

Do Navigation VLA Models Ground What They Are Told? An Empirical Audit of Landmark-Vision Alignment in VLN

Anonymous WACV Datasets Track submission

Paper ID *****

Abstract

Vision-and-language navigation (VLN) requires models to connect route instructions with visual observations when selecting actions, but standard navigation metrics do not reveal which visual evidence supports those actions. We introduce LANDMARKVLN, a 7,228-sample evaluation dataset constructed from VLN benchmarks. Each sample contains a navigation instruction, a singular landmark phrase, current and historical observations, same-instance landmark masks, and local reference actions. Using LANDMARKVLN, we audit eight released navigation VLA models with two complementary analyses. First, attribution-based vision-action (VxA) analysis measures whether visual attribution localizes the named landmark and whether such localization is more frequent for acceptable local actions. Our results here show that only two of 60 VxA evaluations support a positive vision-action association. Second, matched visual intervention compares the output change caused by editing the landmark region with that caused by applying the same edit to a comparable non-landmark region. To distinguish landmark-specific effects from general visual responsiveness, we also apply a full-frame edit that removes image structure from all observations in the evaluated temporal scope. Results here on all 23 configurations show a positive mean response to the full-frame edit, but only six support a larger response to the landmark than to the matched control. These results show that acceptable local actions and broad visual sensitivity are not, by themselves, reliable evidence that existing navigation VLA models actually rely on the instructed landmarks. Code and reproducibility materials are provided in an [anonymous repository](#).

1. Introduction

Vision-and-language navigation (VLN) asks an embodied agent to follow a natural-language route instruction using a stream of egocentric observations [2, 21, 23]. Given an instruction such as *walk past the sofa, turn left at the pi-*

ano, and stop beside the round table, the named landmarks connect successive language segments to visual evidence along the route. Following the route requires the agent to understand the instruction and its landmarks, match those landmarks to changing observations, and produce the corresponding navigation actions.

Recent navigation vision-language-action (VLA) models adapt pretrained vision-language models to generate discrete or continuous actions [32, 39]. Standard metrics such as success rate (SR) and nDTW measure navigation performances [2, 18]. However, an acceptable local action can also arise from recurring route patterns or language regularities in the benchmark, without locating or interpreting the named landmark. Such shortcut-based success does not ensure reliable action selection at decision points where the instructed landmark is needed to distinguish among otherwise plausible actions. Standard navigation metrics therefore do not characterize landmark-vision alignment, and prior VLN works have not directly tested instructed-landmark localization or landmark-specific output sensitivity.

Evaluating this alignment requires a common unit that connects language, visual evidence, and local behaviour. Navigation also introduces a temporal requirement because a landmark may be visible now or may have appeared earlier along the route. We construct LANDMARKVLN from two VLN benchmarks, R2R and RxR, for analysis. Each of its 7,228 samples links a navigation instruction and a singular landmark phrase with current and historical observations, temporally aligned masks of the same physical landmark, and local reference actions.

With this dataset, we audit each model from two angles. First, attribution-based vision-action (VxA) analysis computes a visual attribution map and measures whether it localizes the landmark named in the instruction. It separately compares the predicted local actions with the reference actions, then asks whether landmark localization is more frequent when those actions are acceptable. As a second analysis, matched visual intervention complements VxA analysis with controlled input edits. A full-frame edit mea-

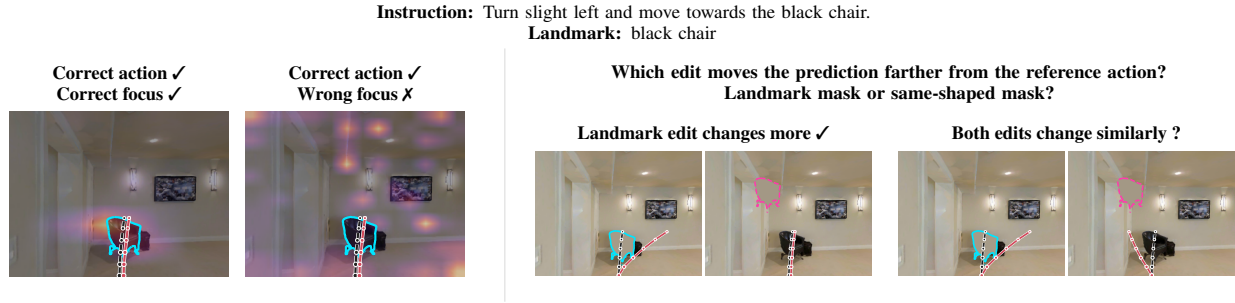


Figure 1. **Left:** The observation is overlaid with the model’s attention heatmap, predicted action in red, and reference action in black. Both predictions are correct, but the right heatmap fails to focus on the landmark. **Right:** We apply per-channel mean edits separately to the landmark mask in cyan and a same-shaped mask in magenta, then compare how far each edited prediction moves from the reference action. A larger change under the landmark edit indicates landmark-specific sensitivity. Similar changes give an inconclusive comparison.

075 sures whether the model responds to visual changes in the
076 selected temporal scope. The matched comparison then asks
077 whether editing the landmark changes the model’s output
078 more than applying the same edit to a same-shaped region
079 elsewhere in the image, measuring landmark-specific out-
080 put sensitivity. We also compare the two analyses to ask
081 whether stronger landmark attribution accompanies a larger
082 landmark-specific intervention response. These two views
083 follow existing visual-explanation research [27, 29, 37].

084 We apply the audit to eight released navigation VLA mod-
085 els: OneVLA [41], InternVLA-N1 [33], OmniVLA [12],
086 Uni-NaVid [40], FantasyVLN [42], StreamVLN [32], VA-
087 MOS [6], and TIC-VLA [17]. They span current-only and
088 history-aware visual inputs as well as discrete and contin-
089 uous navigation outputs. The VxA analysis supports a positive
090 association between landmark localization and acceptable ac-
091 tions in two of 60 panels. Full-frame edits produce a positive
092 mean response in all 23 configurations, whereas matched
093 interventions support a larger response to the landmark than
094 to its matched control in only six.

095 Our contributions are:

- 096 • We introduce LANDMARKVLN, a 7,228-sample evalua-
097 tion dataset that links navigation instructions and singular
098 landmark phrases to current and historical observations,
099 aligned same-instance masks, and local reference actions.
- 100 • We develop an audit framework that reports local action
101 agreement, landmark localization, and landmark-specific
102 output sensitivity.
- 103 • We evaluate eight released navigation VLA models on the
104 same samples and **show that binary localization rarely**
105 **tracks acceptable actions, and landmark-specific sensi-**
106 **tivity is supported in only a subset of configurations.**

107 2. Related Work

108 **VLN tasks and evaluation.** R2R [2] introduced route fol-
109 lowing in Matterport environments, RxR [23] expanded the

language and route diversity, and VLN-CE replaced graph
teleportation with continuous low-level control [14, 15, 21,
22, 30]. Standard metrics such as success rate (SR) and
nDTW measure navigation behaviour [2, 18, 20], but they
do not reveal whether an action uses the landmarks.

Navigation VLA models. Navigation VLAs adapt pre-
trained vision-language models through post-training on nav-
igation instruction-observation-action data. Recent systems
report strong SR across VLN benchmarks [6, 9, 12, 32, 38–
40]. However, none of them evaluate whether the model
connects the observations to the corresponding landmark.

Grounding and intervention tests in VQA. VQA rou-
tinely evaluates localization by comparing attention or
gradient-based importance with human-annotated visual evi-
dence [10, 27, 29, 36, 37]. Complementary intervention tests
retain or mask selected visual regions and measure whether
the answer is preserved or changes [8, 27, 37]. Localization
and intervention thus provide two established views of vi-
sual reliance in VQA. Our VxA and matched intervention
analyses adapt this template to sequential navigation actions,
temporal scopes, and landmark masks tracked across time.

131 3. Constructing LANDMARKVLN

Both analyses require a common evaluation unit that con-
nects one named landmark to local behaviour and visual
evidence. We therefore construct each LANDMARKVLN
sample by pairing a navigation instruction and singular land-
mark phrase with a current RGB observation, its recent ob-
servation history, temporally aligned masks of the same
physical landmark, and a local expert-action horizon.

As the starting pool, we use the continuous-environment
releases of R2R and RxR [2, 21, 23], restricting RxR to
English. FGR2R and Landmark-RxR provide the fine-
grained route supervision for an initial filtering and align-

Table 1. Construction scale by stage. The text-only stage selects singular physical intervention targets. The current-view stage verifies visible, unique, and geometrically usable grounding. The temporal stage resolves visibility across history and verifies the same-instance identity through time. The table reports each stage in its appropriate unit. Pairing the 118,706 text-selected identities with a few candidate frames each produces the 350,660 candidate-frame pairs used by the current-view stage.

Stage	Input	Output
Text only	39,356 English instructions; 374,270 route-local mentions	118,706 singular candidate identities
Image-text, current	350,660 candidate-frame pairs	19,742 current-grounded identities
Image-text, temporal	19,742 current-grounded identities	7,228 intervention-ready samples

ment pass [11, 13]. FGR2R aligns R2R sub-instructions with reference subpaths, while Landmark-RxR provides aligned sub-instruction-subtrajectory units and route-critical points and headings for RxR.

From this aligned base, we construct LANDMARKVLN in three stages. *Text only* selects phrases that can denote one physical intervention target. *Image-text, current* establishes that this target is present, unique, and geometrically usable in the current observation. *Image-text, temporal* resolves landmark visibility in every history frame, requires at least one historical localization, and verifies the same-instance identity through time. It then yields samples for analysis. Table 1 summarizes the scale at each stage.

3.1. Text-only selection

The text-only stage screens landmark mentions in the navigation instructions to retain phrases that can denote one physical object instance and thus serve as candidate intervention targets. It relies exclusively on the instruction text, without consulting RGB observations. The rules below reject invalid or unacceptable landmark phrases.

Object instance rather than region or surface. *The round table* is retained, whereas *the kitchen*, *the left side of the room*, and *the floor* are rejected. These phrases describe regions or surfaces rather than localized objects.

One resolvable object instance. A candidate phrase must resolve from the instruction text to exactly one physical object instance, and it can fail this requirement in three ways. First, an instruction such as *Walk past the chair or the table and enter the hallway* leaves two alternative referents, i.e., *the chair or the table*. Second, *the chair* is rejected when multiple chair mentions in the instruction remain possible referents and no linguistic cue selects one. Third, explicit

counts and plurals, such as *two lamps* and *the chairs*, denote multiple instances and are rejected.

Represented versus representing object. For example, in *Walk past the picture of a boat and turn left*, the physical scene object is the picture, not the depicted boat. The referent is therefore assigned to *the picture*, since the depicted boat has no corresponding 3-D scene instance to mask.

3.2. Current-View Landmark Grounding

The text stage yields a route-local sub-instruction and a singular landmark phrase, but not its visual evidence. The goal here is to determine whether the landmark is visible and uniquely identifiable in the corresponding current RGB observation and, if so, to obtain its landmark mask.

Progressive open-vocabulary grounding. We progressively convert each landmark phrase into a mask through candidate generation, semantic verification, and geometric verification. **Candidate generation.** SAM3 [5] is a text-prompted segmenter that produces open-vocabulary masks conditioned on the landmark phrase. In parallel, YOLOE [31] is an open-vocabulary detector and segmenter that we run with a fixed, prompt-free vocabulary, so that its object inventory is independent of the target phrase. For *the black chair*, SAM3 may propose a chair-shaped mask while YOLOE independently lists the visible objects. Their outputs form the candidate set passed to semantic verification. **Semantic verification.** We then combine the YOLOE inventory with Florence-2 [35], a vision-language model used for prompt-free object detection and region description, and Qwen2.5-VL [4], a multi-modal language model used to assess target presence and enumerate visible objects. For the same black-chair example, Florence-2 describes the proposed region as a chair and Qwen2.5-VL reports one visible black chair. This allows the same SAM3 candidate to proceed. A candidate is rejected when the target is unsupported or its correspondence is ambiguous. **Geometric verification.** Semantic survivors are finally checked using GroundingDINO [24] and OWLv2 [25], two text-conditioned open-vocabulary detectors that independently localize the landmark with bounding boxes. SAM2.1 [26], a class-agnostic promptable segmenter, converts these boxes into masks. The resulting masks must identify the same object as the SAM3 proposal. Otherwise, the candidate is rejected. The corroborated SAM3 proposal becomes the final landmark mask.

Visual uniqueness. If a frame contains several chairs and the phrase does not determine which one is route-relevant, the pair is rejected. A chair is retained only when one individual can be isolated in the current frame.

Table 2. Native action spaces of the evaluated navigation VLA models. Discrete outputs are FORWARD, LEFT 15, RIGHT 15, or STOP. For *continuous* outputs, $N \times D$ denotes N waypoints with D coordinates each.

Model	Venue / year	Action type and horizon
Uni-NaVid [40]	RSS 2025	Discrete, ≤ 4 steps
VAMOS [6]	arXiv 2025	Continuous, 5×2 waypoints
OmniVLA [12]	ICRA 2026	Continuous, 8×4 waypoints
StreamVLN [32]	ICRA 2026	Discrete, 4 steps
InternVLA-N1 [33]	ICLR 2026	Discrete, ≤ 4 steps; Continuous, 33×2 waypoints
FantasyVLN [42]	CVPR 2026	Discrete, ≤ 5 steps
TIC-VLA [17]	ICML 2026	Continuous, 30×2 waypoints
OneVLA [41]	arXiv 2026	Discrete, 5 steps

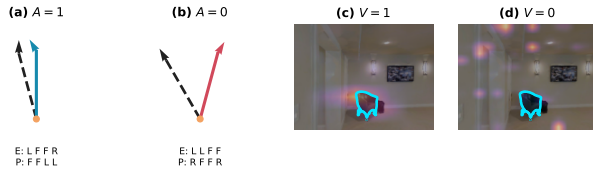


Figure 2. Operational examples of A and V . Panels (a,b) show two discrete action-diagnostic outcomes, and panels (c,d) show the two visual-localization outcomes. Dashed black trajectories are expert references, colored trajectories are model predictions, and E/P list the corresponding actions. Cyan outlines mark target landmarks. Heatmap intensity is normalized within each panel.

3.3. Temporal Landmark Verification

Current-view grounding establishes that a landmark is visible at the decision frame, but some history-based models also require reliable evidence about its earlier visibility and identity. These baselines use four history observations in addition to the current observation, so each candidate must provide a complete five-observation sequence. Of the 19,742 current-grounded identities, 17,813 satisfy this requirement.

We reuse the current-view grounding models to assess each history frame. A frame is *resolved* when their evidence either yields a reliable landmark localization or consistently indicates that the landmark is not visible. Otherwise, it is *unresolved*. We retain only sequences in which all four history frames are *resolved* and the landmark is localized in at least one of them. For localized frames, GroundingDINO and OWLv2 independently produce phrase-conditioned boxes, which SAM2.1 converts into masks; inconsistent localizations are rejected. SAM2.1 is then initialized with the verified current-frame mask and run backward through the history to reject identity switches between similar objects. A human audit of 100 samples is provided in Supplementary Sec. S2.3.

4. Alignment Analysis

The final dataset contains 7,228 landmark-grounded instruction samples. Each sample contains one instruction and its target landmark with the current observation and four history observations. With this dataset, we audit VLA models from two angles: attribution-based vision-action (VxA) analysis (Sec. 4.2) and matched visual intervention (Sec. 4.3).

4.1. Temporal scopes

The models' visual inputs determine which observations are available for analysis. OneVLA, InternVLA-N1, Uni-NaVid, FantasyVLN, StreamVLN, and TIC-VLA consume both current and historical visual observations, so we analyze their current (C), memory (M), and joint current-memory (CM) scopes. OmniVLA and VAMOS depend only on the current frame, so only C is analyzed. Table 2 describes the action spaces of the models.

4.2. Attribution-based vision-action analysis

Action variable. The binary action diagnostic A evaluates whether a model's predicted actions agree with or are acceptably close to the reference actions. If one calls STOP and the other does not, we set $A = 0$. We first convert every model output into discrete navigation actions. We then compare the prediction with the reference based on the resulting position, heading, action order, and collisions. For VAMOS and TIC-VLA, we quantize the first predicted waypoint and compare it with the referenced action. We set $A = 1$ when the prediction satisfies the acceptance criterion and $A = 0$ otherwise. Figure 2(a,b) shows two examples. Supplementary Sec. S3 provides detailed calculations and thresholds.

Visual variable. The binary visual diagnostic V evaluates whether the model's visual attribution is localized on the target landmark. For the current scope, following region-mass localization metrics [3], we sum attribution mass inside the current target mask. For the memory and joint scopes, we sum attribution mass inside the corresponding target masks across the active frames. We set $V = 1$ when the target mass meets the localization criterion and $V = 0$ otherwise. Figure 2(c,d) shows the two visual outcomes. Supplementary Sec. S3.3 provides calculations and thresholds.

VxA outcome. Together, V and A define four outcomes. $(V, A) = (1, 1)$ combines an acceptable action with localized landmark attribution. $(0, 1)$ is an acceptable action without detected landmark localization. $(1, 0)$ combines an action-diagnostic failure with localized landmark attribution. $(0, 0)$ means that both diagnostics are negative.

Attribution estimators. We evaluate attention, gradient, and attention $\times$ gradient for visual attribution. For each

model, s is a differentiable scalar derived from its predicted navigation output. Supplementary Table S5 gives the exact definition of s . Let $a_{\ell,h,q,v}$ denote the attention from action-related query q to visual token v at layer ℓ and head h . *Attention* aggregates how strongly the action-related queries attend to that token [19, 34]. *Gradient* aggregates the absolute sensitivity of s to the same attention probabilities [28]. A general form is

$$H_{\text{att}}(v) \propto \sum_{\ell} \text{mean}_{h,q} a_{\ell,h,q,v},$$

$$H_{\text{grad}}(v) \propto \sum_{\ell} \text{mean}_{h,q} \left| \frac{\partial s}{\partial a_{\ell,h,q,v}} \right|. \quad (1)$$

Attention $\times$ *gradient* is the product of the attention and gradient signals [7]. Supplementary Sec. S3 and Table S6 give the exact construction for every model.

Temporal operations. For C, the estimator is projected only to the current visual-token grid and compared with the current landmark mask. For M, it is projected to the four history grids and compared with the corresponding masks, with the current grid excluded. For CM, all five grids and masks are included. Attribution is normalized over the active visual scope. Supplementary Sec. S3 specifies the model-specific order of aggregation and normalization.

Reported measures. We summarize marginal landmark localization by the localization rate $\pi_V = \Pr(V = 1)$ and vision-action association by the risk difference $\text{RD} = \Pr(V = 1 \mid A = 1) - \Pr(V = 1 \mid A = 0)$. The former records how often attribution localizes to the landmark, while the latter compares this rate between $A = 1$ and $A = 0$ predictions. Each entry in Table 3 pairs the sample estimates $\hat{\pi}_V (\%) \mid \widehat{\text{RD}}$. For example, $32.4 \mid -.023$ means that $V = 1$ occurs in 32.4% of samples overall and is 2.3 percentage points less frequent among $A = 1$ than $A = 0$ predictions. Trajectory-level confidence intervals (CIs) quantify the statistical uncertainty in $\widehat{\text{RD}}$. The exact estimands and statistical protocol appear in Supplementary Secs. S3.3 and S4.

4.3. Matched visual intervention analysis

Matched visual intervention measures landmark-specific output sensitivity. It tests whether the model’s native navigation output responds more to editing the instructed landmark than to applying the same edit at another image location.

Intervention regions. The analysis uses three visual supports (Figure 3). The *target mask* T covers the instruction-named landmark. The *matched-control mask* R is a deterministic translation of the same mask to a location whose bounding box does not overlap T . Target and control therefore contain the same number and arrangement of pixels at



Figure 3. Intervention regions in the current observation. (a) Current RGB, (b) cyan target mask T on the named landmark, (c) dashed magenta same-shaped, non-overlapping control R , and (d) full-frame support F . In (b-d), pixels inside the indicated support are replaced by the per-channel mean. The target and control define the matched effect $E = D^T - D^R$, while the full-frame condition measures broad visual sensitivity.

different locations. Following semantic visual intervention in VQA [1], the local edit removes the selected region by replacing its pixels with the per-channel mean of the unmasked pixels in that frame [16]. Applied to T , it removes the landmark. Applied to R , it changes an equally shaped region elsewhere in the image. Target-edited and control-edited inputs use the same edit rule, while the clean input provides the shared baseline. The *full-frame support* F covers every pixel in the active observation or observations. Its edit replaces each active frame with that frame’s per-channel mean and measures the model’s broad sensitivity to visual input in the selected temporal scope. The matched target-control effect uses T and R , while the full-frame response is reported as a complementary visual sensitivity measure.

Temporal intervention scopes. A C intervention edits the selected support in the current observation and keeps the history unaltered. An M intervention edits the selected support in all four history observations and keeps the current RGB clean. A CM intervention applies the edit to current and history together.

Matched landmark-specific effect. For each intervention, we run the model on the edited input and compare its output with the clean output using a model-specific response measure d_p . Depending on the model’s output format, d_p measures a waypoint or action-vector distance, an action-sequence distance, or the decrease in log probability assigned to the clean prediction. Supplementary Table S5 gives the exact measure for every model.

Let $f_p(X_i)$ denote model p ’s clean navigation output for sample i . For temporal scope $s \in \{C, M, CM\}$ and support $Q \in \{T, R\}$, define

$$D_{i,s}^{Q,(p)} = d_p(f_p(X_i), f_p(\mathcal{I}(X_i, Q_{i,s}))), \quad (2)$$

$$E_{i,s}^{(p)} = D_{i,s}^{T,(p)} - D_{i,s}^{R,(p)}. \quad (3)$$

The effect $E_{i,s}^{(p)}$ compares the response to editing the landmark with the response to editing its matched control. We

report the mean $\bar{E}_s^{(p)}$ over eligible samples. A positive mean indicates greater sensitivity to the landmark than to the control. Table 4 reports these means, and Supplementary Table S9 provides their intervals and statistical tests.

Attribution-intervention agreement. For this comparison, we use attention $\times$ gradient and define the continuous attribution score Δm as the target attribution mass minus the mean mass over the translated spatial controls used by V . Within each model, we compute the Spearman correlation between Δm and the matched target-control effect E over jointly valid samples. This asks whether samples with a stronger attribution preference for the landmark also show a larger landmark-specific output response.

5. Evaluation Design

Analysis grid. We evaluate all eight models on the same 7,228 samples. Their available visual inputs define 20 model-scope combinations. The six temporal models contribute C, M, and CM, while the two current-only models contribute C. For VxA, every combination uses the landmark-bearing local instruction segment with attention, gradient, and attention $\times$ gradient, giving 60 panels. We report localization rates and risk differences for all 60 panels.

The matched intervention analysis keeps the full instruction fixed and uses the same 20 model-scope combinations. InternVLA-N1 exposes continuous and discrete response branches, which are evaluated separately. This produces 23 model-branch-scope estimates.

Metrics. We report results for VxA risk differences, mean target-control effects, and attribution-intervention Spearman correlations. Details are provided in Supplementary Sec. S4.

6. Results

The results answer three linked questions. **Q1:** Does landmark localization occur more often when a model produces an acceptable action? **Q2:** Does a model respond more to editing the instructed landmark than to editing a matched control? **Q3:** Do continuous attribution and intervention response vary together across samples?

6.1. Attribution-Based Vision-Action Analysis

Landmark localization rarely identifies acceptable actions. Landmark localization occurs frequently in several configurations, but evidence that it occurs more often for acceptable than unacceptable actions is rare. Across the 60 model-method panels, localization rates range from 11.4% to 65.4%. These rates describe how often attribution localizes the instructed landmark across all samples. The risk difference then compares localization rates between acceptable and unacceptable actions. The 60 differences range from

Table 3. Landmark-localization rate and vision-action association for attention, gradient, and attention $\times$ gradient using the landmark segment. Each cell reports $\hat{\pi}_V$ (%) and $\widehat{RD}$, separated by |. Dashes indicate scopes that are not applicable. Statistical intervals and adjusted tests for all panels appear in Supplementary Table S7.

Model	Estimator	C		M		CM	
OneVLA	Attention	42.5	[-.008 45.4]	+0.024	47.6	+0.009	
	Gradient	28.8	[-.038 21.5]	+0.008	25.3	-.012	
	Attention $\times$ gradient	32.4	[-.023 25.3]	-.002	29.1	-.023	
InternVLA-N1	Attention	56.0	[+.004 52.3]	-.012	57.3	-.006	
	Gradient	49.9	[-.018 64.2]	-.008	65.4	-.018	
	Attention $\times$ gradient	56.2	[+.004 52.2]	-.011	57.5	-.008	
OmniVLA	Attention	51.7	[+.026 -]	-	-	-	
	Gradient	13.8	[-.020 -]	-	-	-	
	Attention $\times$ gradient	51.4	[+.021 -]	-	-	-	
Uni-NaVid	Attention	54.8	[-.034 47.1]	-.039	54.7	-.035	
	Gradient	39.6	[+.016 34.4]	+0.014	40.3	+0.009	
	Attention $\times$ gradient	55.1	[-.029 47.7]	-.030	54.8	-.033	
FantasyVLN	Attention	56.5	[-.005 54.6]	+0.003	58.8	-.0002	
	Gradient	35.8	[+.024 49.2]	+0.009	50.6	+0.017	
	Attention $\times$ gradient	53.1	[-.023 50.9]	-.032	55.3	-.019	
StreamVLN	Attention	40.2	[+.015 45.5]	+0.010	45.4	+0.006	
	Gradient	38.9	[+.046 55.4]	+0.027	56.0	+0.041	
	Attention $\times$ gradient	40.0	[+.017 44.3]	+0.007	44.7	+0.001	
VAMOS	Attention	28.6	[+.006 -]	-	-	-	
	Gradient	21.1	[+.006 -]	-	-	-	
	Attention $\times$ gradient	30.8	[+.006 -]	-	-	-	
TIC-VLA	Attention	14.7	[+.004 52.4]	-.049	19.7	-.009	
	Gradient	12.4	[+.003 41.9]	-.028	18.5	-.010	
	Attention $\times$ gradient	11.4	[+.008 38.2]	-.022	13.4	-.002	

-.049 to +.046, with 28 positive and 32 negative values. After model-wise Holm correction, two of the 28 positive differences are supported: StreamVLN gradient in C and CM. The remaining 26 do not pass the adjusted Fisher test. This leaves 2 of 60 panels with reliable evidence that landmark localization is more common for acceptable actions. **Frequent landmark localization therefore rarely translates into a supported relationship with action acceptability.**

Different attribution methods give different VxA patterns. The three ways of computing attribution do not always give the same relationship between landmark localization and action acceptability for a fixed model and visual scope. For Uni-NaVid, gradient gives positive values in C, M, and CM, while attention and attention $\times$ gradient give negative values. StreamVLN gives positive values under all three methods and in all three scopes, with the largest values under gradient: +.046 in C, +.027 in M, and +.041 in CM. The attribution calculation therefore changes both the direction and size of the measured relationship.

6.2. Matched Visual Intervention Analysis

The matched intervention analysis asks whether a model responds more to editing the instruction-named landmark

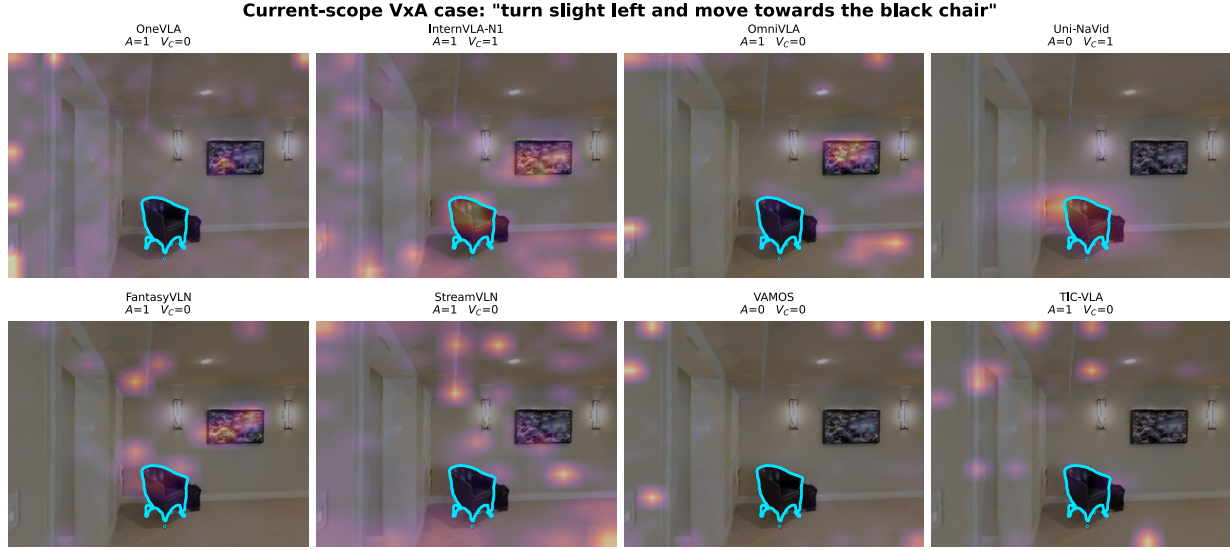


Figure 4. VxA case study for *turn slight left and move towards the black chair*. Panels overlay current-scope attention $\times$ gradient attribution. Cyan outlines the target. Labels give the action diagnostic A and localization decision V_C . Together, the panels show all four combinations of the two diagnostics. Heatmaps are normalized independently within each model.

Table 4. Mean matched target-control response $\bar{E}_s$ for the full dataset. Positive values indicate a larger response to the target edit. Negative values indicate a larger response to the matched control. Complete intervals appear in Supplementary Tables S5 and S9.

Model	C	M	CM
OneVLA	-.0197	+.0176	-.0946
InternVLA-N1 (cont.)	+.00181	+.00595	+.00938
InternVLA-N1 (disc.)	-.0131	-.0234	-.0369
OmniVLA	+.00191	—	—
Uni-NaVid	+.0100	-.0164	-.0050
FantasyVLN	-.0373	+.1068	-.5450
StreamVLN	-.0654	-.0041	-.0784
VAMOS	+.00279	—	—
TIC-VLA	-.00454	+.01075	+.00971

6.3. Case Study

The case study shows how the same diagnostics separate within individual instructions and observations.

All four VxA outcomes in one observation. Figure 4 holds the black-chair instruction and current observation fixed while comparing the eight models. For each panel, the cyan outline and attribution map determine V_C , while the label reports A . InternVLA-N1 produces an acceptable action and localizes the chair. Uni-NaVid localizes the chair but fails the action diagnostic. VAMOS fails both diagnostics. The other five models produce acceptable actions without passing the localization diagnostic. This observation therefore realizes all four action-localization outcomes and makes both directions of the VxA mismatch concrete.

Joint diagnostic profiles. Figure 5 adds the sample-level sign of the matched current-scope effect to the two VxA diagnostics. The three rows under each image show how A , V_C , and $E_C > 0$ divide the eight models. The round-table and chair cases both show high localization with split action outcomes. All eight models localize the round table and seven localize the chair, while four produce acceptable actions in each case. The piano separates these two diagnostics from intervention response: seven models pass A and six pass V_C , while only three have $E_C > 0$. The lamp gives the complementary profile: only one model passes V_C , while five pass A and five have $E_C > 0$. These selected cases provide qualitative examples of diagnostic separation. Tables 3 and 4 provide the aggregate frequency and statistical

than to editing its matched control.

Matched edits isolate landmark-specific response. The matched comparison asks whether editing the landmark produces a larger influence on response than applying the same edit to an equally shaped control region. In Table 4, positive values favor the target and negative values favor the control. Of the 23 mean target-control effects, six support a larger response to the landmark after Holm correction within each native output family. The six supported target-favored effects are OneVLA M (+.0176), InternVLA-N1 continuous CM (+.00938), OmniVLA C (+.00191), VAMOS C (+.00279), and TIC-VLA M (+.01075) and CM (+.00971). The **central intervention result is that supported landmark-specific response appears in only 6 of 23 configurations.**

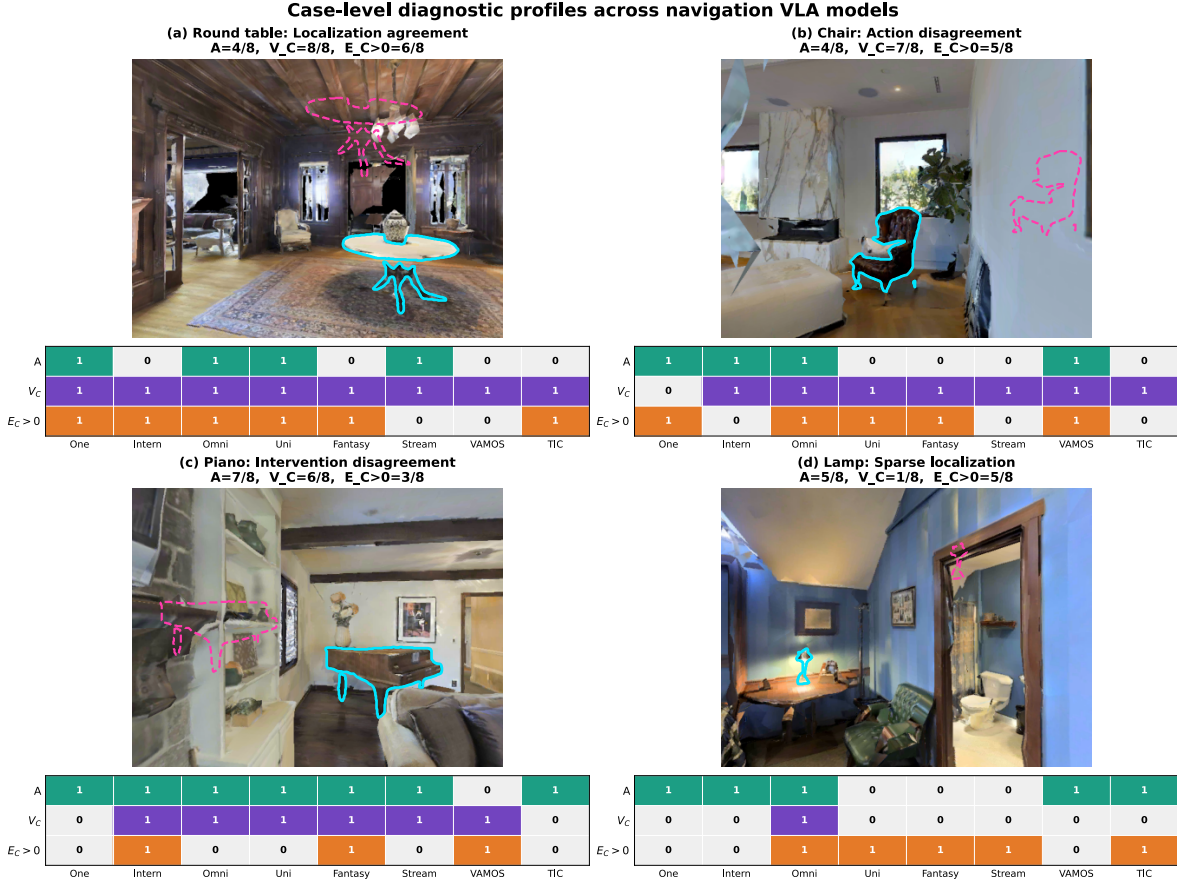


Figure 5. Case-level diagnostic profiles for four selected LANDMARKVLN samples: (a) round table, (b) chair, (c) piano, and (d) lamp. Cyan is the instruction target and dashed magenta is the same-shaped control. Each panel reports the local action diagnostic A , the binary current-scope attention $\times$ gradient result V_C , and whether $E_C > 0$ for all eight models. Here $E_C > 0$ means that the target edit produces a larger native response than the matched control for that sample. The InternVLA-N1 column uses its discrete response branch for E_C .

Table 5. Association between continuous landmark attribution and matched target-control response. Metrics are Spearman ρ . Complete results appear in Supplementary Table S10.

Model/branch	C ρ	M ρ	CM ρ
OneVLA	.100	.105	.106
InternVLA-N1 (cont.)	.087	.119	.144
InternVLA-N1 (disc.)	.031	.047	.073
OmniVLA	.182	—	—
Uni-NaVid	.044	.016	.057

evidence. The binary disagreements motivate the all-sample continuous comparison in the next subsection.

6.4. Attribution-Intervention Association

The two analyses have a weak positive association. Across the 23 configurations, Spearman correlations between continuous landmark attribution and the matched target-control response are all positive but small, ranging from

.016 to .182, with a median of .047 (Table 5). OmniVLA C has the largest correlation at $\rho = .182$.

History does not consistently strengthen the association.

The correlation increases from C to CM for InternVLA-N1 and Uni-NaVid, remains nearly constant for OneVLA and FantasyVLN, and decreases for StreamVLN and TIC-VLA. The weak positive association therefore appears across models, while its variation across visual scopes is model-dependent. Supplementary Table S10 reports sample counts and 95% confidence intervals for every configuration.

7. Conclusion

We introduced LANDMARKVLN to audit whether navigation VLA models rely on instructed landmarks. Across eight models, acceptable local actions and broad visual responses were not, by themselves, reliable evidence of landmark-specific reliance.

References

- [1] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal VQA: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 5
- [2] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sunderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2
- [3] Leila Arras, Ahmed Osman, and Wojciech Samek. CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81:14–40, 2022. 4
- [4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3
- [5] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, pages 138846–138923, 2026. 3
- [6] Mateo Guaman Castro, Sidharth Rajagopal, Daniel Gorbato, Matt Schmitt, Rohan Bajjal, Octi Zhang, Rosario Scalise, Sidharth Talia, Emma Romig, Celso de Melo, et al. Vamos: A hierarchical vision-language-action model for capability-modulated and steerable navigation. *arXiv preprint arXiv:2510.20818*, 2025. 2, 4
- [7] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 5
- [8] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10800–10809, 2020. 2
- [9] 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. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, 2025. 2
- [10] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 932–937, 2016. 2
- [11] Keji He, Yan Huang, Qi Wu, Jianhua Yang, Dong An, Shuanglin Sima, and Liang Wang. Landmark-RxR: Solving vision-and-language navigation with fine-grained alignment supervision. In *Advances in Neural Information Processing Systems*, 2021. 3
- [12] Noriaki Hirose, Catherine Glossop, Dhruv Shah, and Sergey Levine. Omnivla: An omni-modal vision-language-action model for robot navigation. *arXiv preprint arXiv:2509.19480*, 2025. 2, 4
- [13] Yicong Hong, Cristian Rodriguez, Qi Wu, and Stephen Gould. Sub-instruction aware vision-and-language navigation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 3360–3376, 2020. 3
- [14] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. VLN BERT: A recurrent vision-and-language BERT for navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1643–1653, 2021. 2
- [15] Yicong Hong, Zun Wang, Qi Wu, and Stephen Gould. Bridging the gap between learning in discrete and continuous environments for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2
- [16] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 9734–9745, 2019. 5
- [17] Zhiyu Huang, Yun Zhang, Johnson Liu, Rui Song, Chen Tang, and Jiaqi Ma. TIC-VLA: A think-in-control vision-language-action model for robot navigation in dynamic environments. In *Forty-third International Conference on Machine Learning*, 2026. 2, 4
- [18] Gabriel Ilharco, Vihan Jain, Alexander Ku, Eugene Ie, and Jason Baldridge. General evaluation for instruction conditioned navigation using dynamic time warping. In *NeurIPS Workshop on Visually Grounded Interaction and Language*, 2019. 1, 2
- [19] Sarthak Jain and Byron C. Wallace. Attention is not explanation. *Proceedings of NAACL*, 2019. 5
- [20] Vihan Jain, Gabriel Magalhaes, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. Stay on the path: Instruction fidelity in vision-and-language navigation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. 2
- [21] 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*, 2020. 1, 2
- [22] Jacob Krantz, Aaron Gokaslan, Dhruv Batra, Stefan Lee, and Oleksandr Maksymets. Waypoint models for instruction-

- guided navigation in continuous environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15142–15151, 2021. 2
- [23] 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*, 2020. 1, 2
- [24] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Computer Vision – ECCV 2024*, pages 38–55. Springer Nature Switzerland, 2024. 3
- [25] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Advances in Neural Information Processing Systems*, pages 72983–73007, 2023. 3
- [26] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, pages 28085–28128, 2025. 3
- [27] Daniel Reich, Felix Putze, and Tanja Schultz. Measuring faithful and plausible visual grounding in VQA. In *Findings of EMNLP*, pages 3129–3144, 2023. 2
- [28] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 5
- [29] Ramprasaath R. Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a HINT: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2591–2600, 2019. 2
- [30] Josh Qixuan Sun, Huaiyuan Weng, Xiaoying Xing, Chul Min Yeum, and Mark Crowley. View invariant learning for vision-language navigation in continuous environments. *IEEE Robotics and Automation Letters*, 11(5):5861–5868, 2026. 2
- [31] Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. YOLOE: Real-time seeing anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24591–24602, 2025. 3
- [32] 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. *arXiv preprint arXiv:2507.05240*, 2025. 1, 2, 4
- [33] Meng Wei, Chenyang Wan, Peng Peng, Xiqian Yu, Yuqiang Yang, Delin Feng, Wenzhe Cai, Chenming Zhu, Tai Wang, Jiangmiao Pang, et al. Ground slow, move fast: A dual-system foundation model for generalizable vision-language navigation. In *International Conference on Learning Representations*, pages 12380–12396, 2026. 2, 4
- [34] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of EMNLP*, 2019. 5
- [35] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024. 3
- [36] Xiaoying Xing, Chia-Wen Kuo, Li Fuxin, Yulei Niu, Fan Chen, Ming Li, Ying Wu, Longyin Wen, and Sijie Zhu. Where do large vision-language models look at when answering questions? *arXiv preprint arXiv:2503.13891*, 2025. 2
- [37] Zhuofan Ying, Peter Hase, and Mohit Bansal. VisFIS: Visual feature importance supervision with right-for-the-right-reason objectives. In *Advances in Neural Information Processing Systems*, 2022. 2
- [38] Zhuoyuan Yu, Yuxing Long, Zihan Yang, Chengyan Zeng, Hongwei Fan, Jiyao Zhang, and Hao Dong. Correctnav: Self-correction flywheel empowers vision-language-action navigation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18737–18745, 2026. 2
- [39] Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. NaVid: Video-based VLM plans the next step for vision-and-language navigation. In *Robotics: Science and Systems*, 2024. 1
- [40] Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-NaVid: A Video-based Vision-Language-Action Model for Unifying Embodied Navigation Tasks. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, 2025. 2, 4
- [41] Lingfeng Zhang, Xiaoshuai Hao, Yingbo Tang, Lei Zhou, Shuyi Zhang, Jinkun Liu, Hongsheng Li, Chenhao Zhang, Qiang Zhang, Hangjun Ye, et al. Onevla: A unified framework for embodied tasks. *arXiv preprint arXiv:2606.01241*, 2026. 2, 4
- [42] Jing Zuo, Lingzhou Mu, Fan Jiang, Chengcheng Ma, Mu Xu, and Yonggang Qi. Fantasyvln: Unified multimodal chain-of-thought reasoning for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29674–29683, 2026. 2, 4