VLN-Cache: Token Caching for VLN Models with Visual/Semantic Dynamics Awareness

***Abstract*— This is abstract.**

I. INTRODUCTION

Vision-and-Language Navigation (VLN) enables embodied agents to follow natural language instructions in complex, unstructured environments [1], [2]. It has become a mainstream paradigm in the field of embodied intelligence [3], [4]. However, the computational demands of VLN models fundamentally conflict with the real-time navigation requirements of practical robotic deployment [5].

Existing VLN acceleration approaches primarily follow three directions: efficient architectures, model compression, and runtime optimization. Efficient architectures design lightweight networks to reduce inherent complexity [6]–[8]. Compression techniques like quantization shrink model size while preserving predictive capabilities [9], [10]. Runtime optimization dynamically reduces computations during inference without altering model parameters [11]–[13]. These methods have substantially improved VLN inference efficiency.

Token caching has emerged as a highly promising training-free runtime optimization technique.

Token caching exploits temporal coherence: background regions such as walls and floors change little across adjacent frames [14]–[17]. Existing token caching methods follow a detection-reuse paradigm comprising three typical strategies [14], [15], [18]. (1) Content-based similarity detection identifies static tokens via cosine similarity or L2 distance [14], [18]. (2) Task-guided attention filtering uses cross-attention scores or saliency maps to suppress task-irrelevant regions [15], [17]. (3) Heuristic reuse strategies apply fixed thresholds or reuse ratios across layers [16], [19]. These methods have achieved remarkable speedup with negligible accuracy loss in fixed camera or static task scenarios [14], [19].

However, existing token caching methods are built on a static-view assumption: patches at the same image position are expected to remain similar across adjacent frames [14]. This assumption breaks in VLN, where the agent continuously translates and rotates during navigation [20]. As a result, physically static objects can shift substantially in image coordinates, making position-wise matching underestimate reusable tokens. In addition, navigation is instruction-conditioned and temporally staged. A landmark that is critical before a turn can quickly become irrelevant after the agent passes it, so semantic relevance changes even when appearance is stable. These two dynamics jointly limit the direct transfer of static-scene caching strategies to embodied navigation.

To quantify these effects, we analyze patch similarity under ego-motion and track language-vision attention on landmark regions along VLN trajectories. The first finding is a large group of pseudo-dynamic tokens: background regions that are physically unchanged but appear dissimilar due to viewpoint shifts. The second finding is temporally varying landmark relevance: attention rises when approaching a landmark and drops after passing it. Together, these observations suggest two requirements for reliable caching in VLN: viewpoint-aware matching for visual reuse and online semantic refresh for instruction-critical tokens.

Motivated by these findings, we propose *VLN-Cache*, a dual-aware token caching framework for VLN. For visual dynamics, we perform view-aware feature alignment before cross-frame matching, which recovers reusable tokens that would be missed by position-wise comparison. For semantic dynamics, we introduce instruction-guided saliency monitoring that refreshes cached tokens when their task relevance changes. To control the additional overhead, we further adopt an entropy-based layer-adaptive reuse schedule that allocates conservative reuse to sensitive layers and more aggressive reuse to stable layers. The resulting framework delivers practical acceleration while preserving navigation performance.

Our key contributions are as follows.

- We provide empirical evidence that static-scene caching assumptions fail in VLN due to viewpoint-induced mismatch and temporal semantic shift.
- We present *VLN-Cache*, a dual-aware framework that combines view-aware matching with instruction-guided semantic refresh, without architectural changes or retraining.
- We design an entropy-based layer-adaptive reuse strategy

Fig. 1: Overview Placeholder for Observation and Motivation in VLN token caching.

to balance acceleration gain and computational overhead across transformer layers.

Experiments on multiple VLN benchmarks show substantial speedup while maintaining competitive success rates, supporting practical real-time embodied deployment.

II. Preliminary

A. Vision-Language Navigation

Vision-and-Language Navigation (VLN) requires an embodied agent to execute an instruction L while moving in an unseen 3D environment [1], [2]. At navigation step t, the agent receives observation o_t and maintains trajectory history H_t. Action prediction can be written as:

$$a_t = \arg\max_a P\left(a \mid o_t, L, H_t; \theta\right), \tag{1}$$

where θ denotes model parameters. In VLA-style VLN models, each observation is encoded as visual tokens $\{v_t^{(i)}\}_{i=1}^{M}$ and fused with language features for autoregressive reasoning. The model output can be represented as low-level control or waypoint actions, depending on the navigation formulation.

B. Token Caching

Token caching accelerates inference by reusing cross-frame states for redundant tokens [14]. Given adjacent observations, a static-paradigm decision rule is

$$\hat{K}_t^{(i)} = \begin{cases} K_{t-1}^{(i)}, & \text{if sim}\left(p_t^{(i)}, p_{t-1}^{(i)}\right) > \tau, \\ f_K\left(v_t^{(i)}\right), & \text{otherwise}, \end{cases} \tag{2}$$

where $p_t^{(i)}$ is the feature of token i, τ is a reuse threshold, and $f_K(\cdot)$ computes updated keys. The same reuse decision is applied to key-value states across transformer layers. Let $\mathcal{C}_t^{\ell}$ be the reused token set at layer ℓ and step t. Its reuse ratio is

$$r_t^{\ell} = \frac{|\mathcal{C}_t^{\ell}|}{M}, \tag{3}$$

with M visual tokens per step. This static formulation serves as the direct baseline transferred to VLN in our study.

III. Analysis of Visual/Semantic Dynamics

A. Analysis of Visual Dynamics in VLN Inference

B. Analysis of Semantic Dynamics is VLN Inference

IV. VLN-Cache Framework

A. Visual-Dynamic-Aware xxx

1) View-Aware Feature Alignment:

2) Cross-Frame Token Reusing:

B. Semantic-Dynamic-Aware xxx

1) Instruction-Guided Saliency Filtering:

2) Layer-Adaptive Token Reusing:

V. Experiments

A. Setup

B. Main Results

C. Ablation Study

D. Discussion

VI. Conclusion

References

[1] J. Gu, E. Stefani, Q. Wu, J. Thomason, and X. Wang, "Vision-and-language navigation: A survey of tasks, methods, and future directions," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022, p. 7606–7623. [Online]. Available: http://dx.doi.org/10.18653/v1/2022.acl-long.524

[2] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," 2018. [Online]. Available: https://arxiv.org/abs/1711.07280

[3] Y. Du, T. Fu, Z. Chen, B. Li, S. Su, Z. Zhao, and C. Wang, "Vl-nav: Real-time vision-language navigation with spatial reasoning," 2025. [Online]. Available: https://arxiv.org/abs/2502.00931

[4] J. Lee, H. Shin, and J. Ko, "Iros: A dual-process architecture for real-time vlm-based indoor navigation," 2026. [Online]. Available: https://arxiv.org/abs/2601.21506

[5] D. Kang, A. Perincherry, Z. Coalson, A. Gabriel, S. Lee, and S. Hong, "Harnessing input-adaptive inference for efficient vln," 2025. [Online]. Available: https://arxiv.org/abs/2508.09262

[6] D. Zheng, S. Huang, Y. Li, and L. Wang, "Efficient-vln: A training-efficient vision-language navigation model," 2025. [Online]. Available: https://arxiv.org/abs/2512.10310

[7] S. Ye, S. Mao, Y. Cui, X. Yu, S. Zhai, W. Chen, S. Zhou, R. Xiong, and Y. Wang, "Etp-r1: Evolving topological planning with reinforcement fine-tuning for vision-language navigation in continuous environments," 2025. [Online]. Available: https://arxiv.org/abs/2512.20940

[8] J. Hou, Y. Xiao, X. Xue, and T. Zeng, "Log-nav: Efficient layout-aware object-goal navigation with hierarchical planning," 2025. [Online]. Available: https://arxiv.org/abs/2505.06131

[9] J. Zhu, Y. Qiao, S. Zhang, X. He, Q. Wu, and J. Liu, "Minivln: Efficient vision-and-language navigation by progressive knowledge distillation," 2024. [Online]. Available: https://arxiv.org/abs/2409.18800

[10] A.-C. Cheng, Y. Ji, Z. Yang, Z. Gongye, X. Zou, J. Kautz, E. Bıyık, H. Yin, S. Liu, and X. Wang, "Navila: Legged robot vision-language-action model for navigation," 2025. [Online]. Available: https://arxiv.org/abs/2412.04453

[11] W. Qin, A. Burns, B. A. Plummer, and M. Betke, "Walk and read less: Improving the efficiency of vision-and-language navigation via tuning-free multimodal token pruning," 2025. [Online]. Available: https://arxiv.org/abs/2509.15250

[12] H. Hu, L. Huang, X. Wang, Y. Cui, S. Wu, N. Guan, and C. J. Xue, "Nav-ee: Navigation-guided early exiting for efficient vision-language models in autonomous driving," 2025. [Online]. Available: https://arxiv.org/abs/2510.01795

[13] J. Zhang, K. Wang, S. Wang, M. Li, H. Liu, S. Wei, Z. Wang, Z. Zhang, and H. Wang, "Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks," 2025. [Online]. Available: https://arxiv.org/abs/2412.06224

[14] S. Xu, Y. Wang, C. Xia, D. Zhu, T. Huang, and C. Xu, "Vla-cache: Efficient vision-language-action manipulation via adaptive token caching," 2025. [Online]. Available: https://arxiv.org/abs/2502.02175

[15] S. Ding, P. Zhao, X. Zhang, R. Qian, H. Xiong, and Q. Tian, "Prune spatio-temporal tokens by semantic-aware temporal accumulation," 2023. [Online]. Available: https://arxiv.org/abs/2308.04549

[16] S. Chang, P. Wang, M. Lin, F. Wang, D. J. Zhang, R. Jin, and M. Z. Shou, "Making vision transformers efficient from a token sparsification view," 2023. [Online]. Available: https://arxiv.org/abs/2303.08685

[17] J. Li, K. Li, C. Gao, Y. Li, and X. Chen, "Egoprune: Efficient token pruning for egomotion video reasoning in embodied agent," 2025. [Online]. Available: https://arxiv.org/abs/2507.15428

[18] D. Bolya, C.-Y. Fu, X. Dai, P. Zhang, C. Feichtenhofer, and J. Hoffman, "Token merging: Your vit but faster," 2023. [Online]. Available: https://arxiv.org/abs/2210.09461

[19] Y. Wei, J. Fan, J. Guo, R. Zhen, R. Shao, X. Su, Z. Xie, and S. Yang, "Learning to accelerate vision-language-action models through adaptive visual token caching," 2026. [Online]. Available: https://arxiv.org/abs/2602.00686

[20] J. Q. Sun, X. Xing, H. Weng, C. M. Yeum, and M. Crowley, "View invariant learning for vision-language navigation in continuous environments," 2025. [Online]. Available: https://arxiv.org/abs/2507.08831