. For each non-empty group $G_r \in G$, several tokens may still describe nearly identical visual content. To remove this redundancy, we sort the tokens by their absolute responses to the assigned dominant direction. Let $(\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(|G_r|)})$ denote the tokens in G_r, sorted by decreasing response. We retain the first token and then examine the remaining tokens in order. At step k, $\mathcal{V}_r^{(k-1)}$ contains the tokens retained so far. The retained set is updated as:

$$\mathcal{V}_r^{(1)} = \{\mathbf{x}_{(1)}\},\ \mathcal{V}_r^{(k)} = \begin{cases} \mathcal{V}_r^{(k-1)} \cup \{\mathbf{x}_{(k)}\}, & \max\limits_{\mathbf{x} \in \mathcal{V}_r^{(k-1)}} \cos(\mathbf{x}_{(k)}, \mathbf{x}) \le \theta_B, \\ \mathcal{V}_r^{(k-1)}, & \text{otherwise}, \end{cases} \quad k = 2, \ldots, |G_r|, \tag{4}$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity. Thus, a token is retained only when its similarity to every previously retained token is at most θ_B. This greedy procedure is a feature-space form of non-maximum suppression: once a representative token is retained, later tokens with highly similar features are suppressed.

After all tokens have been examined, $\mathcal{V}_r = \mathcal{V}_r^{(|G_r|)}$ contains the selected representatives of group G_r. Rather than use the same cutoff for every frame and token budget, we dynamically compute:

$$\theta_B = \frac{B}{N(N-1)} \sum_{i \ne j} \cos(\mathbf{x}_i, \mathbf{x}_j). \tag{5}$$

This strategy allows a larger token budget to retain more within-group variation. Consequently, $\mathcal{V}_r$ contains representative tokens for group G_r.

Proportional Budget Allocation. After adaptive within-group selection, the token budget must be distributed across groups. An equal quota can waste slots on small groups, whereas a global top-B selection may discard groups that contain complementary information. Thus, we first reserve one slot for each group $\mathcal{V}_r$ and then allocate the remaining slots in proportion to group size.

Specifically, let n_r be the number of tokens remaining in group $\mathcal{V}_r$. If $\sum_{r=1}^{R} n_r < B$, we restore the suppressed tokens with the highest $|P_{i,g(i)}|$ scores until the total reaches B, and update the group counts accordingly. Because $R \le B$, each group can receive at least one slot. The target quota for group $\mathcal{V}_r$ is:

$$q_r = 1 + (B - R) \frac{n_r}{\sum_{j=1}^{R} n_j}. \tag{6}$$

Since q_r may be fractional, we convert the target quotas to integers using the largest-remainder method. If a group cannot fill its rounded quota, we re-assign the unfilled slots to other groups. We then select the highest $|P_{i,g(i)}|$ tokens within each group, yielding exactly B tokens. Setting $B = K$ yields the current-frame set $\mathcal{Q}_t$, while $B = M$ yields the historical pool $\mathcal{C}_h$.

3.3 Inter-frame Temporal Selection

The intra-frame stage preserves complementary content within each observation, but nearby historical frames may still repeat the same visual evidence. For a historical frame $h < t$, $\mathcal{C}_h = \{\mathbf{x}_{h,i}\}_{i=1}^{M}$ is its candidate token pool, with each candidate token retaining its intra-frame group assignment G_r. We score each candidate by its value to the current decision and its novelty relative to $\mathcal{C}_{h-1}$, then apply a diversity-aware selector to retain exactly K tokens.

Group Prototype Construction. Comparing every candidate in $\mathcal{C}_h$ with every candidate in $\mathcal{C}_{h-1}$ requires M^2 similarities. We retain only non-empty groups and let R denote their number. We then represent each group in the preceding frame $\mathbf{X}_{h-1}$ with one prototype. Let $\mathcal{C}_{h-1}^{(r)} \subseteq \mathcal{C}_{h-1}$ denote the tokens assigned to group G_r of frame $\mathbf{X}_{h-1}$. The prototype of each group G_r is the mean of its normalized candidates:

$$\boldsymbol{\ell}_{h-1,r} = \frac{1}{|\mathcal{C}_{h-1}^{(r)}|} \sum_{\mathbf{x} \in \mathcal{C}_{h-1}^{(r)}} \frac{\mathbf{x}}{\|\mathbf{x}\|_2 + \delta}. \tag{7}$$

Here $\delta > 0$ is a numerical stabilizer (here $\delta = 1e-4$). The prototype set $\boldsymbol{\ell}_{h-1} = \{\boldsymbol{\ell}_{h-1,r}\}_{r=1}^{R}$ collects the average feature directions of the R groups in frame $\mathbf{X}_{h-1}$. This reduces the alignment size from $M \times M$ to $M \times R$ without additional clustering or geometric information.

Semi-Relaxed Temporal Alignment. Inter-frame alignment distinguishes repeated content from newly observed evidence by comparing adjacent frames. Balanced optimal transport (OT) fixes the matching capacity of both sets (Cuturi, 2013). This is unsuitable for adjacent frames when objects enter or leave the view and may force new content to match unrelated prototypes. Enforcing fixed prototype capacities can therefore create artificial matches for new content. Instead, semi-relaxed OT keeps every current candidate in the alignment while allowing the prototype capacities to adapt (Fukunaga & Kasai, 2022).

Specifically, for the candidate-prototype pair $(\mathcal{C}_h, \boldsymbol{\ell}_{h-1})$ from frame pair $(\mathbf{X}_h, \mathbf{X}_{h-1})$, let $\mathbf{A} \in \mathbb{R}^{M \times R}$ be the appearance-cost matrix, with $A_{ir} = 1 - \cos(\mathbf{x}_{h,i}, \boldsymbol{\ell}_{h-1,r})$, where i indexes candidate tokens and r indexes prototypes. A small A_{ir} indicates content already represented by $\boldsymbol{\ell}_{h-1}$, whereas a large value indicates a different appearance. Then, each candidate and prototype is represented by a nonnegative mass. We assign uniform mass to the M candidates in $\mathcal{C}_h$, $\mathbf{a} = \mathbf{1}_M/M$, and require the transport plan $\boldsymbol{\Pi} \in \mathbb{R}_+^{M \times R}$ to satisfy the candidate-side constraint $\boldsymbol{\Pi}\mathbf{1}_R = \mathbf{a}$. For prototype $\boldsymbol{\ell}_{h-1,r}$, $\mu_r = |\mathcal{C}_{h-1}^{(r)}|/M$ is the proportion of candidates in $\mathcal{C}_{h-1}$ that belong to group G_r. The R group proportions are collected in $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_R)^\top$. The transport plan $\boldsymbol{\Pi}$ defines a soft alignment between the candidate set $\mathcal{C}_h$ and the preceding-frame prototypes $\boldsymbol{\ell}_{h-1}$:

$$\boldsymbol{\Pi}^\star = \underset{\substack{\boldsymbol{\Pi} \in \mathbb{R}_+^{M \times R} \\ \boldsymbol{\Pi}\mathbf{1}_R = \mathbf{a}}}{\arg\min} \underbrace{\langle \boldsymbol{\Pi}, \mathbf{A} \rangle}_{\text{appearance matching}} + \underbrace{\varepsilon \sum_{i=1}^{M} \sum_{r=1}^{R} \Pi_{ir}(\log \Pi_{ir} - 1)}_{\text{entropy regularization}} + \underbrace{\tau D_{\mathrm{KL}}(\boldsymbol{\Pi}^\top \mathbf{1}_M \,\|\, \boldsymbol{\mu})}_{\text{prototype-marginal regularization}}, \tag{8}$$

where ε and τ are balancing coefficients. The appearance term favors candidate-prototype pairs with low matching costs, while the entropy term smooths the transport plan for stable optimization. Let $\mathbf{p} = \boldsymbol{\Pi}^\top \mathbf{1}_M$ denote the mass received by the R prototypes. The KL term keeps $\mathbf{p}$ close to the preceding group proportions $\boldsymbol{\mu}$ while allowing it to adapt to new visual content. The objective in Equation 8 is solved efficiently using semi-relaxed Sinkhorn iterations (Fukunaga & Kasai, 2022). The converged transport plan $\boldsymbol{\Pi}^\star$ softly aligns each candidate token $\mathbf{x}_{h,i}$ with the preceding-frame prototypes, providing the weights used to compute token-level temporal novelty. The complete derivation and update rules are provided in Appendix A.3.

The temporal novelty of candidate token $\mathbf{x}_{h,i}$ is defined as its transport-weighted mean cost:

$$\nu_{h,i} = \frac{\sum_{r=1}^{R} \Pi_{ir}^\star A_{ir}}{\sum_{r=1}^{R} \Pi_{ir}^\star}. \tag{9}$$

A lower value indicates repeated content, whereas a higher value suggests a potentially novel visual detail. For the first historical frame, no preceding prototype set is available, so this term is omitted.

Task-Aware Token Selection. For each historical candidate token $\mathbf{x}_{h,i} \in \mathcal{C}_h$, temporal novelty captures changes from the preceding observation but not their usefulness for the next action. We therefore add three complementary cues: current-view relevance connects the candidate to the current evidence in $\mathcal{Q}_t$, instruction relevance captures its correspondence to the instruction $\mathcal{L}$, and visual saliency preserves strong visual responses that may be missed by the other two cues. Thus, the candidate utility is computed as:

$$u_{h,i} = \underbrace{\alpha_c \max_{\mathbf{x}_{t,j} \in \mathcal{Q}_t} \cos(\mathbf{x}_{h,i}, \mathbf{x}_{t,j})}_{\text{current-view relevance}} + \underbrace{\alpha_l \max_{\mathbf{e}_m \in \mathcal{L}} \cos(\mathbf{x}_{h,i}, \mathbf{e}_m)}_{\text{instruction relevance}} + \underbrace{\alpha_s \|\mathbf{x}_{h,i}\|_2}_{\text{visual saliency}} + \underbrace{\lambda \nu_{h,i}}_{\text{temporal novelty}}, \tag{10}$$

where α_c, α_l, α_s, and λ are balancing coefficients. Notably, the temporal-novelty term is omitted for the first historical frame because no preceding prototype set is available.

Selecting tokens by utility alone may retain several nearly identical tokens. To this end, we use maximal marginal relevance (MMR) (Carbonell & Goldstein, 1998) to balance utility and diversity. Let $\mathcal{S}_h$ denote the indices of candidate tokens already selected from $\mathcal{C}_h$, initialized with the index of the highest-utility token. At each step, we iteratively select:

$$i^\star = \arg \max_{i \in \{1,\dots,M\} \setminus \mathcal{S}_h} \left[\eta u_{h,i} - (1-\eta) \max_{j \in \mathcal{S}_h} \cos(\mathbf{x}_{h,i}, \mathbf{x}_{h,j}) \right], \tag{11}$$

where $\eta \in [0, 1]$ is a balancing coefficient. The first term favors high-utility candidates, while the second penalizes similarity to tokens already selected from the same frame. Selection continues until K tokens are retained, which are then returned in their original patch order.

3.4 Policy Inference

The frozen navigation policy receives the instruction embeddings $\mathcal{L}$, the selected historical tokens in chronological order, and the current-frame tokens $\mathcal{Q}_t$ to predict the next navigation action. Each frame contributes K original projected tokens, with their features and patch positions unchanged. The agent executes the predicted action and repeats the process with the next observation. No additional training or changes to the policy are required. A detailed ***Complexity Analysis*** is provided in Appendix A.1.

4 Experiments

We evaluate our method on standard VLN benchmarks, analyze its main components and efficiency, and assess its real-world deployment. Further experimental details are provided in Appendix B.

4.1 Experimental Setups

Datasets. We conduct experiments on two public VLN-CE benchmarks: R2R-CE (Krantz et al., 2020) and RxR-CE (Ku et al., 2020). We report results on the validation-unseen splits.

Implementation Details. Following A-MMR (Wang et al., 2026a), we use $N = 729$ tokens per frame and retain $K = 218$, 146, and 72 tokens at pruning ratios of 70%, 80%, and 90%, respectively. The corresponding historical candidate budgets are $M = 436$, 292, and 144. All experiments run on an NVIDIA A100 GPU. We compare against unpruned StreamVLN (Wei et al., 2026), SparseVLM (Zhang et al., 2025c), DivPrune (Alvar et al., 2025), VisPruner (Zhang et al., 2025b), and A-MMR (Wang et al., 2026a).

Evaluation Metrics. We report success rate (SR), success weighted by path length (SPL), and navigation error (NE) on both benchmarks Guo et al. (2026). R2R-CE additionally reports oracle success rate (OS), while RxR-CE reports normalized dynamic time warping (nDTW) (Ilharco et al., 2019). SR measures successful stopping near the goal, SPL further accounts for path efficiency, and NE measures the final geodesic distance to the goal. OS records whether the trajectory reaches the goal region at any time, whereas nDTW measures agreement with the reference route.

4.2 Experimental Results

As shown in Table 1, our method improves SPL over A-MMR across all six pruning settings by an average of 3.30 points (8.1% relative). On R2R-CE, it achieves an average SPL gain of 3.36 points and reduces NE by 0.55 m. At 80% pruning, the gains reach 4.79 SR, 4.02 SPL, and 4.73 OS points, with a 0.67 m reduction in NE. Notably, the 218-token configuration retains **50.59**% SPL, exceeding the unpruned 729-token backbone by 0.93 percentage points. On RxR-CE, the average SPL gain is 3.24 points (8.6% relative), increasing from 1.56 to 4.68 points as pruning becomes more aggressive. At 90% pruning, our method yields a 14.2% relative SPL gain, a 3.94-point nDTW improvement, and a 0.36 m NE reduction. These gains arise because semantic coverage preserves complementary visual regions, while instruction-aware scoring and temporal alignment retain route-relevant landmarks and suppress repeated history, particularly under tight token budgets and on longer RxR-CE routes.

Table 1: Navigation performance on the R2R-CE (Krantz et al., 2020) and RxR-CE (Ku et al., 2020) validation-unseen splits. The unpruned StreamVLN (Wei et al., 2026) result and all competing results are reported by A-MMR (Wang et al., 2026a). **Bold** values indicate the best result.

Method	Venue	R2R (val-unseen)				RxR (val-unseen)			
		SR↑	SPL↑	OS↑	NE↓	SR↑	SPL↑	nDTW↑	NE↓
Unpruned · 729 tokens per frame									
Unpruned	ICRA'26	55.74	49.66	63.60	5.06	56.53	47.26	64.14	5.71
70% pruning · 218 tokens per frame									
SparseVLM (Zhang et al., 2025c)	ICML'25	46.82	42.18	**61.83**	5.48	40.87	35.02	53.01	7.18
DivPrune (Alvar et al., 2025)	CVPR'25	48.78	38.14	61.34	5.52	44.48	32.15	50.62	7.29
VisPruner (Zhang et al., 2025b)	ICCV'25	48.61	44.43	54.05	5.76	51.34	43.32	60.55	6.43
A-MMR (Wang et al., 2026a)	IROS'26	51.50	47.28	58.02	5.56	53.19	45.13	61.76	6.26
Ours	-	**55.25**	**50.59**	61.07	**5.04**	**53.80**	**46.69**	**62.56**	**5.89**
80% pruning · 146 tokens per frame									
SparseVLM (Zhang et al., 2025c)	ICML'25	45.19	40.63	**60.09**	5.55	34.59	30.10	52.89	7.64
DivPrune (Alvar et al., 2025)	CVPR'25	39.31	28.64	55.14	6.47	35.13	23.92	42.68	8.49
VisPruner (Zhang et al., 2025b)	ICCV'25	50.52	43.92	58.51	5.59	49.16	40.15	57.64	6.73
A-MMR (Wang et al., 2026a)	IROS'26	52.58	46.44	60.09	5.38	53.36	43.64	61.02	6.16
Ours	-	**57.37**	**50.46**	**64.82**	**4.71**	**56.26**	**47.13**	**63.14**	**5.73**
90% pruning · 72 tokens per frame									
SparseVLM (Zhang et al., 2025c)	ICML'25	34.91	31.08	45.95	5.98	23.17	20.87	50.06	8.51
DivPrune (Alvar et al., 2025)	CVPR'25	27.57	18.55	46.06	7.36	23.22	14.56	33.73	9.83
VisPruner (Zhang et al., 2025b)	ICCV'25	41.16	29.27	66.83	6.43	37.34	25.34	41.05	7.60
A-MMR (Wang et al., 2026a)	IROS'26	47.63	36.36	68.46	5.69	45.71	32.91	47.69	6.90
Ours	-	**51.77**	**39.10**	**71.72**	**5.23**	**47.18**	**37.59**	**51.63**	**6.54**

Table 2: Ablation studies on the R2R-CE validation-unseen split. 'LRG' and 'WGS' denote low-rank grouping and within-group selection, respectively. 'Equal quota' assigns the same token budget to every feature group. 'w/o Text' removes instruction relevance; 'w/o Temporal' removes OT-based temporal novelty; 'w/o MMR' directly selects the top-K candidates by utility; and 'Balanced OT' replaces semi-relaxed OT with balanced OT. **Bold** values indicate the best result.

Setups	218 tokens (↓ 70%)				146 tokens (↓ 80%)				72 tokens (↓ 90%)			
	SR↑	SPL↑	OS↑	NE↓	SR↑	SPL↑	OS↑	NE↓	SR↑	SPL↑	OS↑	NE↓
w/o LRG	51.22	46.62	57.04	5.48	54.98	47.99	64.55	5.09	45.95	34.26	67.65	5.69
w/o WGS	54.49	49.93	60.52	5.09	54.98	48.43	63.02	5.04	48.94	37.70	70.09	5.55
Equal quota	54.87	50.29	60.96	**4.90**	57.15	50.23	**65.14**	4.80	50.95	**39.27**	70.58	5.34
w/o Text	53.89	49.30	60.41	5.20	56.28	49.21	63.62	4.78	48.61	36.64	69.55	5.47
w/o Temporal	53.51	48.90	60.63	5.11	56.44	49.85	65.04	**4.71**	49.10	37.47	68.57	5.61
w/o MMR	54.16	49.58	60.36	5.06	57.21	50.41	64.38	4.75	49.76	38.47	69.98	5.45
Balanced OT	53.13	48.47	59.92	5.13	56.39	49.64	64.60	4.78	49.76	37.83	69.98	5.38
Full model	**55.25**	**50.59**	**61.07**	5.04	**57.37**	**50.46**	64.82	4.71	**51.77**	39.10	**71.72**	**5.23**

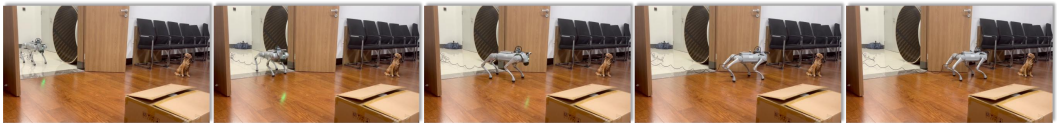
4.3 Ablation Studies

The ablation results show that both intra- and inter-frame components become increasingly important under tighter token budgets. Removing low-rank grouping causes the largest degradation; at 72 tokens, SPL drops by 4.84 points and NE increases by 0.46 m. Within-group selection also provides larger gains at higher pruning ratios, while proportional allocation achieves a more balanced trade-off across all metrics than equal allocation. For historical selection, removing text relevance, temporal novelty, or MMR at 72 tokens reduces SPL by 2.46, 1.63, and 0.63 points, respectively. Replacing semi-relaxed OT with balanced OT decreases SPL by 2.12, 0.82, and 1.27 points across the three budgets and consistently increases NE. These results validate the complementary roles of semantic coverage, instruction guidance, temporal novelty, and adaptive prototype matching.

Efficiency Analysis. Table 3 reports inference efficiency on an NVIDIA A100 GPU. Relative to unpruned StreamVLN, our method improves throughput by 41.1%, reduces latency by 24.4%, and lowers TFLOPs by 46.5%. Although DivPrune is faster and VisPruner requires marginally fewer TFLOPs, both achieve substantially lower navigation performance in Table 1. Overall, our method provides a stronger navigation–efficiency trade-off among the evaluated pruning methods.



Instruction: *Walk straight out of the room through the open wooden doorway. Continue across the open area, then turn right at the end into the corridor beside the red fire-safety cabinet. Follow the corridor and stop in front of the dog sitting along the left wall.*



Instruction: *Go straight through the open doorway into the wood-floored meeting room. Walk toward the large cardboard box, then turn left and pass the box on its left. Enter the aisle between the rows of chairs, continue forward, and stop in front of the dog sitting ahead.*



Instruction: *Go through the wooden doorway into the hall, keeping the large circular panel on your right. Continue straight to the doorway beside the blue umbrella. Go through the doorway and turn right. Follow the passage between the office chair and the gray cabinet, then stop in front of the office chair placed on the red-and-black box.*

Figure 3: Real-world deployment on the Unitree Go2. Each row shows sampled observations from one instruction-following trajectory. Red text marks intermediate route cues, while green text identifies the final target. Additional real-world evaluations are provided in Appendix B.2.3.

4.4 Real-world Deployment

We deploy the proposed method on a Unitree Go2 quadruped robot. StreamVLN inference runs on a remote desktop workstation equipped with an NVIDIA RTX 4090 GPU, while a lightweight control client runs onboard the robot's NVIDIA Jetson Orin NX. A single Intel RealSense D435i camera captures RGB observations, which are transmitted to the workstation over a local network. The predicted actions are returned to the onboard client for execution, and robot motion is estimated using the Go2's proprioceptive odometry.

Table 3: Efficiency analysis of inference throughput (FPS), computational cost (TFLOPs), and CUDA inference latency for all methods on the StreamVLN backbone using an NVIDIA A100 GPU. The best and second-best results among pruning methods are shown in **bold** and underlined, respectively.

Method	FPS↑	TFLOPs↓	Latency (ms)↓
Unpruned	3.14	3.76	328.34
SparseVLM	4.03	7.62	248.26
DivPrune	**4.68**	<u>2.01</u>	**213.64**
VisPruner	4.04	**2.00**	<u>247.57</u>
Ours	<u>4.43</u>	<u>2.01</u>	248.26

Figure 3 presents three representative instruction-following trajectories in indoor environments. The routes require the robot to pass through multiple doorways and corridors, follow intermediate landmarks such as cabinet, box, panel, and umbrella, and stop at the specified target. The sampled observations show that the compressed visual history retains route-relevant evidence over extended trajectories and supports the required turning and stopping behavior.

5 Conclusion

We present a training-free spatio-temporal visual-token compressor for efficient VLN. Low-rank grouping, within-group selection, and proportional allocation preserve complementary evidence within each frame, while instruction-aware scoring and prototype-based semi-relaxed OT retain task-relevant history and suppress cross-frame redundancy. By selecting original visual tokens, the compressor requires neither retraining nor additional prior knowledge. Experiments on R2R-CE and RxR-CE demonstrate consistent gains across compression ratios and a favorable navigation and efficiency trade-off. Deployment on a Unitree Go2 further validates its practicality for low-latency embodied navigation.

AI USE STATEMENT

In this work, we used generative AI tools to assist with understanding existing code, diagnosing and fixing implementation bugs, and improving the clarity and grammar of the manuscript. All AI-assisted code changes were manually reviewed and validated through experiments, and all AI-assisted writing was checked and revised by the authors. We also use a pretrained generative vision-language model as the navigation Path Planner in real deployment, as described in the paper. Generative AI was not used to fabricate data, results, citations, or scientific claims. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9392–9401, 2025.

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 *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 3674–3683. IEEE, 2018.

Jaime G Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, volume 98, pp. 290941–291025, 1998.

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*, pp. 19–35. Springer, 2024.

An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Biyik, Hongxu Yin, Sifei Liu, and Xiaolong Wang. Navila: Legged robot vision-language-action model for navigation. *RSS*, 2025.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Zhengyao Fang, Pengyuan Lyu, Chengquan Zhang, Guangming Lu, Jun Yu, and Wenjie Pei. Prune redundancy, preserve essence: Vision token compression in vlms via synergistic importance-diversity. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, 2026.

Takumi Fukunaga and Hiroyuki Kasai. On the convergence of semi-relaxed sinkhorn with marginal constraint and ot distance gaps. *arXiv preprint arXiv:2205.13846*, 2022.

Wenxuan Guo, Xiuwei Xu, Yichen Liu, Xiangyu Li, Hang Yin, Huangxing Chen, Wenzhao Zheng, Jianjiang Feng, Jie Zhou, and Jiwen Lu. Awarevln: Reasoning with self-awareness for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4065–4075, June 2026.

Gabriel Ilharco, Vihan Jain, Alexander Ku, Eugene Ie, and Jason Baldridge. General evaluation for instruction conditioned navigation using dynamic time warping. *arXiv preprint arXiv:1907.05446*, 2019.

Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 374(2065):20150202, 2016.

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

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4392–4412, 2020.

Yang Li, Aming Wu, Zihao Zhang, and Yahong Han. Towards open environments and instructions: General vision-language navigation via fast-slow interactive reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25184–25192, June 2026.

Yantao Lu, Shiqi Sun, Ning Liu, Bo Jiang, Ying Zhang, Jinchao Chen, and Chenglie Du. Stepnav: Spatial-temporal efficient visual token pruning for vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24097–24105, 2026.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9338–9346. IEEE, 2019.

Xinshuai Song, Weixing Chen, Yang Liu, Weikai Chen, Guanbin Li, and Liang Lin. Towards long-horizon vision-language navigation: Platform, benchmark and method. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12078–12088. IEEE, 2025.

Qitong Wang, Yijun Liang, Ming Li, Tianyi Zhou, and Christopher Rasmussen. History-conditioned spatio-temporal visual token pruning for efficient vision-language navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2026a.

Shuo Wang, Yongcai Wang, Zhaoxin Fan, Yucheng Wang, Maiyue Chen, Kaihui Wang, Zhizhong Su, Wanting Li, Xudong Cai, Yeying Jin, and Deying Li. Monodream: Monocular vision-language navigation with panoramic dreaming. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10074–10082, 2026b.

Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, and Shuqiang Jiang. Gridmm: Grid memory map for vision-and-language navigation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15579–15590. IEEE, 2023.

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

Zihao Xin, Wentong Li, Yixuan Jiang, Bin Wang, Runmin Cong, Jie Qin, and Shengjun Huang. Decovln: Decoupling observation, reasoning, and correction for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 32410–32420, June 2026.

Yunzhe Xu, Yiyuan Pan, and Zhe Liu. Dream to recall: Imagination-guided experience retrieval for memory-persistent vision-and-language navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Chenyu Yang, Xuan Dong, Xizhou Zhu, Weijie Su, Jiahao Wang, Hao Tian, Zhe Chen, Wenhai Wang, Lewei Lu, and Jifeng Dai. Pvc: Progressive visual token compression for unified image and video processing in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24939–24949, June 2025.

Jiahao Yang, Zihan Wang, Xiangyang Li, Xing Zhu, Yujun Shen, Yinghao Xu, and Shuqiang Jiang. Ga-vln: Geometry-aware bev representation for efficient vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40053–40063, 2026.

Shuang Zeng, Dekang Qi, Xinyuan Chang, Feng Xiong, Shichao Xie, Xiaolong Wu, Shiyi Liang, Mu Xu, and Xing Wei. Janusvln: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. In *International Conference on Learning Representations*, volume 2026, pp. 33001–33026, 2026.

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. *RSS*, 2025a.

Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20857–20867. IEEE, 2025b.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. In *International Conference on Machine Learning*, pp. 74840–74857. PMLR, 2025c.

APPENDIX

A THEORETICAL ANALYSIS AND ALGORITHMS

This appendix provides additional analysis of the two compression stages and complete pseudocode. Unlike an abstract information-set argument, the results below state properties that follow directly from the operations used by our method. Throughout the appendix, R_0 denotes the rank retained by the truncated decomposition, and $R \leq R_0$ denotes the number of non-empty feature groups after empty assignments are removed. Table 4 summarizes the notation used in the following analyses.

Table 4: Notation used throughout Appendix A.

Symbol	Meaning
N, d, B, R_0, R	Number of tokens, feature dimension, output budget, retained rank, and number of non-empty feature groups.
H, K, I, n_r	Number of historical frames, final per-frame token budget, Sinkhorn iterations, and tokens assigned to group r.
$C_{\text{intra/inter/total/A-MMR}}$	Intra-frame, inter-frame, total, and A-MMR computational complexities.
$\mathbf{X}, \widehat{\mathbf{X}}$	Original token matrix and its bounded, centered version.
$\mathbf{P}, \mathbf{\Lambda}, \mathbf{D}$	Patch loadings, singular values, and dominant feature directions from truncated SVD.
$G_r, \widehat{\mathcal{V}}_r$	Token indices assigned to group r and the subset retained by within-group selection.
$\theta_B, s(\cdot,\cdot), \bar{\theta}_B$	Cosine-similarity threshold, shifted cosine similarity, and its shifted threshold.
$b_r, f_r, \mathcal{S}_B$	Number of initially retained and restored tokens in group r, and the final set of B tokens.
$\mathcal{C}_h, M, \mathbf{x}_{h,i}$	Candidate pool of historical frame h, its size, and candidate token i.
$\boldsymbol{\ell}_{h-1,r}, \mathbf{A}$	Prototype r from the preceding historical frame and the candidate-prototype cost matrix.
$\mathbf{a}, \boldsymbol{\mu}$	Candidate masses and preceding-frame group proportions.
$\mathbf{\Pi}, \mathbf{p}$	Transport plan and its induced prototype-side mass.
ε, τ	Entropy coefficient and prototype-marginal regularization coefficient.
$\mathbf{H}, \mathbf{u}, \mathbf{v}$	Positive transport kernel and its row- and column-scaling vectors.
$\nu_{h,i}$	Temporal novelty of candidate $\mathbf{x}_{h,i}$.

A.1 COMPLEXITY ANALYSIS

For clarity, let H denote the number of historical frames. The intra-frame stage is applied once to the current frame and each of the H historical frames. For one frame, low-rank grouping costs $O(NdR)$, within-group selection costs $O\left(d\sum_{r=1}^{R} n_r^2\right)$, and sorting and quota allocation cost $O(N \log N)$. Thus, its total cost is:

$$C_{\text{intra}} = O\left((H+1)\left[NdR + d\sum_{r=1}^{R} n_r^2 + N\log N\right]\right). \tag{12}$$

The inter-frame stage is applied only to the H historical frames after each frame has been reduced to M candidates. Prototype construction and semi-relaxed alignment cost $O(HMR(d+I))$, while task-aware scoring and MMR cost $O(HMKd)$. Hence,

$$C_{\text{inter}} = O(HMR(d+I) + HMKd)\,,\; C_{\text{total}} = C_{\text{intra}} + C_{\text{inter}}, \tag{13}$$

where I is the number of Sinkhorn iterations.

A-MMR (Wang et al., 2026a) directly performs relevance scoring and diversity selection over all N tokens, giving an overall complexity of $C_{\text{A-MMR}} = O((H+1)NKd)$. For approximately balanced groups, $\sum_r n_r^2 \approx N^2/R$, so our intra-frame cost becomes $O\big((H+1)(NdR + N^2d/R)\big)$,

excluding the lower-order sorting term. With $R \ll K$ and $N/R < K$, this cost is no larger than A-MMR's global selection. Our method additionally performs semi-relaxed temporal alignment, but only between M candidates and R prototypes. Its $O(HMR(d+I))$ cost is small relative to the $O(HMKd)$ scoring-and-selection term when $R(d+I) \ll Kd$. Moreover, because $M < N$, this dominant historical term is lower than A-MMR's $O(HNKd)$. Therefore, the savings from candidate compression largely offset the added alignment cost: explicit cross-frame modeling introduces only a small overhead and does not increase the dominant overall complexity.

Computational benefit of Prototypes. Constructing the R prototypes costs $O(Md)$ once group memberships are known. Computing $\mathbf{A}$ costs $O(MRd)$, and each semi-relaxed Sinkhorn iteration costs $O(MR)$. The transport matrix requires $O(MR)$ memory. Direct token-to-token transport would instead require an $M \times M$ cost matrix, $O(M^2 d)$ cost construction, and $O(M^2)$ work per iteration. Prototype transport is therefore cheaper whenever $R \ll M$.

A.2 Theoretical Analysis of Intra-frame Semantic Coverage

The intra-frame stage has a simple goal: represent the different visual patterns in one frame using only B original tokens. Its justification requires three claims. First, the low-rank representation must retain the strongest variations in the original features. Second, selection within each group must remove near-duplicate tokens. Third, budget allocation must keep every non-empty group while returning exactly B tokens. We prove these claims in the same order as the method.

Low-rank preservation. Let $\widehat{\mathbf{X}} \in \mathbb{R}^{N\times d}$ be the bounded and centered token matrix defined in Eq. (1), and let $\lambda_1 \geq \cdots \geq \lambda_{\min(N,d)} \geq 0$ be its singular values. For $1 \leq R_0 \leq \operatorname{rank}(\widehat{\mathbf{X}})$, truncated SVD keeps the first R_0 directions and gives

$$\widehat{\mathbf{X}}_{R_0} = \mathbf{P}\mathbf{\Lambda}\mathbf{D}^\top. \tag{14}$$

where $\mathbf{P} \in \mathbb{R}^{N\times R_0}$ contains the token loadings, $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_{R_0})$ contains the retained singular values, and $\mathbf{D} \in \mathbb{R}^{d\times R_0}$ contains the corresponding feature directions. Thus, $\widehat{\mathbf{X}}_{R_0}$ summarizes the R_0 strongest variations among the N token features.

Proposition 1 (Optimality of Truncated SVD)**.** *Among all matrices of rank at most R_0, $\widehat{\mathbf{X}}_{R_0}$ has the smallest reconstruction error:*

$$\widehat{\mathbf{X}}_{R_0} \in \operatorname*{arg\,min}_{\operatorname{rank}(\mathbf{Y})\leq R_0} \|\widehat{\mathbf{X}} - \mathbf{Y}\|_{\mathrm{F}}^2, \qquad \|\widehat{\mathbf{X}} - \widehat{\mathbf{X}}_{R_0}\|_{\mathrm{F}}^2 = \sum_{g>R_0} \lambda_g^2. \tag{15}$$

where $\mathbf{Y} \in \mathbb{R}^{N\times d}$ is any competing matrix of rank at most R_0, $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm, and λ_g is the g-th singular value of $\widehat{\mathbf{X}}$.

Proof. SVD orders feature directions by decreasing singular value. The Eckart-Young theorem (Eckart & Young, 1936) states that retaining the first R_0 directions minimizes the squared reconstruction error, whose value is the energy in the omitted directions, $\sum_{g>R_0} \lambda_g^2$. Thus, the retained subspace loses the least possible feature energy among all rank-R_0 linear summaries. This does not imply that a singular direction is a predefined object category. □

Remark. This result explains why low-rank grouping is useful before pruning. It exposes the strongest shared response patterns without labels or retraining, allowing the token budget to cover complementary visual evidence rather than only one conspicuous region. In VLN, this gives a small but distinctive landmark its own route into the retained set.

Within-group separation. After grouping, similar tokens may still occur repeatedly within the same group. The selection rule therefore scans each group in descending loading order and retains a token only when it is not too similar to any token already retained. For a non-empty index set G_r, let $\widehat{\mathcal{V}}_r \subseteq G_r$ contain the indices retained before optional fallback restoration.

Proposition 2 (Within-Group Separation Guarantee)**.** *Any two distinct tokens retained in the same group satisfy*

$$\cos(\mathbf{x}_i, \mathbf{x}_j) \leq \theta_B, \qquad \forall\, i, j \in \widehat{\mathcal{V}}_r, \quad i \neq j. \tag{16}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the features of two distinct retained patches, and θ_B is the adaptive similarity threshold for budget B.

Proof. The first token is always kept. A later token is kept only if its cosine similarity to every previously retained token is at most θ_B. Consequently, each accepted token satisfies the threshold with all earlier survivors. The same check is applied throughout the scan, so every pair in the final set satisfies Eq. equation 16. □

Remark. The proposition gives a direct redundancy guarantee within each feature group: under normal selection, two patches more similar than θ_B cannot both survive. Applying this rule within, rather than across, groups removes repeated evidence without allowing a large region to suppress a different landmark represented by another group.

The pairwise guarantee also yields a group-level redundancy bound. Define $s(\mathbf{x}_i, \mathbf{x}_j) = [1 + \cos(\mathbf{x}_i, \mathbf{x}_j)]/2$ and $\bar{\theta}_B = (1 + \theta_B)/2$. If $b_r = |\widehat{\mathcal{V}}_r|$, then

$$\sum_{r=1}^{R} \sum_{\{i,j\} \subseteq \widehat{\mathcal{V}}_r} s(\mathbf{x}_i, \mathbf{x}_j) \leq \bar{\theta}_B \sum_{r=1}^{R} \binom{b_r}{2}. \tag{17}$$

where $s(\mathbf{x}_i, \mathbf{x}_j) \in [0, 1]$ is the shifted cosine similarity, $\bar{\theta}_B$ is its threshold, $\{i, j\} \subseteq \widehat{\mathcal{V}}_r$ denotes an unordered index pair, and $\binom{b_r}{2}$ is the number of retained pairs in group r. If the fallback restores f_r suppressed tokens to group r, the threshold guarantee still holds for the original b_r survivors. Only pairs involving a restored token may reach the maximum value $s = 1$, giving the valid bound

$$\sum_{r=1}^{R} \left[\bar{\theta}_B \binom{b_r}{2} + b_r f_r + \binom{f_r}{2} \right]. \tag{18}$$

where f_r is the number of restored tokens in group r. The three terms bound pairs of two original survivors, one original and one restored token, and two restored tokens, respectively. Thus, fallback relaxes the separation guarantee only for the tokens needed to make the requested budget feasible.

Coverage and budget. Suppressing duplicates is not sufficient by itself: a global top-B rule could still spend most of the budget on one large group. The allocation step must therefore prove two properties simultaneously: every non-empty group remains represented, and the total output contains exactly B tokens.

Proposition 3 (Group Coverage and Exact Budget Guarantee)**.** *Assume $B \geq R$ and that at least B eligible tokens are available after optional restoration. Proportional Budget Allocation returns a set $\mathcal{S}_B$ satisfying*

$$|\mathcal{S}_B| = B, \qquad |\mathcal{S}_B \cap G_r| \geq 1, \quad r = 1, \ldots, R. \tag{19}$$

where $\mathcal{S}_B$ is the final selected set and $\mathcal{S}_B \cap G_r$ is the subset selected from non-empty group r.

Proof. Because $B \geq R$, one slot can first be reserved for each non-empty group. This already guarantees coverage. The largest-remainder rule assigns the remaining $B - R$ slots. If a group has too few candidates, its unused slots are passed to groups that still contain unselected candidates. Since at least B candidates are available, this redistribution continues until all B slots are filled. The output therefore contains every group and exactly B tokens. □

Remark. This proposition connects semantic coverage to the fixed interface required by the VLN policy. Every discovered feature pattern contributes evidence, while the compressor still returns exactly K current-frame tokens or M historical candidates. Coverage is therefore improved without changing the policy input format.

Together, these results explain the role of the complete intra-frame stage. Low-rank decomposition preserves the dominant feature patterns, within-group selection limits repeated evidence inside each pattern, and proportional allocation prevents a complementary pattern from disappearing under a global ranking.

A.3 Theoretical Analysis of Inter-frame Semi-relaxed OT

The inter-frame stage measures whether candidate $\mathbf{x}_{h,i}$ repeats visual content already present in the preceding historical frame. A valid novelty measure must do four things: evaluate every candidate, allow objects to appear or disappear, have a stable and efficient solution, and give each token an interpretable score. We therefore show that the semi-relaxed alignment has a unique solution, explain how Sinkhorn iterations compute it, prove that the resulting novelty is bounded, and clarify how τ controls the relaxation.

Adaptive alignment. Let $\mathbf{A} \in \mathbb{R}^{M\times R}$ be the appearance-cost matrix, with $A_{ir} = 1 - \cos(\mathbf{x}_{h,i}, \boldsymbol{\ell}_{h-1,r})$. A small A_{ir} means that candidate $\mathbf{x}_{h,i}$ resembles preceding-frame prototype $\boldsymbol{\ell}_{h-1,r}$. Every candidate receives mass $1/M$, collected in $\mathbf{a} = \mathbf{1}_M/M$. Because only non-empty groups produce prototypes, their preceding-frame proportions satisfy $\mu_r = |\mathcal{C}_{h-1}^{(r)}|/M > 0$ and $\sum_r \mu_r = 1$. Semi-relaxed OT solves

$$\boldsymbol{\Pi}^\star = \underset{\substack{\boldsymbol{\Pi}\in\mathbb{R}_+^{M\times R}\\ \boldsymbol{\Pi}\mathbf{1}_R=\mathbf{a}}}{\arg\min} \left\{ \langle \boldsymbol{\Pi}, \mathbf{A}\rangle + \varepsilon \sum_{i,r} \Pi_{ir}(\log \Pi_{ir} - 1) + \tau D_{\mathrm{KL}}\big(\boldsymbol{\Pi}^\top \mathbf{1}_M \,\big\|\, \boldsymbol{\mu}\big) \right\}. \tag{20}$$

where $\langle \boldsymbol{\Pi}, \mathbf{A}\rangle = \sum_{i=1}^{M}\sum_{r=1}^{R} \Pi_{ir}A_{ir}$, $\mathbf{1}_q$ is a length-q all-ones vector, and $D_{\mathrm{KL}}(\mathbf{p}\|\boldsymbol{\mu}) = \sum_{r=1}^{R} p_r \log(p_r/\mu_r)$ for $\mathbf{p} = \boldsymbol{\Pi}^\top \mathbf{1}_M$. We use the continuous convention $0\log 0 = 0$ in both the entropy and KL terms. Since $\mathbf{p}$ and $\boldsymbol{\mu}$ have unit mass, this definition is equivalent on the feasible set to the generalized KL divergence $\sum_r [p_r \log(p_r/\mu_r) - p_r + \mu_r]$. The coefficients $\varepsilon > 0$ and $\tau \geq 0$ balance transport smoothness and adherence to the preceding-frame proportions, respectively. The first term favors low-cost visual matches, the second spreads mass smoothly, and the third discourages unnecessary changes in prototype usage. The constraint $\boldsymbol{\Pi}\mathbf{1}_R = \mathbf{a}$ includes every candidate, whereas the prototype-side mass $\mathbf{p}$ is allowed to adapt when content appears or disappears.

Proposition 4 (Existence and Uniqueness of the Semi-Relaxed Transport Plan)**.** *Assume that $\mathbf{A}$ is finite, $\varepsilon > 0$, $\tau \geq 0$, and every entry of $\mathbf{a}$ and $\boldsymbol{\mu}$ is positive. Equation equation 20 then has a unique minimizer $\boldsymbol{\Pi}^\star$ with $\Pi_{ir}^\star > 0$ for all i and r. It satisfies $\boldsymbol{\Pi}^\star\mathbf{1}_R = \mathbf{a}$, while the prototype-side marginal $\mathbf{p}^\star = (\boldsymbol{\Pi}^\star)^\top\mathbf{1}_M$ remains adaptive.*

Proof. The feasible set is non-empty because $\mathbf{a}\boldsymbol{\mu}^\top$ satisfies the row constraint. It is closed and bounded, hence compact. Under the convention $0\log 0 = 0$, the objective is continuous and therefore attains a minimum. The entropy term is strictly convex for $\varepsilon > 0$, while the transport and KL terms are linear and convex, respectively. Their sum is thus strictly convex, which makes the minimizer unique.

It remains to show that the minimizer is positive. Suppose that $\Pi_{ir}^\star = 0$. Since $a_i > 0$, the same row contains some q with $\Pi_{iq}^\star > 0$. For sufficiently small $\delta > 0$, moving mass δ from (i, q) to (i, r) preserves feasibility. The new entry contributes $\varepsilon\delta\log\delta$, whereas the changes of the finite transport cost and all terms evaluated at positive masses are $O(\delta)$. If the receiving prototype has zero marginal, its KL contribution also has order $\delta\log\delta$ and is non-positive for small δ. Because $\delta\log\delta$ dominates any linear term as $\delta \downarrow 0$, this perturbation decreases the objective, contradicting optimality. Hence every entry of $\boldsymbol{\Pi}^\star$ is positive, so the first-order optimality conditions used below are valid. The row constraint fixes the candidate-side mass, whereas the prototype-side marginal is regularized rather than fixed. □

Remark. Uniqueness makes the temporal novelty score deterministic for fixed features and hyperparameters. More importantly, the adaptive prototype marginal matches the VLN setting: every historical candidate is examined, but newly visible content is not forced to occupy the same group proportions as the preceding view.

Sinkhorn solution. The optimal plan can be computed by repeated matrix scaling rather than by directly optimizing all MR transport variables. First, the costs are converted into positive affinities using $\mathbf{H} = \exp(-\mathbf{A}/\varepsilon)$ element-wise: a smaller cost produces a larger entry in $\mathbf{H}$. The first-order optimality conditions give the scaled form

$$\boldsymbol{\Pi}^\star = \mathrm{diag}(\mathbf{u}^\star)\mathbf{H}\,\mathrm{diag}(\mathbf{v}^\star). \tag{21}$$

where $\mathrm{diag}(\mathbf{u}^\star)$ and $\mathrm{diag}(\mathbf{v}^\star)$ are diagonal matrices formed from the positive scaling vectors $\mathbf{u}^\star \in \mathbb{R}_+^M$ and $\mathbf{v}^\star \in \mathbb{R}_+^R$, respectively. Starting from any positive $\mathbf{v}^{(0)}$, the two scaling vectors are updated by

$$\mathbf{u}^{(s+1)} = \mathbf{a} \oslash (\mathbf{H}\mathbf{v}^{(s)}), \qquad \mathbf{v}^{(s+1)} = \left[\boldsymbol{\mu} \oslash (\mathbf{H}^\top \mathbf{u}^{(s+1)})\right]^{\frac{\tau}{\tau+\varepsilon}}, \tag{22}$$

where s is the iteration index, $\oslash$ denotes element-wise division, and the power is applied element-wise. The $\mathbf{u}$ update enforces the candidate-side mass, while the $\mathbf{v}$ update softly adjusts prototype usage. After the final $\mathbf{v}$ update, $\mathbf{u}$ is recomputed and Eq. equation 21 recovers the plan. For positive $\mathbf{a}$, $\boldsymbol{\mu}$, and $\mathbf{H}$, these iterations converge to the unique solution of Eq. equation 20 (Fukunaga & Kasai, 2022). Only matrix-vector products are required; no VLN parameter is updated.

Novelty bounds. Once the alignment is obtained, the novelty of candidate $\mathbf{x}_{h,i}$ is its transport-weighted appearance cost:

$$\nu_{h,i} = \frac{\sum_{r=1}^{R} \Pi^\star_{ir} A_{ir}}{\sum_{r=1}^{R} \Pi^\star_{ir}}. \tag{23}$$

where the numerator accumulates candidate-prototype costs using the transport weights, and the denominator is the total mass of candidate $\mathbf{x}_{h,i}$, equal to $1/M$ under the row constraint.

Proposition 5 (Bounds on Temporal Novelty)**.** *For every candidate* $\mathbf{x}_{h,i}$*,*

$$\min_r A_{ir} \le \nu_{h,i} \le \max_r A_{ir}, \qquad 0 \le \nu_{h,i} \le 2. \tag{24}$$

where the minimum and maximum are taken over the R *preceding-frame prototypes.*

Proof. The row constraint gives $\sum_r \Pi^\star_{ir} = 1/M$. Therefore, $w_{ir} = M\Pi^\star_{ir}$ forms nonnegative weights that sum to one. The novelty score is consequently a weighted average, $\nu_{h,i} = \sum_r w_{ir} A_{ir}$. A weighted average cannot be smaller than its minimum input or larger than its maximum input. Since cosine similarity lies in $[-1, 1]$, every cost $A_{ir} = 1 - \cos(\cdot, \cdot)$ lies in $[0, 2]$, which gives the second bound. □

Remark. The bounded score has a direct operational meaning. It stays on the same fixed scale for every token and frame, so it can be combined with current-view relevance, instruction relevance, and saliency without depending on the number of prototypes. High novelty identifies evidence not already summarized by the preceding history.

Thus, $\nu_{h,i}$ is easy to interpret: a small value indicates repeated content, whereas a large value indicates content that is not well represented by the preceding-frame prototypes.

Role of τ. The coefficient τ controls how strongly the preceding-frame group proportions influence the new alignment.

Proposition 6 (Limiting Behavior of Prototype-Marginal Relaxation)**.** *Let* $\mathbf{\Pi}^\star_\tau$ *denote the minimizer of Eq. equation 20, and let* $\mathbf{p}_\tau = (\mathbf{\Pi}^\star_\tau)^\top \mathbf{1}_M$*. As* $\tau \to \infty$*,* $\mathbf{p}_\tau \to \boldsymbol{\mu}$ *and* $\mathbf{\Pi}^\star_\tau$ *converges to the unique entropy-regularized balanced OT plan with marginals* $\mathbf{a}$ *and* $\boldsymbol{\mu}$*. When* $\tau = 0$*, each candidate is matched independently:*

$$\Pi^\star_{ir} = \frac{1}{M} \frac{\exp(-A_{ir}/\varepsilon)}{\sum_{q=1}^{R} \exp(-A_{iq}/\varepsilon)}. \tag{25}$$

where q *indexes the* R *prototypes in the normalizing denominator.*

Proof. Let $\mathcal{K} = \{\mathbf{\Pi} \ge 0 : \mathbf{\Pi}\mathbf{1}_R = \mathbf{a}\}$ and let $F(\mathbf{\Pi})$ denote the transport and entropy terms in Eq. equation 20. Since $\mathcal{K}$ is compact and F is continuous, $m = \min_{\mathbf{\Pi}\in\mathcal{K}} F(\mathbf{\Pi})$ is finite. A balanced plan exists because $\mathbf{Q} = \mathbf{a}\boldsymbol{\mu}^\top$ has marginals $\mathbf{a}$ and $\boldsymbol{\mu}$. For any balanced feasible plan $\mathbf{Q}$, optimality of $\mathbf{\Pi}^\star_\tau$ gives

$$0 \le D_{\mathrm{KL}}(\mathbf{p}_\tau \| \boldsymbol{\mu}) \le \frac{F(\mathbf{Q}) - m}{\tau} \longrightarrow 0. \tag{26}$$

Thus, $\mathbf{p}_\tau \to \boldsymbol{\mu}$. By compactness, every sequence with $\tau \to \infty$ has a subsequence for which $\mathbf{\Pi}^\star_\tau$ converges to some $\overline{\mathbf{\Pi}}$. The row constraint is preserved by continuity, and $\mathbf{p}_\tau \to \boldsymbol{\mu}$ gives the required column marginal, so $\overline{\mathbf{\Pi}}$ is balanced. Moreover, optimality and the non-negativity of KL

Algorithm 1 Intra-frame Semantic Coverage

Require: $\mathbf{X} \in \mathbb{R}^{N \times d}$, budget B, rank R_0, threshold θ_B
Ensure: $\mathcal{S}_B$, $|\mathcal{S}_B| = B$

$\widehat{\mathbf{X}} \leftarrow \mathrm{Center}(\sigma(\mathbf{X}))$ ▷ bound and center features
$(\mathbf{P}, \mathbf{\Lambda}, \mathbf{D}) \leftarrow \mathrm{SVD}_{R_0}(\widehat{\mathbf{X}})$
$g(i) \leftarrow \arg\max_r |P_{ir}|, \quad G_r \leftarrow \{i : g(i) = r\}$ ▷ form feature groups
Remove empty G_r and relabel the remaining groups as $r = 1, \ldots, R$
for $r = 1, \ldots, R$ **do**
  $\widehat{\mathcal{V}}_r \leftarrow \varnothing$
  for $i \in \mathrm{argsort}_{j \in G_r}(-|P_{j,g(j)}|)$ **do**
    if $\widehat{\mathcal{V}}_r = \varnothing$ or $\max_{j \in \widehat{\mathcal{V}}_r} \cos(\mathbf{x}_i, \mathbf{x}_j) \le \theta_B$ **then**
      $\widehat{\mathcal{V}}_r \leftarrow \widehat{\mathcal{V}}_r \cup \{i\}$
    end if
  end for
end for
if $\sum_r |\widehat{\mathcal{V}}_r| < B$ **then**
  Restore the highest-loading suppressed indices until $\sum_r |\widehat{\mathcal{V}}_r| = B$
end if
$\mathbf{q} \leftarrow \mathrm{LargestRemainder}(B, \{|\widehat{\mathcal{V}}_r|\}_{r=1}^{R})$ ▷ cap quotas and reassign unfilled slots
$\mathcal{S}_B \leftarrow \mathrm{Order}\Big(\bigcup_r \mathrm{Top}_{q_r}(\widehat{\mathcal{V}}_r, |P_{i,g(i)}|)\Big)$
return $\mathcal{S}_B$

imply $F(\mathbf{\Pi}_\tau^\star) \le F(\mathbf{Q})$ for every balanced $\mathbf{Q}$. Taking limits shows that $\overline{\mathbf{\Pi}}$ minimizes the entropy-regularized balanced OT objective. This balanced minimizer is unique because $\varepsilon > 0$ makes F strictly convex. Hence every convergent subsequence has the same limit, and the full sequence $\mathbf{\Pi}_\tau^\star$ converges to that balanced plan.

When $\tau = 0$, the KL term vanishes and the rows become independent. Minimizing each row subject to $\sum_r \Pi_{ir} = 1/M$ gives the normalized exponential weights in Eq. equation 25. □

Remark. The limits clarify why a finite τ is used for temporal alignment. A very large value reproduces the rigid capacity matching that can create artificial correspondences when a landmark enters or leaves view; zero discards all information about the preceding group distribution. A finite value keeps that distribution as guidance while remaining responsive to genuine visual change.

A finite τ therefore lies between the two extremes: preceding-frame proportions guide the alignment without being imposed as hard capacities.

A.4 Algorithm Analysis

Algorithm 1 applies the complete intra-frame procedure to one frame. It forms non-empty low-rank groups, scans each group from its strongest loading, and suppresses only tokens that duplicate an already retained token in the same group. If suppression leaves fewer than B tokens, the highest-loading suppressed tokens are restored. Largest-remainder allocation then assigns an integer quota to every group and redistributes any quota that a small group cannot fill. The selected tokens are finally returned in their original patch order. Using $B = K$ produces $\mathcal{Q}_t$; using $B = M$ produces $\mathcal{C}_h$.

Algorithm 2 reduces a historical pool from M candidates to K tokens. For $h > 1$, it averages the normalized candidates in each reused group of $\mathcal{C}_{h-1}$ to obtain prototypes, solves the semi-relaxed alignment, and converts the aligned costs into token-wise novelty. The four utility cues are then computed for every candidate. MMR starts from the highest-utility token and adds candidates that are useful but not too similar to those already selected. For $h = 1$, novelty is zero because no preceding historical pool exists.

Algorithm 2 Inter-frame Temporal Selection

Require: $\mathcal{C}_h$, optional $\mathcal{C}_{h-1}$, $\mathcal{Q}_t$, instruction feature $\mathcal{L}$, budget K, iterations I, ε, τ, $(\alpha_c, \alpha_l, \alpha_s)$, λ, η
Ensure: $\mathcal{S}_h$, $|\mathcal{S}_h| = K$
if $h > 1$ **then**
  $\boldsymbol{\ell}_{h-1,r} \leftarrow |\mathcal{C}_{h-1}^{(r)}|^{-1} \sum_{\mathbf{x} \in \mathcal{C}_{h-1}^{(r)}} \mathbf{x}/\|\mathbf{x}\|_2$ ▷ build prototypes
  $A_{ir} \leftarrow 1 - \cos(\mathbf{x}_{h,i}, \boldsymbol{\ell}_{h-1,r}), \quad a_i \leftarrow 1/M, \quad \mu_r \leftarrow |\mathcal{C}_{h-1}^{(r)}|/M$
  $\mathbf{H} \leftarrow \exp(-\mathbf{A}/\varepsilon), \quad \mathbf{v} \leftarrow \mathbf{1}_R$
  for $s = 1, \ldots, I$ **do**
    $\mathbf{u} \leftarrow \mathbf{a} \oslash (\mathbf{H}\mathbf{v})$
    $\mathbf{v} \leftarrow [\boldsymbol{\mu} \oslash (\mathbf{H}^\top \mathbf{u})]^{\tau/(\tau+\varepsilon)}$
  end for
  $\mathbf{u} \leftarrow \mathbf{a} \oslash (\mathbf{H}\mathbf{v}), \quad \boldsymbol{\Pi}^\star \leftarrow \mathrm{diag}(\mathbf{u})\mathbf{H}\,\mathrm{diag}(\mathbf{v})$
  $\nu_{h,i} \leftarrow \sum_r \Pi_{ir}^\star A_{ir} / \sum_r \Pi_{ir}^\star, \quad i = 1, \ldots, M$ ▷ temporal novelty
else
  $\nu_{1,i} \leftarrow 0, \quad i = 1, \ldots, M$
end if
Compute $c_{h,i}$, $l_{h,i}$, and $s_{h,i}$ from $\mathcal{Q}_t$, $\mathcal{L}$, and $\mathbf{x}_{h,i}$ ▷ three task cues
$u_{h,i} \leftarrow \alpha_c c_{h,i} + \alpha_l l_{h,i} + \alpha_s s_{h,i} + \lambda \nu_{h,i}$ ▷ task-aware utility
$\mathcal{S}_h \leftarrow \{\arg\max_i u_{h,i}\}$
while $|\mathcal{S}_h| < K$ **do**
  $i^\star \leftarrow \arg\max_{i \notin \mathcal{S}_h} [\eta u_{h,i} - (1-\eta) \max_{j \in \mathcal{S}_h} \cos(\mathbf{x}_{h,i}, \mathbf{x}_{h,j})]$
  $\mathcal{S}_h \leftarrow \mathcal{S}_h \cup \{i^\star\}$
end while
return $\mathrm{Order}(\mathcal{S}_h)$

Both algorithms operate only on projected token features and their original indices. They require no gradient update, depth estimate, camera pose, or additional clustering stage.

B Experiment

B.1 Experimental Setups

B.1.1 Datasets

We evaluate on two continuous-environment VLN benchmarks, R2R-CE (Anderson et al., 2018; Krantz et al., 2020) and RxR-CE (Ku et al., 2020; Krantz et al., 2020). Both use photo-realistic Matterport3D indoor scans, but replace the discrete jumps between panoramic viewpoints with continuous motion in the Habitat simulator (Savva et al., 2019). The agent must therefore follow an instruction using low-level movement and turning actions while handling longer trajectories, partial observations, and possible collisions.

R2R-CE is the continuous counterpart of Room-to-Room (R2R). The original R2R benchmark contains 7,189 routes across 90 indoor environments, with three English instructions for each route (Anderson et al., 2018). Its VLN-CE release contains 10,819 training episodes, 778 validation-seen episodes, 1,839 validation-unseen episodes, and 3,408 test episodes (Krantz et al., 2020). R2R instructions are generally concise and describe relatively short routes. We report results on the validation-unseen split, whose 11 environments do not appear during training.

RxR-CE is derived from Room-Across-Room (RxR), which contains 126,069 instruction-path pairs based on approximately 16.5K paths (Ku et al., 2020). RxR provides English, Hindi, and Telugu instructions, together with dense temporal alignment between instruction words and annotator poses. Its routes are longer and more varied than those in R2R, and the instructions usually contain more detailed landmarks and intermediate actions. RxR-CE transfers these trajectories to Habitat for continuous control. Following the StreamVLN protocol (Wei et al., 2026), we evaluate on 1,517 validation-unseen episodes. Together, the two benchmarks test whether compression remains reliable for both concise goal-directed instructions and longer, landmark-rich navigation histories.

B.1.2 IMPLEMENTATION DETAILS

We insert the compressor after the frozen StreamVLN visual projector and before the navigation policy (Wei et al., 2026). No model parameter is trained or updated, and the StreamVLN checkpoint, action decoder, and VLN-CE evaluation protocol remain unchanged. Compression is applied independently at every policy step; selected tokens retain their original features, patch order within each frame, and temporal order across frames.

Each frame contains $N = 729$ visual tokens. For pruning ratios of 70%, 80%, and 90%, the policy receives $K = 218$, 146, and 72 tokens per frame, respectively. The corresponding historical candidate budgets are $M = 436$, 292, and 144. We set the utility weights to $\alpha_c = 0.60$, $\alpha_l = 0.20$, $\alpha_s = 0.30$, and $\lambda = 0.20$. For semi-relaxed OT, we use $\varepsilon = 0.08$, $\tau = 0.08$, and $I = 20$ Sinkhorn iterations. The MMR coefficient is $\eta = 0.65$. The retained rank and adaptive within-group threshold are fixed across both datasets and all pruning ratios; no benchmark-specific tuning is used.

Navigation results are computed on the official validation-unseen episodes using the standard VLN-CE evaluator. All experiments use an NVIDIA A100 GPU. Inference measurements use the same model checkpoint and token budgets as the navigation evaluation, so any difference is attributable to token selection rather than retraining or a change in the policy.

B.1.3 COMPARED METHODS

Unpruned StreamVLN (Wei et al., 2026) provides the reference performance without visual-token compression. We further compare with four training-free pruning methods. SparseVLM (Zhang et al., 2025c) uses language-conditioned relevance to remove less useful visual tokens. DivPrune (Alvar et al., 2025) emphasizes diversity among the retained tokens, while VisPruner (Zhang et al., 2025b) selects informative visual tokens without additional training. A-MMR (Wang et al., 2026a) is designed for VLN history compression: it uses the current view to score historical tokens and balances relevance against redundancy. All methods are compared at the same retained-token budgets. The unpruned and baseline results in Table 1 are reported by Wang et al. (Wang et al., 2026a); a fully controlled reproduction additionally requires identical checkpoints, evaluators, history configurations, and decoding settings.

B.1.4 EVALUATION METRICS

Success rate (SR) is the percentage of episodes in which the agent stops within the goal threshold. Success weighted by path length (SPL) additionally penalizes unnecessarily long trajectories. Oracle success rate (OS) records whether the trajectory enters the goal region at any point, regardless of the final stopping location. Navigation error (NE) is the geodesic distance, in meters, from the final position to the goal. Normalized dynamic time warping (nDTW) measures agreement with the reference trajectory while accounting for the order of visited locations (Ilharco et al., 2019). We report SR, SPL, OS, and NE on R2R-CE, and SR, SPL, nDTW, and NE on RxR-CE. All metrics except NE are percentages; higher SR, SPL, OS, and nDTW are better, whereas lower NE is better.

B.2 ADDITIONAL EXPERIMENTS

B.2.1 SENSITIVITY TO SINKHORN ITERATIONS

We evaluate the number of semi-relaxed Sinkhorn iterations $I \in \{5, 10, 20, 30, 50\}$ on the R2R-CE validation-unseen split with 72 retained tokens. As shown in Figure 4, $I = 20$ achieves the best result on all four metrics, reaching 51.77% SR, 39.10% SPL, 71.72% OS, and 5.229 m NE. Increasing the iteration count beyond 20 provides no further gain while incurring additional computation. We therefore use $I = 20$ in all experiments.

B.2.2 ADDITIONAL RESULTS ON NAVILA

To examine whether the proposed compressor transfers across navigation policies, we further integrate it into NaVILA (Cheng et al., 2025) and compare it with DivPrune (Alvar et al., 2025) and VisPruner (Zhang et al., 2025b) under identical token budgets. Table 5 reports the R2R results.

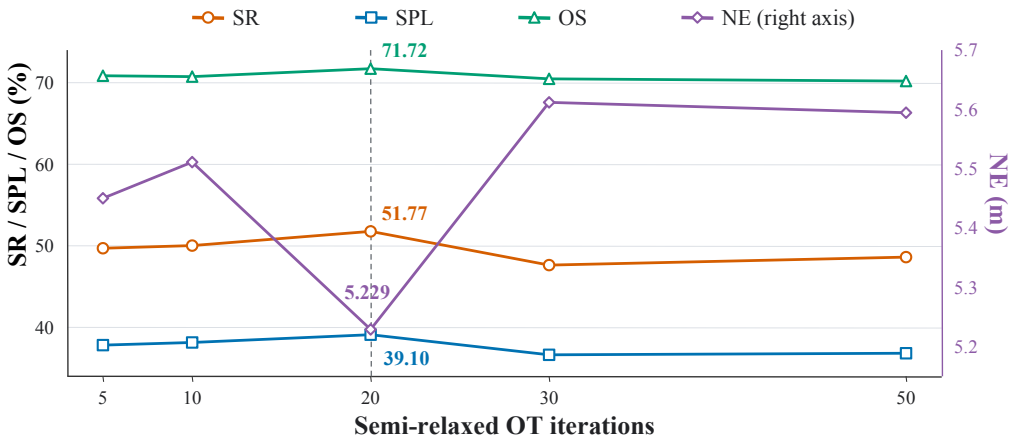


Figure 4: Sensitivity to the number of semi-relaxed Sinkhorn iterations on the R2R-CE validation-unseen split with 72 retained tokens. Higher values are better for SR, SPL, and OS, whereas lower values are better for NE.

Table 5: Navigation performance on R2R using the NaVILA backbone. The unpruned result is reported by NaVILA (Cheng et al., 2025); all pruning methods are evaluated with the corresponding per-frame token budget. Bold values indicate the best pruning result within each budget.

Method	Venue	SR (%)↑	SPL (%)↑	OS (%)↑	NE↓
Unpruned · 729 tokens per frame					
Unpruned	RSS'25	54.00	49.00	62.50	5.22
70% pruning · 218 tokens per frame					
DivPrune (Alvar et al., 2025)	CVPR'25	0.11	0.10	4.24	9.71
VisPruner (Zhang et al., 2025b)	ICCV'25	27.51	20.71	**47.41**	7.70
Ours	-	**30.40**	**25.48**	46.55	**6.93**
80% pruning · 146 tokens per frame					
DivPrune (Alvar et al., 2025)	CVPR'25	0.76	0.74	4.57	9.69
VisPruner (Zhang et al., 2025b)	ICCV'25	18.81	13.80	41.44	8.19
Ours	-	**27.57**	**22.50**	**46.82**	**7.38**
90% pruning · 72 tokens per frame					
DivPrune (Alvar et al., 2025)	CVPR'25	0.11	0.10	4.24	9.71
VisPruner (Zhang et al., 2025b)	ICCV'25	9.41	6.07	41.60	8.78
Ours	-	**17.46**	**12.00**	**46.49**	**8.39**

Our method achieves the best SR and SPL at all three pruning ratios. Relative to the strongest pruning baseline, VisPruner, it improves SR and SPL by 6.57 and 6.47 percentage points on average, respectively. The advantage increases under tighter budgets: at 90% pruning, SR rises by 8.05 points and SPL by 5.93 points, nearly doubling the retained SPL (97.7% relative), while NE decreases by 0.39. These results indicate that the proposed semantic coverage and temporal selection remain effective when transferred from StreamVLN to a different navigation policy.

B.2.3 Additional Real-World Deployment Results

Figures 5 and 6 provide additional Unitree Go2 trajectories collected under the same deployment setup as the main paper. These examples cover further indoor routes and instructions, complementing the representative results in Figure 3.

C Limitations and Future Work

Although the proposed compressor improves the navigation and efficiency trade-off without retraining, it has several limitations. First, all evaluated pruning methods operate on the same eight frames sampled uniformly along the trajectory. This provides a controlled comparison, but uniform sam-

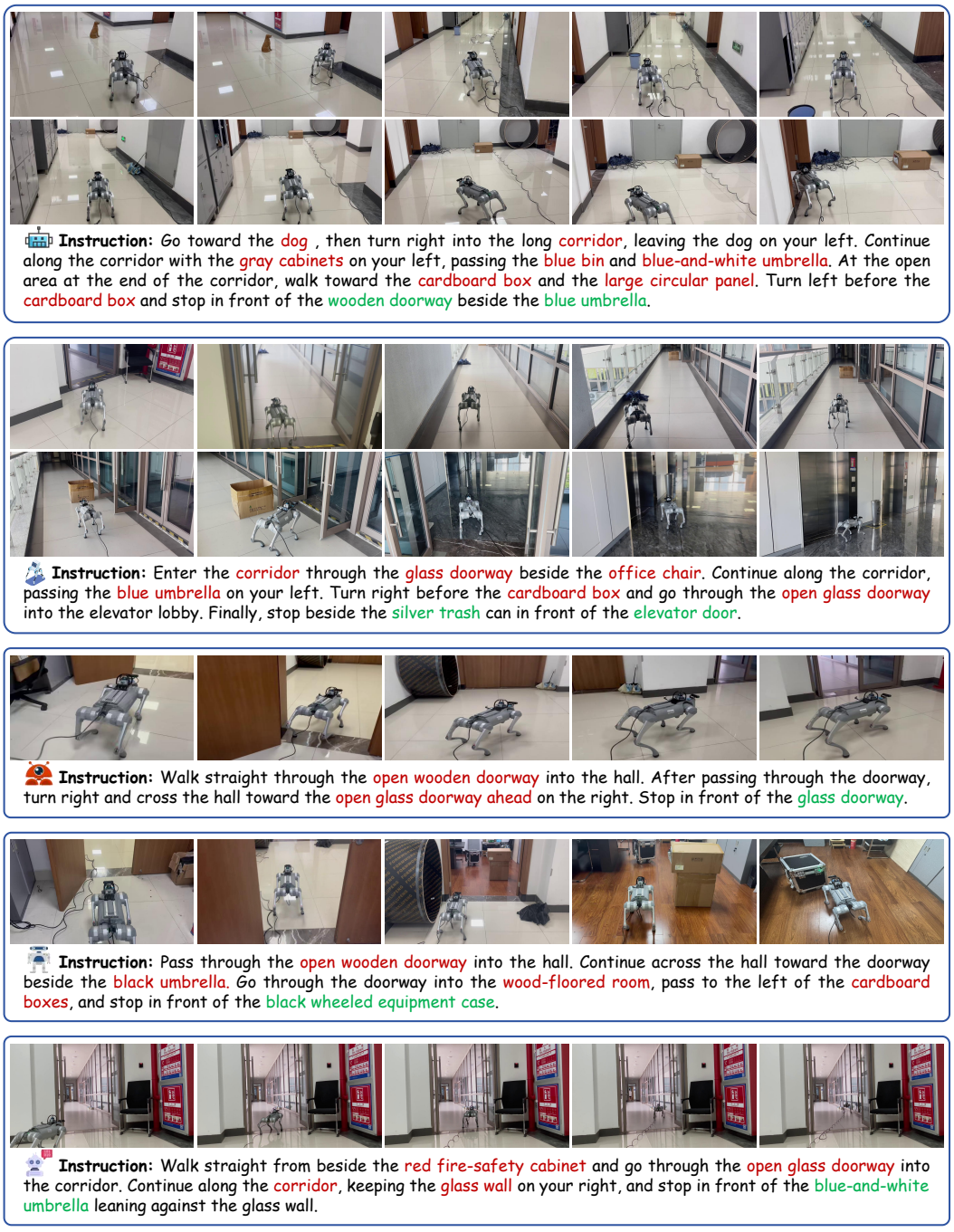


Figure 5: Additional real-world deployment results on the Unitree Go2. Each row shows sampled observations from one instruction-following trajectory under the same hardware and inference setup used in the main experiments.

pling treats every time interval equally. A landmark may appear only briefly between two sampled frames; if this key frame is omitted, subsequent token pruning cannot recover its evidence. Future work should therefore combine informative-frame selection with token pruning, so the history first retains decisive observations and then removes redundant tokens within them.

Second, our temporal module compares each historical frame only with group prototypes from the immediately preceding sampled frame. This local comparison is efficient, but it may miss redun-

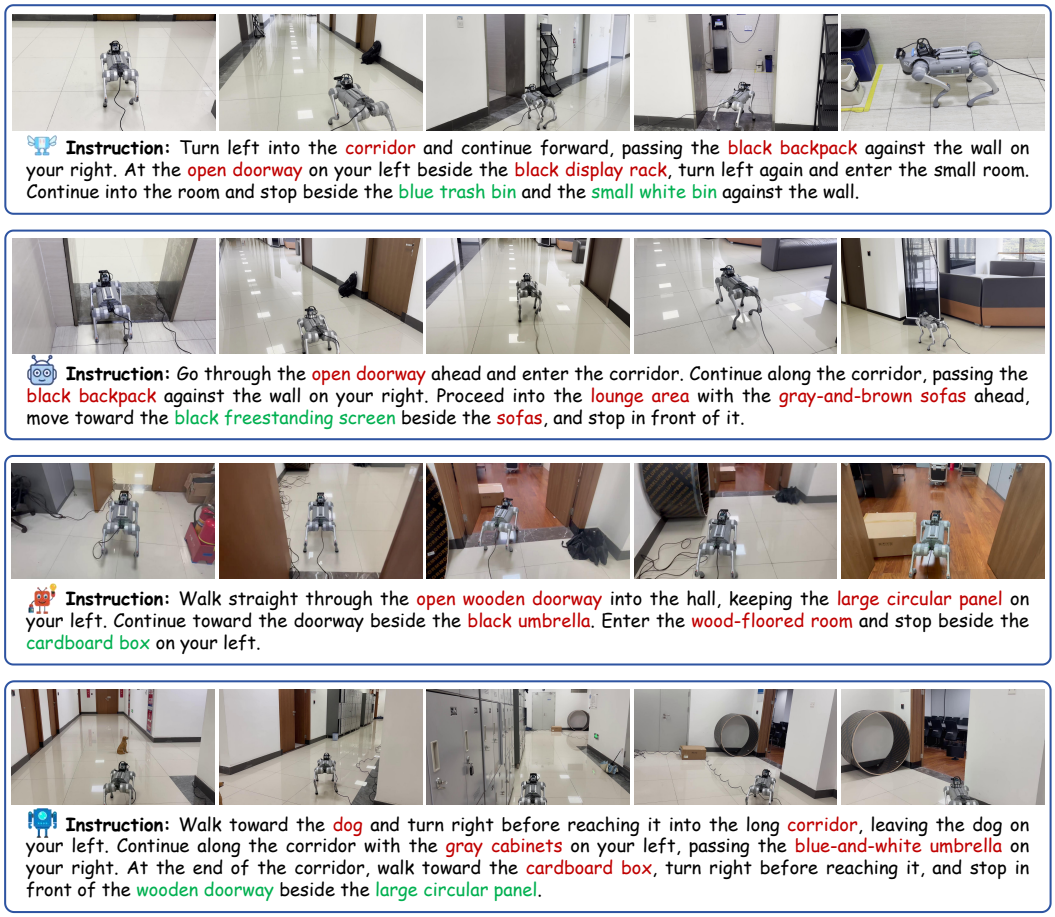


Figure 6: Additional real-world deployment results on the Unitree Go2 (continued). Each row shows sampled observations from one instruction-following trajectory under the same hardware and inference setup used in the main experiments.

dancy that reappears after a longer interval, while averaging a group into one prototype can weaken rare details. Moreover, appearance-based alignment may confuse visually similar locations or treat viewpoint and illumination changes as new content. Multi-timescale memory and optional geometric or uncertainty cues could address these cases without making pose or depth mandatory inputs.

Finally, instruction relevance is computed from a single representation of the complete instruction. It does not explicitly determine which sub-instruction is active at the current navigation step, such as leaving a room or turning at a particular landmark. Step-aware language grounding could connect historical evidence more precisely to the agent's immediate objective. The grouping and transport stages also introduce preprocessing overhead. Because physical robots require both accurate decisions and real-time responses, reducing this cost without losing critical visual evidence remains an important direction.