NATIONAL CENTER FOR UNDERSTANDING FUTURE
TRAVEL BEHAVIOR AND DEMAND

Final Project Report

Exploring Top-Down Visual Attention for Transportation Behavior Analysis

BY

Bilal Abdulrahman[1]
Email: babdulrahman@gc.cuny.edu

Gong Qi Chen[1]
Email: gongqi.chen14@gc.cuny.edu

Zhigang Zhu[1,2]
Email: zzhu@ccny.cuny.edu

Alison Conway[2]
Email: aconway@ccny.cuny.edu

[1]Department of Computer Science, The CUNY Graduate Center
365 Fifth Avenue, New York, NY 10016
[2]Grove School of Engineering, The City College of New York
160 Covent Avenue, New York, NY 10031

May 2026

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. N/A	**2. Government Accession No.** N/A	**3. Recipient's Catalog No.** N/A	
4. Title and Subtitle Exploring Top-Down Visual Attention for Transportation Behavior Analysis: Probing Top-Down Attention Mechanisms in Vision-Language Models for Pedestrian Behavior Analysis		**5. Report Date** May 2026 **6. Performing Organization Code** N/A	
7. Author(s) Bilal AbdulRahman, https://orcid.org/0000-0002-1164-1976 Gong Qi Chen, https://orcid.org/0009-0007-8415-2915 Alison Conway, Ph.D. https://orcid.org/0000-0002-2538-4501 Zhigang Zhu, Ph.D. https://orcid.org/0000-0002-9990-1137		**8. Performing Organization Report No.** N/A	
9. Performing Organization Name and Address Grove School of Engineering The City College of New York 160 Convent Avenue, New York, NY 10031		**10. Work Unit No. (TRAIS)** N/A **11. Contract or Grant No.** 69A3552344815 and 69A3552348320	
12. Sponsoring Agency Name and Address U.S. Department of Transportation, University Transportation Centers Program, 1200 New Jersey Ave, SE, Washington, DC 20590		**13. Type of Report and Period Covered** Final Report, 2024-2026 **14. Sponsoring Agency Code** USDOT OST-R	
15. Supplementary Notes N/A			
16. Abstract Human action recognition is relevant to a number interactions that occur in transportation systems, including those between a driver and their vehicle, between drivers and pedestrians, and between other humans and existing vehicles or traffic control devices. As action recognition systems scale to handle complex urban environments, understanding exactly where a model focuses its attention is critical for ensuring reliability and fairness. This study proposes a framework for integrating top-down visual attention for transportation applications, focusing on the case of pedestrians navigating complex environments such as signalized intersections. The project developed a specialized probing tool for Vision-Language Models (VLMs) to visualize cross-modal attention maps and employs a bio-inspired feedback mechanism to enhance and improve the attention mechanism in the vision backbone. The study established a framework to track how text prompts guide visual focus. The core technical achievement is an algorithmic pipeline that accurately maps 1D sequence tokens back to 2D pixel coordinates, overcoming the spatial distortions caused by aggressive image padding, resizing, and sliding-window cropping. This tool significantly enhances the interpretability of VLMs used in transportation applications.			
17. Key Words Vision-Language Models; Top-Down Attention; Pedestrian Behavior; Interpretability; Neural Networks		**18. Distribution Statement** No restrictions.	
19. Security Classif.(of this report) Unclassified	**20. Security Classif.(of this page)** Unclassified	**21. No. of Pages** 24	**22. Price** N/A

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, under Grant No. 69A3552344815 and 69A3552348320 from the U.S. Department of Transportation's University Transportation Centers Program. The U.S. Government assumes no liability for the contents or use thereof.

ACKNOWLEDGMENTS

This research was partially supported by the National Center for Understanding Future Travel Behavior and Demand (TBD), a National University Transportation Center sponsored by the U.S. Department of Transportation (USDOT) under grant numbers 69A3552344815 and 69A3552348320. The authors would like to thank the TBD National Center, USDOT, The City College of New York and the CUNY Graduate Center for their support of university-based research in transportation, particularly for the funding provided for this project.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

EXECUTIVE SUMMARY

Exploring top-down visual attention for transportation behavior analysis, this study stands at the intersection of cognitive psychology, AI and computer vision, and transportation safety and efficiency. Human action recognition is relevant to a number interactions that occur in transportation systems, including those between a driver and their vehicle, between drivers and pedestrians, and between other humans and existing vehicles or traffic control devices.

As action recognition systems scale to handle complex urban environments, understanding exactly where a model focuses its attention is critical for ensuring reliability and fairness. This study proposes a framework for integrating top-down visual attention for transportation applications. In implementing the idea within a one-year timeframe, after a comprehensive survey on human action recognition for all applications, we focused on applying the proposed framework to the case of pedestrians navigating complex environments such as signalized intersections.

The project developed a specialized probing tool for Vision-Language Models (VLMs) to visualize cross-modal attention maps. We also employ a bio-inspired feedback mechanism to enhance and improve the attention mechanism in the vision backbone. We established a framework to track how text prompts guide visual focus.

The core technical achievement is an algorithmic pipeline that accurately maps 1D sequence tokens back to 2D pixel coordinates, overcoming the spatial distortions caused by aggressive image padding, resizing, and sliding-window cropping. This tool significantly enhances the interpretability of VLMs used in transportation applications, including analyzing pedestrian behavior. We also employ this probing tool to compare our feedback mechanism with baseline.

INTRODUCTION

Action recognition technologies have profound implications for intelligent human-machine systems, particularly in the assessment of urban mobility and transportation safety. The capacity to accurately identify and understand human actions holds substantial promise across diverse sectors, including autonomous driving, automated surveillance, and automated behavioral analysis for street design and traffic control applications. Modern methodologies predominantly leverage deep learning models to autonomously learn these features from massive datasets. However, as these models are increasingly deployed in high-stakes environments, such as monitoring pedestrian interactions at crosswalks or navigating the intricate "sidewalk friction" created by varying crowd densities, the "black-box" nature of deep neural networks remains a significant hurdle for trust and safety validation.

To be specific, this project provides two unique contributions to the community of using AI for transportation applications. First, a critical component of the proposed multimodal (text + image) architecture is the integration of attention mechanisms, which mimic the human cognitive ability to focus on specific aspects of an environment while ignoring irrelevant distractors. Despite their efficacy, the internal decision-making processes of large-scale Vision-Language Models (VLMs) are often opaque. This project focuses on the development of a probing framework designed to interrogate these mechanisms within a custom VLM pipeline. By mapping text tokens back onto visual attention layers, we can systematically evaluate how models perceive complex street scenes. This layer-by-layer analysis allows us to isolate which components most significantly impact the model's response to specific queries, bridging the gap between raw pixel data and high-level linguistic output.

Second, to further enhance the perceptual accuracy of the system, we introduce a bio-inspired feedback mechanism into the vision backbone by adopting the **Analysis-by-Synthesis Vision Transformer (AbsViT)** (Shi et al., 2023) within the **Molmo** framework (Deitke, et al, 2024). This is illustrated in Figure 1. Unlike standard VLMs where the vision encoder acts as a passive, feedforward feature extractor, our implementation uses a top-down pathway to **condition the vision backbone on the text prompt**. This ensures that the linguistic goal is communicated back to the vision encoder during the feature extraction process. By providing the backbone with information about the prompt, the model can selectively prioritize visual features that are semantically relevant to the task at hand. This goal-directed conditioning improves the quality of visual tokens before they are even processed by the LLM. To ensure robust performance, we pretrained this modified backbone on the **VQAv2** dataset (Goyal et al., 2017), allowing the model to learn the intricate relationship between query-driven attention and visual feature selection.

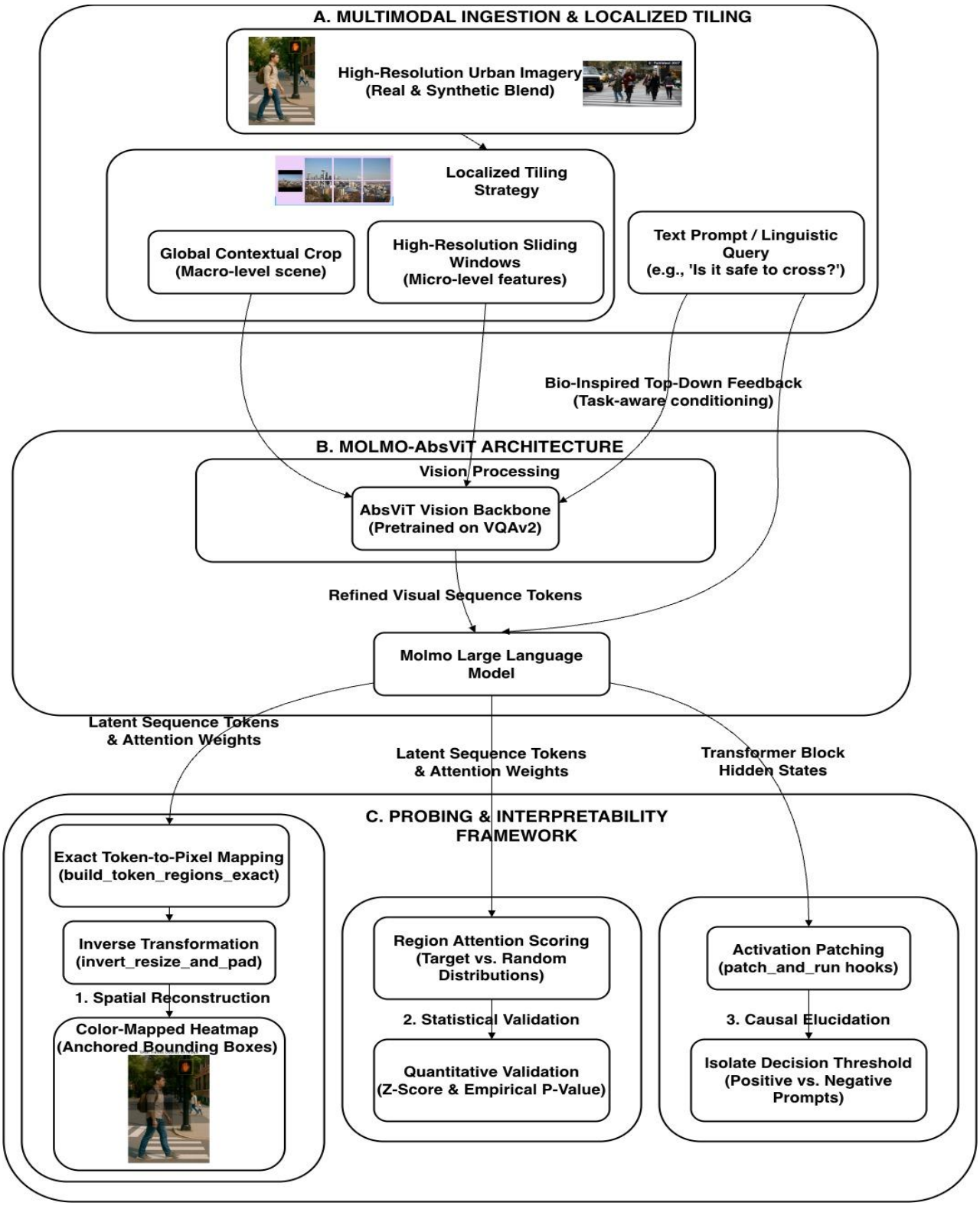

Figure 1. A probing framework in a proposed Molmo-AbsViT architecture designed to interrogate these mechanisms within a custom VLM

To summarize, the rapid scaling of VLM parameters has led to significant computational overhead, often making real-time deployment in urban environments challenging. Our probing method serves a dual purpose: in addition to providing explainability, it identifies redundant layers that contribute minimally to the final inference for a specific downstream task. By leveraging these insights, we can implement targeted pruning strategies that reduce the model's footprint with minimal performance loss. Furthermore, by injecting tokens from contrasting prompt inferences, such as positive and negative prompts, we further elucidate the model's internal decision-making boundaries. Finally, qualitative comparisons between the baseline Molmo (Deitke, et al, 2024) and our modified backbone demonstrate that the inclusion of bio-inspired feedback leads to more focused visual attention and minor qualitative improvements in output text, providing a more reliable and efficient framework for urban action recognition.

LITERATURE REVIEW

This focused literature review is based on a more comprehensive survey on human action recognition and the use of attention mechanism (AbdulRahman et al, 2026), which was also completed as part of this project. Research on pedestrian behaviors and urban walkability spans multiple domains, typically bifurcating into studies of **urban form/infrastructure** or **human factors/technology**. While both fields have advanced significantly, a critical gap remains in integrating high-level semantic understanding with low-level visual attention, a gap that modern Vision-Language Models (VLMs) are only beginning to address.

Urban Design and Walkability Metrics

Traditional planning and urban design studies have established various walkability indices to quantify how built environments influence movement. For instance, metrics focusing on network connectivity and proximity have been validated in cities like Rome and Venice (Gori, Nigro, & Petrelli, 2014). Similarly, the Pedestrian Accessibility Tool (Erath et al., 2017) incorporates detailed design attributes, such as crossings and sidewalk widths, into GIS-based walkability analyses. While these tools assist planners in evaluating safety and Level of Service (Ryus et al., 2022), they often treat the urban environment as a static set of features, failing to account for the dynamic, real-time "sidewalk friction" caused by moving obstacles like opposing pedestrians, hand carts, or varying group densities.

Data-Driven and AI-Based Behavioral Analysis

The rise of deep learning has shifted the focus toward automated behavioral analysis. LiDAR-based systems have been used at intersections to collect granular data on walking speeds and start-up delays (Li et al., 2023), revealing that actual pedestrian behaviors frequently diverge from guideline assumptions. In parallel, computer vision (CV) models have been deployed on traffic camera feeds to autonomously detect and count pedestrians and bicyclists (Pourhomayoun, 2020), while autonomous vehicle research utilizes vision-based models to predict trajectories by inferring intent from visual context (Zhong et al., 2023).

However, as noted in the development of our **Molmo-based** framework (Deitke, et al, 2024), these models often function as "black boxes." While they are capable of high-accuracy detection, they lack **explainability**. They cannot easily articulate which visual features (e.g., a specific obstacle or a distant vehicle) triggered a particular prediction. This highlights the necessity for probing methods that can map model attention back to specific image regions layer-by-layer.

Human-Centered Studies and Visual Attention

Human-centered research focuses on the cognitive mechanics of decision-making. Cognitive modeling, combined with reinforcement learning, has been used to explain how pedestrians decide to cross streets under visual uncertainty (Wang & Srinivasan, 2024). Experimental psychology approaches, utilizing eye-tracking and Virtual Reality (VR), show that pedestrians allocate gaze strategically to navigate obstacles (de Winter et al., 2021). Findings suggest that distraction, such as texting while walking, significantly reduces situational awareness and prolongs gaze fixation (Krishna & Choudhary, 2025). Furthermore, highly salient distractors, like car body lights, can capture top-down focus, disrupting safety-critical tasks (Yi, Zhao, & Lin, 2024).

The Gap: Integrating Bio-Inspired Feedback and VLMs

Despite these advancements, a disconnect persists. Standard CV architectures in VLMs are typically query-agnostic. The vision backbone extracts the same set of features regardless of what the user asks. This research addresses this limitation by introducing a **feedback loop** that informs the **AbsViT** backbone of the specific prompt. By conditioning the vision stage on the "goal" (the prompt), the model performs **task-aware feature selection**, mimicking the way human vision filters noise based on current objectives. By pretraining this backbone on the **VQAv2** dataset (Goyal et al., 2017), we align the model's internal visual attention with the complex, goal-oriented tasks performed by humans in urban environments, ultimately providing a more robust and explainable tool for assessing transportation safety.

DATA

The Computational Challenge of Urban Imagery

The dataset for this analysis comprises a curated, self-collected repository of high-resolution imagery and video montages capturing pedestrian dynamics and complex urban street interactions. To ensure an unambiguous evaluation of the model's internal probing mechanisms, this dataset incorporates a deliberate blend of synthetic (AI-generated) and carefully sourced real-world images. These scenes were specifically selected for their clear, unobstructed vantage points of high-density traffic and localized pedestrian scenarios, providing an ideal, noise-reduced baseline for visually isolating and mapping the model's attention gradients.

Analyzing these highly detailed scenes poses a significant computational challenge: most Vision-Language Models (VLMs) operate on a fixed visual token budget, often resizing input images to a standard, low-fidelity grid (e.g., 336 × 336 pixels). While this downsampling is sufficient for identifying macro-level objects, it effectively destroys the fine-grained spatial details required to analyze nuanced phenomena, such as individual pedestrian grouping, behavioral reactions to environmental stimuli, and the presence of small but high-impact obstacles. In a typical urban scene, an individual pedestrian may occupy only a tiny fraction of the total pixel area; compressing these images leads to a critical loss of the exact spatial features necessary for accurate action recognition and rigorous safety assessment.

Multi-Scale Localized Tiling Strategy

To resolve this, the data is processed using a localized tiling strategy designed to maximize feature density without exceeding the model's computational limits. Rather than a single-pass inference, the input imagery is padded and strategically divided:

1. **Global Contextual Crop:** A down-sampled version of the entire frame provides the model with macro-level scene context, allowing it to understand the general layout of the street, the location of crosswalks, and the overall traffic flow.
2. **High-Resolution Sliding Windows:** A series of overlapping, high-resolution tiles are extracted from the original image. These sliding windows preserve the micro-level pedestrian features and textural details of the environment.

By integrating these crops into a multi-layered data structure, the **Molmo-AbsViT** architecture (as detailed previously in Figure 1) can effectively "zoom in" on areas of interest. This strategy is particularly powerful when combined with our **prompt-conditioned feedback mechanism**: the model uses the text prompt to prioritize which high-resolution tiles contain the most semantically relevant information, effectively focusing its computational "attention" on the most critical parts of the street scene.

Finetuning on VQA v2.0 for Perceptual Alignment

For finetuning the feedback mechanism, we employ the **VQA v2.0 dataset** (Goyal et al., 2017). As one of the most robust benchmarks for training LLMs, it consists of open-ended questions grounded in images from the **MS COCO dataset** (Lin, et al, 2014). Crucially, version 2.0 was designed to be "balanced" for every question; it includes pairs of similar images that lead to different answers for the same query.

This balancing is essential for our research because it forces the model to move beyond simple language biases, such as the tendency to answer "yes" to "Is there a...?" questions regardless of visual evidence. For a model employing **AbsViT**, this dataset provides an excellent training ground for our top-down feedback loop. It requires the vision backbone to actively select and refine features based on the specific prompt to distinguish between subtle visual differences. With over 1.1 million questions and 10 ground-truth human-annotated answers per image, VQA v2.0 ensures

that the model's internal decision-making process is aligned with human consensus, making the final system more reliable for real-world urban behavioral analysis.

ANALYSIS

The core analytical challenge of interpreting Vision-Language Models (VLMs) lies in token-to-pixel grounding. Molmo is a family of state-of-the-art open vision-language models that achieves frontier-level performance by prioritizing high-quality, information-dense training data over massive parameter counts (Deitke, et al, 2024). When the Molmo VLM processes tiled, high-resolution images, the visual data is heavily transformed and flattened into sequence tokens. To achieve true interpretability, these latent tokens must be reverse-mapped to the original, high-fidelity image space. The probing tools and associated architectural adapters utilized and developed in this project are described below. We have made these tools openly available on GitHub via the "Feedback_molmo" repository (AbdulRahman, 2026), and the names of the developed algorithms/pipelines are noted below in bold/italic fonts in the form of ***algorithm*** (*code_name*).

Spatial Reconstruction and Exact Token Grounding

Because the input imagery undergoes localized tiling, bilinear resizing, and padding to fit the model's rigid input constraints, naive reverse-mapping leads to severe spatial misalignment. To counter this, our framework employs ***a rigorous reconstruction algorithm*** (*build_token_regions_exact* and *build_token_to_pixel_map_from_outputs*). This algorithm mathematically reconstructs the spatial bounding box (y_0, y_1, x_0, x_1) for every individual token across both the macro-level "Global" crop and all high-resolution sliding windows.

This mapping accounts for the complex transformations applied during the forward pass, tracking scale shifts, padding insertions, and stride overlaps. Once the exact regions are anchored, the system extracts the raw attention weights from specific language query tokens. Using ***an inverse transformation pipeline*** (*invert_resize_and_pad*), the system averages the overlapping token weights, normalizes them, and projects a color-mapped heatmap directly back onto the unpadded, original image space, ensuring zero loss of spatial fidelity.

Statistical Validation of Attention Mechanisms

Visualizing heatmaps is qualitatively useful, but rigorous behavioral analysis requires quantitative validation to ensure the model's focus is intentional rather than an artifact of network noise. To achieve this, the analysis pipeline implements ***a robust region-scoring algorithm*** (*region_attention_score*).

When analyzing a specific target within a complex street scene, a localized mask is generated around the area of interest. The system extracts the attention density within this mask and compares it against hundreds of randomly sampled spatial distributions across the broader image. By

computing a Z-score, a percentile ranking, and an empirical p-value, we can definitively prove whether the VLM's visual attention is statistically significant and strongly correlated with the linguistic query. Furthermore, attention rollout is computed across the transformer layers to track how visual focus coalesces and evolves as it travels deeper into the network.

Causal Elucidation via Activation Patching

To move beyond correlational heatmaps and understand *where* the model actually makes its decisions, we *probe* the causal mechanics of the transformer blocks using ***activation patching*** (*patch_and_run*).

This experiment involves pairing *contrasting linguistic queries*, a "positive" prompt targeting one specific action or object, and a "negative" prompt targeting another. During inference, we register forward hooks onto the VLM's transformer blocks. As the model processes the negative prompt, we systematically inject the cached hidden states (activations) from the positive prompt, sweeping through the model layer-by-layer. By monitoring the output coordinates, we can isolate the exact layer threshold where the model's semantic interpretation flips. This technique successfully elucidates the "black box," pinpointing the specific computational depth at which high-level decision-making and spatial localization occur.

Evaluating the Bio-Inspired Feedback Architecture

The final phase of the analysis leverages this entire probing framework to evaluate the integration of the Analysis-by-Synthesis Vision Transformer (AbsViT). Using ***a custom modular adapter*** (*FrozenMolmoBackboneWithFeedback*), we dynamically swap the standard, feedforward vision backbone with our pretrained, bio-inspired module.

By running identically processed urban scenes through both architectures, we generate comparative, layer-by-layer attention overlays. This head-to-head analysis empirically measures the impact of prompt-conditioning on the vision encoder. By tracking changes in the statistical focus and the activation patching thresholds, we can quantify how injecting top-down feedback refines early-stage feature selection and ultimately yields higher-quality visual tokens for the language model to interpret.

The probing method and the associated architectures can inform the development of multimodal human-machine interface dashboards for traveler behavior analysis, making them more intuitive for human users. A sample dashboard will be produced during Phase 2 of this project, when more application-specific outcomes are produced for pedestrian behavior analysis against urban sidewalks/crosswalks.

RESULTS

The results of the spatial mapping, statistical validation, and causal attention overlays confirm the effectiveness of the top-down probing mechanism and the integration of the bio-inspired feedback architecture. By successfully reverse-mapping latent sequence tokens back to the physical image space, we can empirically evaluate the VLM's perceptual accuracy in complex urban environments.

Grounding Tokenized Data

A primary hurdle in VLM interpretability is accurately grounding tokenized data back to high-resolution input imagery, especially after scaling and localized tiling transformations. As demonstrated in **Figure 2**, the spatial reconstruction algorithm successfully superimposes the token index grid directly over the original pedestrian scene. The overlay establishes a precise spatial link between the flattened sequence data and the physical layout of the environment. **Figure 3** shows attention with feedback (left side) compared to how the model focuses with the baseline visual encoder (right side).

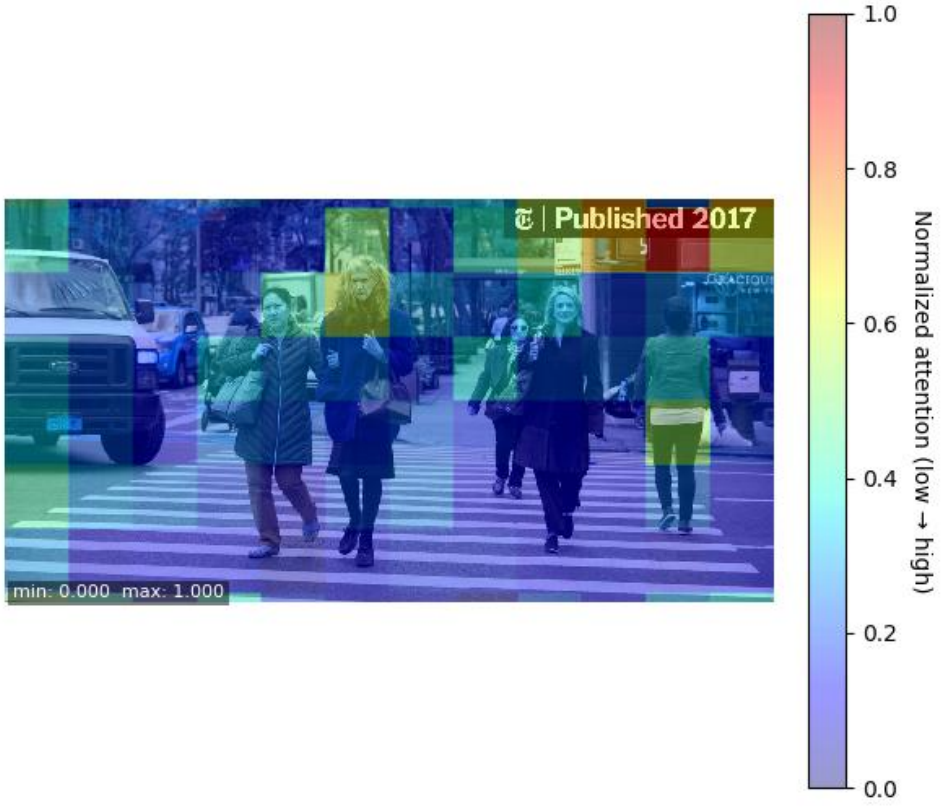

Figure 2. Token indices overlay on image, where the overlapping window patches map perfectly to critical semantic features, such as the pedestrian mid-stride, the crosswalk markings, and the traffic signal.

Figure 3. Comparison of model focuses with attention with feedback (left) vs. the baseline visual encoder (right).

In this scene, the overlapping window patches map perfectly to critical semantic features, such as the pedestrian mid-stride, the crosswalk markings, and the traffic signal (Figure 2). This visual confirmation shows that the coordinate mapping functions effectively account for scale shifts and stride overlaps, ensuring that when the model attributes attention to a specific token, we can trace it back to the exact pixel coordinates of the real-world obstacle or actor.

Figure 4 illustrates the underlying token matrix structure generated by our localized tiling strategy, displaying the 13 distinct crops processed during a single forward pass. This consists of one "Global" contextual crop and 12 high-resolution "Slide" crops.

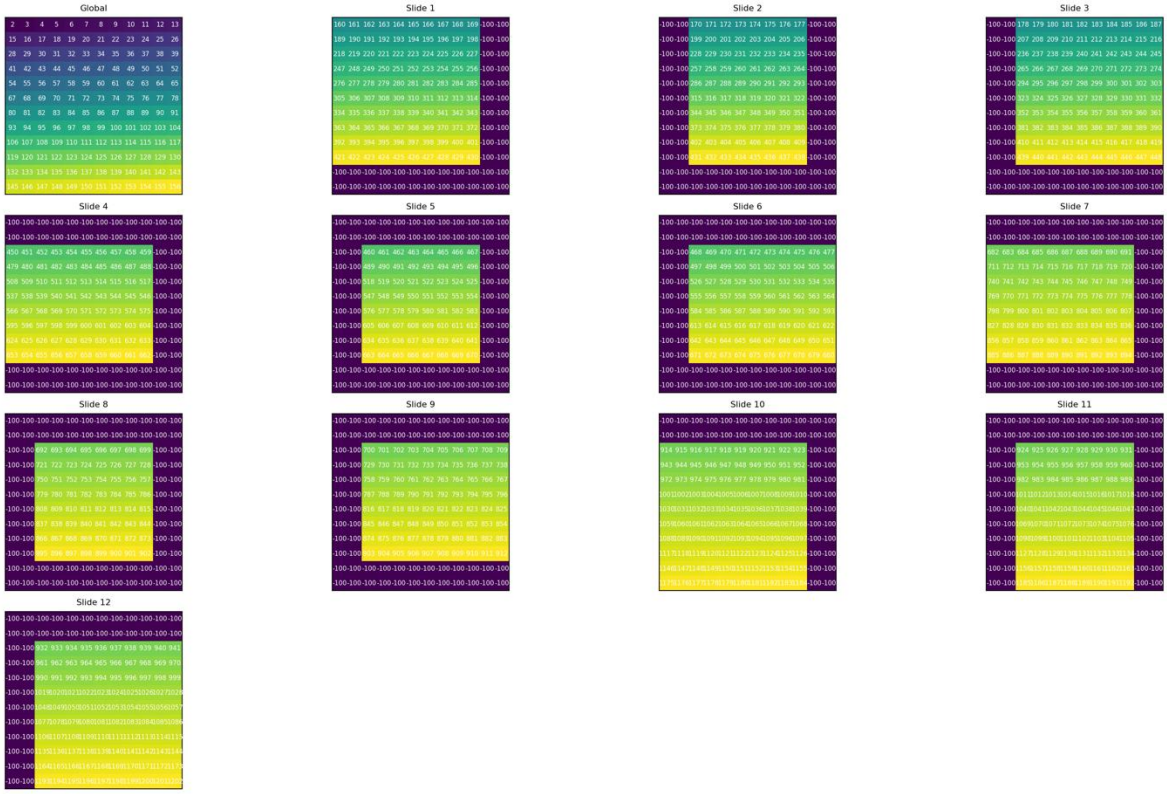

Figure 4. Multi-Scale Token Matrices and Tiling Strategy, consisting of one "Global" contextual crop and 12 high-resolution "Slide" crops.

The visualization highlights the precise boundaries of our data structure:

- **The Global Crop:** Shows a continuous sequence of token indices, representing the macro-level scene context downsampled to fit the model's base resolution.
- **The 12 Slide Crops:** Display the dense, high-resolution sliding windows. The presence of `-100` values along the borders visually confirms the model's handling of padding and overlapping margins. These margins prevent the loss of semantic continuity at the edges of the sliding windows.

By processing these overlapping matrices, the model successfully extracts both broad contextual awareness and highly localized attention gradients without exceeding its token budget.

Cognitive Alignment and Goal-Directed Focus

When querying the model with goal-directed prompts, such as "Is it safe to cross the street?", the projected heatmaps derived from these token matrices allow us to visually and statistically verify the network's internal reasoning, as demonstrated in the following examples.

Figure 5 shows the attention distribution of the word ‘safe’ on layer 11 head 16 for a simpler pedestrian crosswalk scene than the one in Figure 2. In this and the following figures (Figure 6

and Figure 7 for the word “cross”), red means high while green means medium and blue means low. **Table 1** highlights the most specialized attention heads across the transformer blocks on **Figure 5**. While hundreds of heads exhibit statistically significant spatial awareness, these specific heads demonstrate extreme, localized focus during critical semantic and syntactic sequence steps.

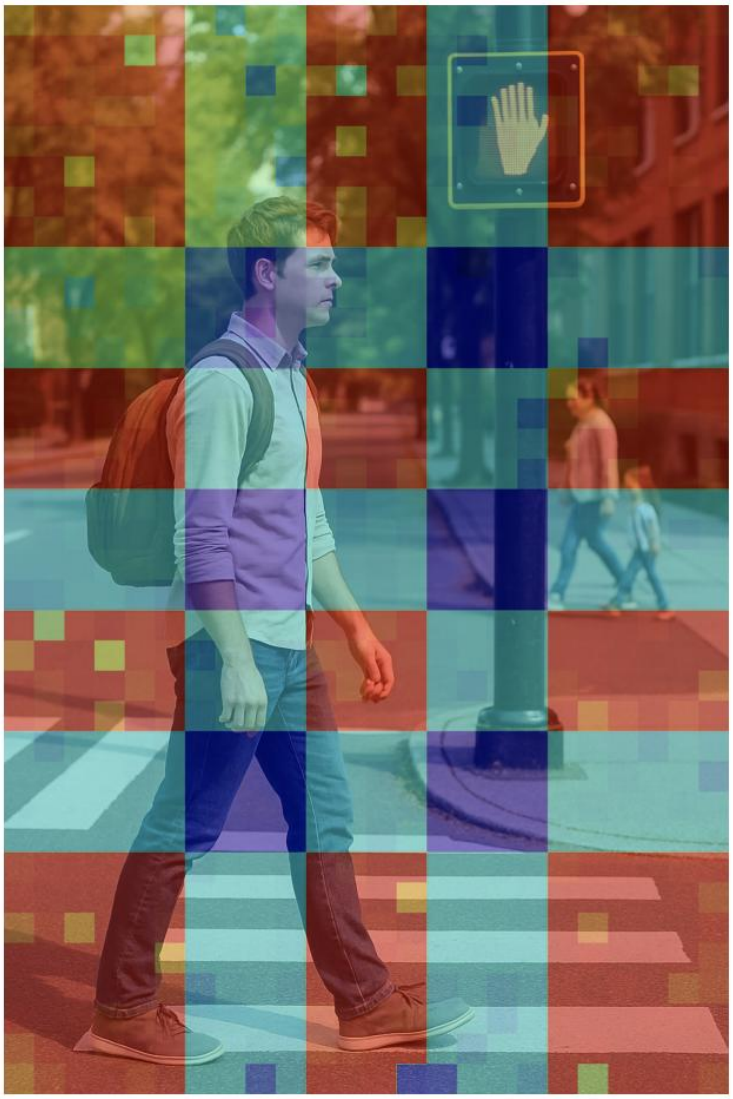

Figure 5. Attention distribution of the word ‘safe’ on layer 11 head 16 for a simple pedestrian scene but with a "Don't Walk" hand signal. The token “safe” is the first one among those highlighted in Table 1 as the most specialized attention heads.

Table 1. Extreme Localization Heads (z > 1.5, p < 0.01) for Figure 5

Layer	Token	Head	Percentile	z-score	p-value
11	safe	16	100.0	1.96	0.003
17	safe	2	100.0	1.86	0.003
19	street	26	100.0	1.78	0.003
19	?	21	100.0	1.81	0.003
19	?	26	100.0	2.37	0.003
21	to	25	100.0	1.95	0.003
23	it	19	100.0	1.84	0.003
26	to	23	100.0	1.96	0.003

Figure 6 shows the attention distribution of the 'cross' word on layer 14 head 5 for the more complicated pedestrian scene as shown in Figure 2. Table 2 highlights the most specialized attention heads across the transformer blocks on Figure 6. While hundreds of heads exhibit statistically significant spatial awareness, these specific heads demonstrate extreme, localized focus during critical semantic and syntactic sequence steps.

Figure 6. Attention distribution of the word 'cross' on layer 14 head 5 for the more complicated but clean pedestrian scene as shown in Figure 2, where the attention distribution is more uniform.

Table 2. Extreme Localization Heads (z > 1.5, p < 0.01) for Figure 6

Layer	Token	Head	Percentile	z-score	p-value
15	street	4	100.0	2.79	0.003
15	street	3	100.0	2.79	0.003
16	street	3	100.0	2.74	0.003
17	street	26	100.0	2.39	0.003
14	cross	5	100.0	2.20	0.003
13	?	13	100.0	1.97	0.003
12	?	19	100.0	1.96	0.003
13	:	13	100.0	1.87	0.003

Figure 7 shows the attention distribution of the 'cross' word on layer 10 head 22 for a more intriguing pedestrian scene with broken crosswalk stripes and heavy vehicles like a bus on the left. Again, red means high while green means medium and blue means low. Table 3 highlights the most specialized attention heads across the transformer blocks on Figure 7. While hundreds of heads exhibit statistically significant spatial awareness, these specific heads demonstrate extreme, localized focus during critical semantic and syntactic sequence steps.

Figure 7. Attention distribution of the word 'cross' on layer 10, head 22 for a more intriguing pedestrian scene with broken crosswalk stripes and a bus on the left, where attention distribution in more complicated.

Table 3. Extreme Localization Heads (z > 1.5, p < 0.01) for Figure 7

Layer	Token	Head	Percentile	z-score	p-value
4	:	23	100.0	3.60	0.003
10	cross	22	100.0	3.48	0.003
2	Is	12	100.0	3.40	0.003
10	to	22	100.0	3.37	0.003
10	it	22	100.0	3.06	0.003
6	street	0	100.0	3.02	0.003
10	the	21	100.0	2.95	0.003
10	Assistant	21	100.0	2.92	0.003

Rather than dispersing attention uniformly across the scene, the prompt-conditioned feedback mechanism actively drives the vision backbone to prioritize task-relevant features. The attention maps confirm that the network's reasoning tightly aligns with human cognitive expectations. The model demonstrably anchors its highest attention weights on safety-critical elements: the illuminated "Don't Walk" hand signal (Figure 5) and the trajectory of the moving pedestrians (Figures 5-7). This highly focused, top-down attention distribution stands in contrast to the broader, query-agnostic feature extraction typical of baseline, feedforward VLMs, indicating that the bio-inspired feedback loop actively improves task-aware visual perception.

Semantic and Syntactic Triggers in Distributed Spatial Grounding

To evaluate how the architecture grounds visual concepts to text, we probed the attention maps across all 28 transformer layers during the query "Is it safe to cross the street?". The data reveals that spatial localization is not isolated to a single "vision head," but is rather a distributed, highly sparse process triggered by specific lexical categories.

Semantic Concept Grounding and Navigable Affordances

During the processing of core semantic tokens (safe, cross, street), middle and late-stage transformer layers deploy specialized heads that heavily index the target visual region. For example, the token safe triggers extreme localization at Layer 11, Head 16 ($z = 1.96$) for the token “safe”, isolating the pedestrian and the crossing signal as shown in Figure 5.

Crucially, the attention distributed during action-oriented tokens reveals an optimization for physical constraints over social dynamics. For the token “cross” (e.g., Layer 10, Head 22, $z = 3.48$) in Figure 7, the model treats the concept as a spatial and navigational verb. The network effectively operates as a "navigable terrain head," grounding its highest attention across macro-environmental hazards, the boundaries of the asphalt, crosswalk stripes, and the trajectories of heavy vehicles like the bus on the left.

Simultaneously, the model exhibits active suppression (cool masking) over the core masses and faces of the pedestrians. When assessing urban sidewalk friction, localized identities or emotional expressions provide near-zero actionable data regarding crossing safety. Because the prompt does not require modeling inter-agent relational dynamics (e.g., gaze tracking or social cueing), the architecture deliberately down-weights high-frequency human details, prioritizing the physical geometry of the scene.

Syntactic Tokens as Information Sinks and Kinematic Volumes

A counter-intuitive finding from the attention distributions is the high degree of spatial localization triggered by syntactic and functional tokens (e.g., “to”, “?”, “:”). Rather than ignoring visual input, the model utilizes these relational tokens as information "sinks" to aggregate spatial context before sequence generation. The token “to”, representing the relational trajectory of crossing, exhibits massive attention spikes in the final layers, notably at Layer 26, Head 23 ($z = 1.96$) and Layer 21, Head 25 ($z = 1.95$) in Table 1.

Furthermore, punctuation tokens act as global aggregators while strictly enforcing physical abstraction. The terminal “?” yields the highest spatial concentration in the entire forward pass (Layer 19, Head 26, $z = 2.37$). Similarly, the “:” token (Layer 13, Head 13, $z = 1.87$) in Table 2 pools a macro-level summary of the scene while applying precise negative masks over localized "noise," such as pedestrian faces and the frontal grille of the approaching truck. This confirms that these aggregator heads calculate hazard potential by evaluating pedestrians and vehicles strictly as kinematic volumes. By discarding identifying visual textures, the network streamlines the extraction of the geometric data necessary to answer the spatial prompt.

CONCLUSIONS AND POLICY IMPLICATIONS

The integration of language-guided, top-down attention mechanisms, specifically through the novel pairing of the **Molmo** framework with an **AbsViT** vision backbone, provides a crucial pathway to nuanced, context-aware understanding in action recognition. By shifting from passive, feedforward feature extraction to goal-directed, prompt-conditioned perception, this architecture more closely mirrors human cognition. However, as these advanced technologies become increasingly prevalent in critical fields like urban mobility, algorithmic accountability is paramount. The inherent opacity of large-scale deep learning models, often termed "black-box" systems, complicates the tracing of decisions and the assignment of responsibility, which can severely undermine public trust in automated assessments.

The probing and spatial mapping tools developed in this work directly dismantle this opacity. By successfully bridging the gap between abstract latent token sequences and mathematically exact, interpretable spatial heatmaps, we render the model's internal reasoning transparent. Furthermore, by utilizing causal techniques like activation patching, we can pinpoint exactly where and how a

model makes its semantic decisions, moving beyond mere correlation to true algorithmic explainability.

From a policy perspective, tools that enforce and enhance model interpretability are non-negotiable for the responsible deployment of AI in public spaces. Transportation agencies and city planners require definitive proof that an intelligent system is not just making accurate predictions, but making them for the right reasons. By ensuring that these systems correctly perceive high-stakes urban dynamics, such as navigating sidewalks while identifying transient obstacles, and interpreting pedestrian safety cues at crosswalks, agencies can confidently utilize these technologies. Ultimately, this explainable VLM framework empowers municipalities to make data-driven decisions that inform safer street design and more inclusive accessibility policies, all while maintaining robust ethical oversight.

REFERENCES

AbdulRahman, B. (2026). *feedback_molmo* [Computer software]. GitHub. https://github.com/bilalze/feedback_molmo/

AbdulRahman, B., Conway, A., & Zhu, Z. (2026). A Survey on Action Recognition: Multimodal Approaches, Ethical Considerations, and Feedback Mechanisms. *Special Issue on Advances in Deep Learning for Open-World Computer Vision and Pattern Recognition, Electronics (ISSN 2079-9292) (under review)*.

Angel, A., Cohen, A., & Nelson, T. (2024). Evaluating the Relationship Between Walking and Street Characteristics Based on Big Data and Machine Learning Analysis. *Cities*, 132.

Angel, A., & Plaut, P. (2023). Pedestrian Movement and the Built Environment – A Big Data-Based Analysis. *Advances in Mobility*, Springer.

de Winter, J., Bazilinskyy, P., et al. (2021). How Do Pedestrians Distribute Their Visual Attention When Walking Through a Parking Garage? An Eye-Tracking Study. *Ergonomics*, 64(6).

Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., et al. (2024). Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.48550/arxiv.2409.17146

Erath, A., van Eggermond, M., Ordóñez, S., & Axhausen, K. W. (2017). Introducing the Pedestrian Accessibility Tool: Walkability Analysis for a Geographic Information System. *Transportation Research Record No. 2661*.

Gori, S., Nigro, M., & Petrelli, M. (2014). Walkability Indicators for Pedestrian-Friendly Design. *Transportation Research Record No. 2464*.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017). Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6904-6913).

Krishna, K. V., & Choudhary, P. (2025). Unravelling Situational Awareness of Multi-Tasking Pedestrians Through Average Gaze Fixation Durations: An Accelerated Failure Time Modelling Approach. *Accident Analysis & Prevention*, 182.

Li, P., Kothuri, S., Keeling, K., Yang, X., & Chowdhury, F. (2023). Pedestrian Behavior Study to Advance Pedestrian Safety in Smart Transportation Systems Using Innovative LiDAR Sensors. *NITC-RR-1393*.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., et al. (2014). Microsoft COCO: Common Objects in Context. *Proceedings of the European Conference on Computer Vision (ECCV)*, 740–755. https://doi.org/10.48550/arxiv.1405.0312

Pourhomayoun, M. (2020). Automatic Traffic Monitoring and Management for Pedestrian and Cyclist Safety Using Deep Learning and Artificial Intelligence. *Mineta Transportation Institute*, Report CA20-3501.

Ryus, P., Musunuru, A., Bonneson, J., et al. (2022). Guide to Pedestrian Analysis. *NCHRP Research Report 992*, Transportation Research Board.

Shi, B., Darrell, T., & Wang, X. (2023). Top-Down Visual Attention from Analysis by Synthesis. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2102–2112. https://doi.org/10.1109/cvpr52729.2023.00209

Wang, Y., & Srinivasan, A. R. (2024). Pedestrian Crossing Decisions Can Be Explained by Bounded Optimal Decision-Making Under Noisy Visual Perception. *Transportation Research Part C*, 150.

Yi, X., Zhao, R., & Lin, Y. (2024). The Impact of Nighttime Car Body Lighting on Pedestrians' Distraction: A VR Simulation Based on Bottom-Up Attention Mechanism. *Safety Science*, 180.

Zhong, X., Yan, X., Yang, Z., et al. (2023). Visual Exposes You: Pedestrian Trajectory Prediction Meets Visual Intention. *IEEE Transactions on Intelligent Transportation Systems*, 24(9).