HARNESSING INPUT-ADAPTIVE INFERENCE FOR EFFICIENT VISION-AND-LANGUAGE NAVIGATION

Dongwoo Kang, Akhil Perincherry, Zachary Coalson,
Aiden Gabriel, Stefan Lee, and Sanghyun Hong

Oregon State University

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Oregon State University

TRUE Trustworthy and Responsible AI Lab

VLN MODELS ARE COMPUTATIONALLY DEMANDING

- Modern VLN agents employ large vision and language encoders
  - Initial model size[1]: 73M parameters
  - Current size[2,3,4]: 160–181M parameters

- These agents are compute-intensive
  - An average episode ≈ same compute as generating 5 tokens with GPT-3
  - Prevents deployment in resource-constrained environments

Walk toward the railing and right past the stairs. Walk into the living room and turn right. Stop by the end table.[1]

[1]Anderson et al., *Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments*, CVPR 2018.
[2]Chen et al., *History Aware Multimodal Transformer for Vision-and-Language Navigation*, NeurIPS 2021.
[3]Chen et al., *Think Global, Act Local: Dual-scale Graph Transformer for Vision-and-Language Navigation*, CVPR 2022.
[4]Hong et al., *A Recurrent Vision-and-Language BERT for Navigation*, CVPR 2021.

Our approach: input-adaptive inference

- Insight: not every input requires the same amount of compute

★ = correct navigation decision

OUR APPROACH: INPUT-ADAPTIVE INFERENCE – CONT'D

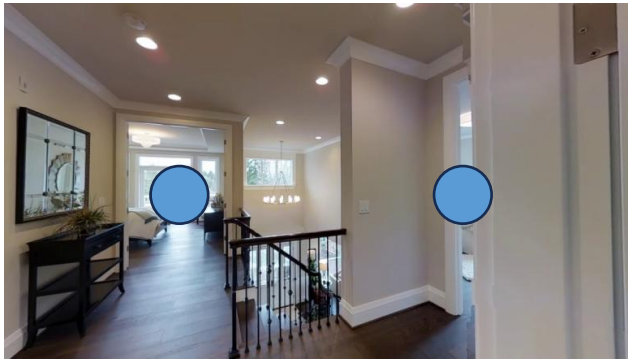
- Input-adaptive inference[1,2,3,4]:
  - Stop forward pass when internal decision converges
  - 2–3× computational savings with small performance loss

- Adaption to VLN:
  - Candidate model: visual encoder
    - Accounts for >99.5% of computations
  - Existing techniques[3] lead to inaccurate navigation

[1]Huang et al., *Multi-Scale Dense Networks for Resource Efficient Image Classification*, ICLR 2018.
[2]Liu et al., *FastBERT: a Self-distilling BERT with Adaptive Inference Time*, ACL 2020.
[3]Tang et al., *You Need Multiple Exiting: Dynamic Early Exiting for Accelerating Unified Vision Language Model*, CVPR 2023.
[4]Xin et al., *DeeBERT: Dynamic Early Exiting for Accelerating BERT Inference*, ACL 2020.

HARNESSING SPATIAL LOCALITY

- *k*-extension
  - Motivation: not all views in a panorama are navigable
    - Processing only navigable views degrades performance
  - We process each navigable view + *k* adjacent views
  - Adjacent views are *early-exited*
    - Adaptive exit thresholds based on proximity to navigable view

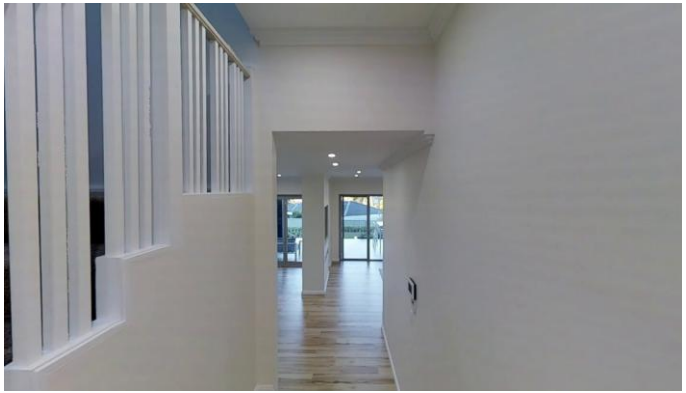


● = navigable view

HARNESSING TEMPORAL LOCALITY

- Locality-sensitive (view) hashing
  - Motivation: an agent encounters similar views across navigation steps
  - Storage: Hash RGB inputs; store processed embeddings
  - Retrieval: Fetch embedding from bucket with most similar RGB
    - Similarity threshold to filter out bad candidates

nav. step



RESULT HIGHLIGHT: PERFORMANCE AND COMPUTATIONS NEEDED

- Setup:
  - Environments: standard and continuous
  - Tasks: R2R, R2R-Back, R2R-Last, REVERIE, CVDN, SOON, and R2R-CE
  - Agents: HAMT, DUET, and VLN-CE-BERT
  - Metrics:
    - Performance: Success Rate (SR)
    - Efficiency: GFLOPs
- Results:
  - On R2R: 59–60% reduction in GFLOPs with 9–12% drop in SR
  - On R2R-CE: 87% reduction in GFLOPs with 8% drop in SR

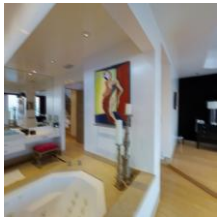
[1]Krantz et al., *Beyond the Nav-Graph: Vision-and-Language Navigation in Continuous Environments*, ECCV 2020.

RESULT HIGHLIGHT: NAVIGATION UNDER VISUAL CORRUPTIONS

- Setup
  - Standard environment
  - HAMT on R2R
  - Metrics:
    - Performance: SR
    - Efficiency: GFLOPs

- Results:
  - Our agent is vulnerable to speckle noise: 31% SR loss with 21% increase in GFLOPs
  - Denoising recovers SR by 18% and reduces GFLOPs by 6%

Clean


Motion Blur

Speckle Noise

Low Lighting

Spatter

Defocus Blur

THANK YOU!

Zachary Coalson

See You All at Our Poster Session!
Exhibit Hall I #761 | Tuesday @ 6:15PM

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Oregon State University

TRUE Trustworthy and Responsible AI Lab