

# Do Navigation VLA Models Ground What They Are Told?

## An Empirical Audit of Landmark-Vision Alignment in VLN

### Supplementary Material

#### S1. Limitations

Attribution visualizations measure where computation concentrates. Matched interventions measure the sensitivity of model outputs to controlled pixel edits. Pixel replacement can introduce visual inputs outside the model’s training distribution, which limits the connection between these effects and closed-loop navigation behaviour.

#### S2. Dataset Documentation

##### S2.1. Composition and Source Coverage

**Source coverage.** This subsection supplements the construction summary in main Sec. 3 and Table 1. The 7,228 samples draw on 5,922 English instructions and 4,025 source navigation trajectories across 68 Matterport environments. Table S1 reports their distribution across the R2R and RxR source splits. Environment counts overlap across sources and splits, so the total is computed over their union.

Table S1. Composition and source coverage of LANDMARKVLN. Environment counts give the number represented within each source split.

| Source | Split      | Samples | Instructions | Trajectories | Environments |
|--------|------------|---------|--------------|--------------|--------------|
| R2R    | Train      | 578     | 540          | 432          | 51           |
| R2R    | Val-seen   | 37      | 36           | 27           | 22           |
| R2R    | Val-unseen | 111     | 103          | 77           | 9            |
| RxR    | Train      | 5,044   | 4,068        | 2,700        | 57           |
| RxR    | Val-seen   | 578     | 459          | 304          | 54           |
| RxR    | Val-unseen | 880     | 716          | 485          | 10           |
| Total  | All        | 7,228   | 5,922        | 4,025        | 68           |

##### S2.2. Released Sample Schema

**Sample fields.** This subsection expands the evaluation unit defined in main Sec. 3. Table S2 groups the fields released for each sample. Together they connect the instructed landmark to its visual evidence and the local reference behaviour used by the two analyses.

Table S2. Fields released for each LANDMARKVLN sample.

| Field group         | Contents                                                                                                                  |
|---------------------|---------------------------------------------------------------------------------------------------------------------------|
| Source metadata     | Source benchmark and split, sample identifier, and source-trajectory identifier.                                          |
| Language            | Full navigation instruction, singular landmark phrase, and the landmark-bearing local segment.                            |
| Visual evidence     | Current RGB observation and four preceding RGB observations.                                                              |
| Landmark annotation | Same-instance landmark masks aligned across the five observations. A resolved nonvisible history frame has an empty mask. |
| Reference behaviour | A local reference horizon of up to five navigation actions aligned with the current observation.                          |

##### S2.3. Human Quality Audit

**Audit design and results.** This subsection supplements the dataset construction in main Sec. 3. We conducted a stratified human audit of 100 samples drawn from distinct source trajectories. One annotator reviewed the landmark phrase, current observation and mask, four history observations and their masks, and same-instance consistency across time. Table S3 reports the results.

Table S3. Human quality audit results. Four history frames are evaluated per sample. History-mask usability is evaluated for the 399 frames automatically labelled `present`. A complete sample satisfies every applicable audit criterion.

| Criterion                          | Correct | Rate (%) |
|------------------------------------|---------|----------|
| Phrase denotes a physical object   | 100/100 | 100.0    |
| Phrase denotes one instance        | 94/100  | 94.0     |
| Phrase has an unambiguous referent | 94/100  | 94.0     |
| Current-mask category              | 97/100  | 97.0     |
| Current-mask instance              | 95/100  | 95.0     |
| Current-mask usability             | 98/100  | 98.0     |
| History visibility                 | 382/400 | 95.5     |
| History-mask usability             | 381/399 | 95.5     |
| Same-instance consistency          | 99/100  | 99.0     |
| Complete sample                    | 94/100  | 94.0     |

## S3. Operational Definitions and Model-Specific Processing

### S3.1. Temporal Verification

**Temporal state resolution.** This paragraph gives the exact rules behind Temporal Landmark Verification in main Sec. 3.3. The state `resolved` groups two implementation states, `present` and `absent`. Uncertainty is never treated as absence. A history frame is `present` when both SAM3 presence and joint scores are at least .75, its mask passes the frozen geometry checks, and its mask overlaps at least one target-blind source at .25 IoU or higher. The sources are YOLOE, prompt-free Florence-2 outputs, and a single-instance Qwen2.5-VL inventory. A frame is `absent` when both SAM3 scores are at most .20 and YOLOE and the prompt-free Florence-2 branches return no target boxes. All other outcomes are `unresolved` and are recorded as `uncertain` in the frozen artifacts. All four history frames must be `resolved`, at least one must be `present`, and every nonempty history mask must pass DINO/OWL-to-SAM2.1 geometry and current-mask-anchored reverse tracking.

**Temporal geometry and identity.** This paragraph supplies the overlap thresholds referenced in main Sec. 3.3. A selected SAM3 mask must cover .1%-15% of the frame, touch at most one image border, and place at most 1% of its pixels on the border. For every `present` frame, GroundingDINO and OWLv2 detections must score at least .20 and .05. Their SAM2.1 masks must each have predicted IoU at least .75, overlap one another by at least .50 IoU, and overlap the selected SAM3 mask by at least .25 IoU. Reverse tracking uses the current SAM3 mask as its anchor and requires SAM2.1 predicted IoU at least .75 and at least .25 IoU with each selected history mask. Target-conditioned Qwen2.5-VL responses and blind montage labels are retained as quality metadata.

### S3.2. Model Inputs

**Released input processing.** This paragraph expands the temporal scopes in main Sec. 4.1 and the analysis grid in main Sec. 5. VxA uses the landmark-bearing local instruction segment, while matched interventions keep the full instruction fixed. Each model retains its released instruction template, image processing, temporal interface, and action decoder. The shared language and observations are formatted for each model while preserving its released navigation output format.

### S3.3. Action and Visual Diagnostics

**Exact action diagnostic.** This paragraph expands the action variable in main Sec. 4.2 and the native outputs in main Table 2. For the six models whose released outputs are converted to an ordered action chunk  $a_{1:h}$ ,  $h \in \{1, \dots, 5\}$ , let  $S_i(a_{1:j})$  be the simulated state after executing prefix  $j$  and let  $z_{i,h}$  be the oriented endpoint of the expert stage containing the horizon endpoint. Prefix quality is

$$Q_i(a_{1:j}) = \frac{d_{\text{geo}}(S_i(a_{1:j}), z_{i,h})}{0.25 \text{ m}} + \frac{\Delta\theta(S_i(a_{1:j}), z_{i,h})}{\alpha_p} + 4c_i(a_{1:j}), \quad (\text{S1})$$

$$R_i(a_{1:h}) = \sum_{j=1}^h .9^{j-1} \left[ Q_i(a_{1:j}) - \min_{b_{1:j} \in \{F,L,R\}^j} Q_i(b_{1:j}) \right]_+. \quad (\text{S2})$$

Here  $c_i$  is the number of collisions and  $\alpha_p$  is the model’s native  $15^\circ$  or  $30^\circ$  turn. Target heading replaces the next-path heading within 0.5 m of the endpoint. The stop decision is evaluated before this trajectory comparison: predicted and expert stop status must agree. The evaluation samples are nonterminal, so a predicted STOP fails. Otherwise,  $A = 1$  iff  $R_i \leq 4.0742669$  and no prefix collides. VAMOS and TIC-VLA instead use their released first-waypoint resolution: the waypoint bearing is quantized at  $\pm 7.5^\circ$  and must equal the first expert primitive. Table S4 specifies native decoding, chunk truncation, malformed-output handling, and model-specific application.

**Exact visual diagnostic and VxA estimand.** This paragraph expands the visual variable and reported measures in main Sec. 4.2 and the examples in main Fig. 1. Let  $G_{i,s}^{(p)}(x) \geq 0$  be the action-conditioned attribution map for model  $p$  and temporal scope  $s$ , normalized once over the complete active visual domain  $\Omega_s$ :

$$\sum_{x \in \Omega_s} G_{i,s}^{(p)}(x) = 1. \quad (\text{S3})$$

For target mask or mask sequence  $T_{i,s}$ , its attribution overlap is the mass

$$m_{i,s}^{(p)} = \sum_{x \in T_{i,s}} G_{i,s}^{(p)}(x). \quad (\text{S4})$$

We translate the same target shape to  $K \geq 19$  valid, bbox-disjoint locations  $N_{i,s}^{(1)}, \dots, N_{i,s}^{(K)}$  and compute their masses  $m_{i,s,k}^{0,(p)}$ . The empirical spatial rank and binary decision are

$$r_{i,s}^{(p)} = \frac{1 + \sum_{k=1}^K \mathbf{1}[m_{i,s,k}^{0,(p)} \geq m_{i,s}^{(p)}]}{K + 1}, \quad (\text{S5})$$

$$V_{i,s}^{(p)} = \mathbf{1}[r_{i,s}^{(p)} \leq .05]. \quad (\text{S6})$$

This is an attribution-mass comparison, not IoU with a thresholded heatmap. With  $K = 19$ ,  $V = 1$  only when the target strictly outranks every translation. most samples use  $K = 100$ . For a panel of  $N$  valid samples, the marginal localization rate and VxA risk difference are

$$\hat{\pi}_{V,s}^{(p)} = \frac{1}{N} \sum_{i=1}^N V_{i,s}^{(p)}, \quad (\text{S7})$$

$$\widehat{\text{RD}}_s^{(p)} = \Pr(V_s^{(p)} = 1 \mid A^{(p)} = 1) - \Pr(V_s^{(p)} = 1 \mid A^{(p)} = 0). \quad (\text{S8})$$

We call  $r$  an empirical rank rather than an exact  $p$ -value because translated locations are designed spatial controls, not exchangeable generative draws.

### S3.4. Intervention and Cross-Analysis Measures

**Exact intervention edits.** This paragraph specifies the edits introduced in main Sec. 4.3 and Fig. 2. For target and matched-control supports, selected pixels are replaced by the rounded per-channel mean of the pixels outside that support in the same frame. For full-frame support, every pixel is replaced by the rounded per-channel mean of the original frame. Memory and joint interventions apply the operation independently to each active frame using that frame’s own mean. The instruction and all nonvisual inputs remain fixed.

**Exact attribution-intervention agreement.** This paragraph expands the continuous comparison in main Sec. 6.4. The 23 reported correlations use the landmark-bearing local instruction segment and attention $\times$ gradient. For each sample, we define landmark-specific attribution as the target mass minus the mean translated-control mass:

$$\Delta m_{i,s}^{(p)} = m_{i,s}^{(p)} - \frac{1}{K} \sum_{k=1}^K m_{i,s,k}^{0,(p)}. \quad (\text{S9})$$

The intervention keeps the full instruction fixed and uses the same temporal scope  $s$ . For each model, native output branch, and scope, we compute Spearman correlation between  $\Delta m_{i,s}^{(p)}$  and the matched target-control effect  $E_{i,s}^{(p)}$  over their jointly valid samples.

808 **S3.5. Model-Specific Inference and Attribution**

809 **Native action mappings.** Table S4 supplements the released action outputs in main Table 2 and the common action diagnostic  
 810 in main Sec. 4.2.

Table S4. Exact native decoding and mappings to the local action variable. *Ordered regret* refers to the ordered-prefix diagnostic defined above, including collision and nonterminal STOP checks.

| Model        | Native inference and parsing                                                                                                                                    | Mapping to $A$                                                                                                                                                              |
|--------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| OneVLA       | Argmax over the first four rows of the released $5 \times 11$ action probabilities.                                                                             | Four-action ordered regret with $15^\circ$ turns.                                                                                                                           |
| InternVLA-NI | Final System-2 output after at most one look-down step, mapped by the released parser to an action sequence.                                                    | Released discretizer, horizon $h \leq 4$ , then ordered regret with $15^\circ$ turns. Empty or malformed output gives $A = 0$ .                                             |
| OmniVLA      | Eager forward pass; the $8 \times 4$ waypoint output uses the released 0.1 m position scale.                                                                    | Released waypoint-to-action discretizer, $h \leq 4$ , then ordered regret with $15^\circ$ turns.                                                                            |
| Uni-NaVid    | Sampling with temperature .5, maximum 1024 tokens, and a reset decoder cache; strict parsing retains at most four text actions.                                 | Ordered regret with $30^\circ$ turns. STOP truncates the sequence; invalid or nonterminal-STOP outputs fail the diagnostic.                                                 |
| FantasyVLN   | SDPA generation with temperature $10^{-6}$ , repetition penalty 1.05, maximum 64 tokens. The released wrapper is parsed strictly with random fallback disabled. | Predicted chunk of at most five actions, ordered regret with $30^\circ$ turns. STOP is absorbing; invalid parses give $A = 0$ and an attribution abstention.                |
| StreamVLN    | Greedy decoding, one beam, maximum 32 tokens. Strict four-action parsing.                                                                                       | Four-action ordered regret with $15^\circ$ turns; malformed output or nonterminal STOP fails.                                                                               |
| VAMOS        | Greedy decoding, one beam, maximum 16 tokens. Exactly five image-coordinate waypoints are required.                                                             | The first waypoint defines a bearing from image bottom-center. Quantization at $\pm 7.5^\circ$ must match the first expert primitive; later waypoints are not part of $A$ . |
| TIC-VLA      | Sampling with temperature .7, maximum 200 tokens. Strict parsing yields 30 cumulative robot-frame $(x, y)$ waypoints.                                           | The first waypoint bearing is quantized at $\pm 7.5^\circ$ and compared with the first expert primitive; later waypoints are not part of $A$ .                              |

**Attribution scores and intervention responses.** Table S5 supplements the attribution estimator in main Sec. 4.2 and the native response measure in main Sec. 4.3. Attribution requires a differentiable scalar, while intervention requires a native response measure. The table records both choices. For temporal models, model-specific processing exposes current and history visual grids separately so that C, M, and CM remain scope matched.

811  
812  
813  
814

Table S5. Model-native attribution targets and intervention response measures. Log-probability harm is the decrease assigned to the clean decoded action/path under an edited input.

| Model        | Scalar differentiated for VxA                                                                                                          | Native response measure $d_p$                                                                                                                                                                                               |
|--------------|----------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| OneVLA       | At each of five output steps, select the clean argmax among the first four navigation channels, then sum the five selected components. | L2 change over the navigation-only $5 \times 4$ output.                                                                                                                                                                     |
| InternVLA-N1 | Sum of the teacher-forced log-probabilities of the final greedy System-2 continuation. The sampled System-1 response is fixed.         | Waypoint RMS in metres when clean and edited outputs both remain continuous; normalized edit distance when both remain discrete. Branch/shape changes are retained separately, and the two native metrics are never pooled. |
| OmniVLA      | Inner product of the differentiable $8 \times 4$ waypoint output with the detached unit direction of the clean output.                 | XY waypoint RMS in metres.                                                                                                                                                                                                  |
| Uni-NaVid    | Sum of teacher-forced clean-action token log-probabilities.                                                                            | Clean-action log-probability harm.                                                                                                                                                                                          |
| FantasyVLN   | Sum of clean special-action-token log-probabilities, excluding wrapper tokens.                                                         | Clean-action log-probability harm.                                                                                                                                                                                          |
| StreamVLN    | For each generated navigation-action token, its teacher-forced log-probability. Synthetic padding STOP tokens are excluded.            | Clean-action log-probability harm.                                                                                                                                                                                          |
| VAMOS        | Mean teacher-forced log-probability of the generated waypoint/path continuation.                                                       | Clean-path log-probability harm.                                                                                                                                                                                            |
| TIC-VLA      | Mean elementwise inner product between the differentiable $30 \times 2$ waypoint output and the detached clean output.                 | Position RMS over all 30 cumulative waypoints.                                                                                                                                                                              |

**Exact attribution-map construction.** This paragraph and Table S6 supplement the attribution estimators and temporal operations in main Sec. 4.2. Table S5 defines the model-specific action score  $s$ . For OneVLA, InternVLA-N1, OmniVLA, Uni-NaVid, FantasyVLN, StreamVLN, and VAMOS, the differentiated object is the post-softmax attention probability  $a_{\ell,h,q,v}$ . Here  $\ell$  indexes a selected layer or action-head capture,  $h$  an attention head,  $q$  an action-related query, and  $v$  a visual token. Before normalization, these seven implementations compute

815  
816  
817  
818  
819

$$\begin{aligned}
 R_{\text{att}}(v) &\propto \sum_{\ell} \text{mean}_{h,q} a_{\ell,h,q,v}, \\
 R_{\text{grad}}(v) &\propto \sum_{\ell} \text{mean}_{h,q} \left| \frac{\partial s}{\partial a_{\ell,h,q,v}} \right|, \\
 R_{\text{att} \times \text{grad}}(v) &\propto \sum_{\ell} \text{mean}_{h,q} \left| a_{\ell,h,q,v} \frac{\partial s}{\partial a_{\ell,h,q,v}} \right|.
 \end{aligned} \tag{S10}$$

820

For StreamVLN, these equations are evaluated separately with  $s = s_t$  and  $R_m = R_{m,t}$  for each generated action token. Except for the action-wise StreamVLN order below, each method  $m \in \{\text{att}, \text{grad}, \text{att} \times \text{grad}\}$  is normalized over the model-specific visual-token domain  $\Omega_{p,s}$ :

821  
822  
823

$$G_m(v) = \frac{R_m(v)}{\sum_{u \in \Omega_{p,s}} R_m(u)}. \tag{S11}$$

824

Implementations may average rather than sum over  $\ell$ . The resulting positive scale cancels under this normalization. StreamVLN separately normalizes the map for each of its  $T$  generated action tokens and then averages the maps:

825  
826

$$G_m(v) = \frac{1}{T} \sum_{t=1}^T \frac{R_{m,t}(v)}{\sum_{u \in \Omega_{p,s}} R_{m,t}(u)}. \tag{S12}$$

827

TIC-VLA instead differentiates  $s$  with respect to the projected source-token feature  $z_{v,d}$ . It first aggregates attention and feature gradient as

$$\begin{aligned}\bar{a}(v) &= \text{mean}_{\ell,q} a_{\ell,q,v}, \\ g(v) &= \text{mean}_d \left| \frac{\partial s}{\partial z_{v,d}} \right|, \\ R_{\text{att}}(v) &= \bar{a}(v), \\ R_{\text{grad}}(v) &= g(v), \\ R_{\text{att} \times \text{grad}}(v) &= \bar{a}(v)g(v).\end{aligned}\tag{S13}$$

The three TIC-VLA raw maps then use the same L1 normalization over its visual domain. Table S6 records the aggregation order and visual domain used by every model.

Table S6. Model-specific attribution-map construction. Seven models differentiate the action score with respect to post-softmax attention probabilities, while TIC-VLA differentiates it with respect to projected source-token features. The final column gives the full visual input.  $\Omega_{p,s}$  is its active current, memory, or joint subset for temporal scope  $s$ .

| Model        | Differentiated variable | Aggregation order                                                                                                                                                  | Full visual input                                                                   |
|--------------|-------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
| OneVLA       | Attention probability   | Mean heads and action queries.<br>Aggregate DiT cross-attention captures across all blocks and denoising steps.<br>Multimodal backbone detached.                   | Visual tokens from the current and four history frames.<br>All input visual tokens. |
| InternVLA-N1 | Attention probability   | Mean heads and continuation queries.<br>Aggregate the final four decoder layers.                                                                                   | Released LOOKDOWN tokens included when present.<br>256 current-image tokens.        |
| OmniVLA      | Attention probability   | Mean heads and action-query positions.<br>Aggregate the final four decoder layers.                                                                                 | Goal-image and proprioceptive tokens excluded.                                      |
| Uni-NaVid    | Attention probability   | Mean heads and action-token positions.<br>Aggregate the final four decoder layers.                                                                                 | 84 visual tokens: 16 history and 68 current.                                        |
| FantasyVLN   | Attention probability   | Mean heads and action-token positions.<br>Aggregate the final four decoder layers.<br>For each action token, mean heads and the final four layers, normalize, then | Seven grids: four history and left, front, and right current views.                 |
| StreamVLN    | Attention probability   | mean the action-token maps.<br>Mean heads and continuation queries.                                                                                                | Five $14 \times 14$ visual grids.                                                   |
| VAMOS        | Attention probability   | Aggregate the final four decoder layers.<br>Mean head-averaged attention over 30 action queries and three cross-attention layers. Mean                             | 256 current-image tokens.                                                           |
| TIC-VLA      | Source-token feature    | absolute feature gradients over channels. Multiply the two token-level scores after aggregation.                                                                   | Five $16 \times 16$ visual grids.<br>State tokens excluded.                         |

## S4. Statistical Protocol

### S4.1. Trajectory-Group Confidence Intervals

**Bootstrap intervals.** This paragraph expands the confidence intervals reported in main Sec. 5 and main Tables 3 and 4. For each analysis, we first restrict to valid samples and group them by source navigation trajectory. Each bootstrap replicate samples the same number of trajectory groups with replacement and retains every sample from each selected group. The statistic is recomputed from the resulting sample rows. Matched mean-effect replicates average over all resampled samples. We use 5,000 replicates for VxA risk differences and matched mean effects, and 2,000 replicates for Spearman correlations. The 95% confidence interval is given by the 2.5th and 97.5th percentiles of the bootstrap replicates for which the statistic is defined. Marginal localization rates are reported as observed fractions.

### S4.2. VxA Association Tests

**Fisher tests and multiplicity.** This paragraph defines the tests behind the VxA results in main Sec. 6.1 and Table 3. Each panel uses a two-sided Fisher exact test on its sample-level  $A \times V$  table. For each temporal model, the nine reported

attribution-method and scope panels form one Holm family. For each current-only model, the three attribution-method panels form one family. We call an association positively supported when  $\widehat{RD} > 0$  and  $p_H < .05$ , and negatively supported when  $\widehat{RD} < 0$  and  $p_H < .05$ . The Fisher values determine the Holm decision. Trajectory-group bootstrap intervals report uncertainty in  $\widehat{RD}$ , and  $\widehat{\pi}_V$  is reported as the observed localization rate.

### S4.3. Matched-Effect Tests

**Trajectory sign randomization and multiplicity.** This paragraph defines the tests behind the intervention results in main Sec. 6.2 and Table 4. For source trajectory  $g$ , let  $u_g = \sum_{i \in g} E_i$  and let  $N$  be the number of eligible samples. The observed statistic is

$$T_{\text{obs}} = N^{-1} \sum_{g \in \mathcal{G}} u_g. \quad (\text{S14})$$

For each of  $B = 10,000$  replicates, independent Rademacher signs  $\eta_g \in \{-1, +1\}$  produce  $T_b = N^{-1} \sum_g \eta_g u_g$ . The one-sided empirical score is

$$p = \frac{1 + \sum_{b=1}^B \mathbf{1}[T_b \geq T_{\text{obs}}]}{B + 1}. \quad (\text{S15})$$

For each model and native output branch, the available scopes form one Holm family. A configuration supports a landmark-favored mean when  $\bar{E}_s > 0$  and its Holm-adjusted one-sided randomization value is below .05. Trajectory-group bootstrap intervals report uncertainty in  $\bar{E}_s$ .

860 **S5. Complete Results**861 **S5.1. Complete VxA Results**

862 **All attribution methods.** Two of 60 panels support a positive association after within-model Holm correction: StreamVLN  
 863 gradient in C and CM. Table S7 supplements main Sec. 6.1 and Table 3 with all methods, scopes, intervals, and adjusted tests.

Table S7. Complete inference for the 60 VxA panels in main Table 3. Each method cell reports  $\hat{\pi}_V$  (%) |  $\widehat{RD}$  [trajectory-group 95% CI] |  $p_H$ . Here  $p_H$  is the within-model Holm-adjusted two-sided Fisher value. The correction covers nine method-scope panels for each temporal model and three method panels for each current-only model.  $N$  is the valid sample count and  $A$  is the action-acceptance rate (%). Only applicable scopes are shown.

| Model        | Scope | $N$   | $A$  | Attention |         |                  | Gradient |       |         | Attention $\times$ gradient |       |       |         |                  |       |
|--------------|-------|-------|------|-----------|---------|------------------|----------|-------|---------|-----------------------------|-------|-------|---------|------------------|-------|
| OneVLA       | C     |       |      | 42.47     | [-.0084 | [-.0369,.0200]]  | 1.000    | 28.76 | [-.0375 | [-.0627,-.0122]]            | .0111 | 32.44 | [-.0230 | [-.0501,.0052]]  | .3765 |
|              | M     | 7,228 | 69.0 | 45.43     | [-.0241 | [-.0049,.0530]]  | .3765    | 21.47 | [-.0076 | [-.0156,.0303]]             | 1.000 | 25.29 | [-.0018 | [-.0264,.0221]]  | 1.000 |
|              | CM    |       |      | 47.59     | [-.0094 | [-.0190,.0392]]  | 1.000    | 25.26 | [-.0119 | [-.0366,.0123]]             | 1.000 | 29.11 | [-.0235 | [-.0499,.0035]]  | .3505 |
| InternVLA-N1 | C     |       |      | 55.96     | [-.0036 | [-.0247,.0321]]  | 1.000    | 49.90 | [-.0176 | [-.0450,.0098]]             | 1.000 | 56.17 | [-.0038 | [-.0245,.0314]]  | 1.000 |
|              | M     | 7,228 | 71.4 | 52.34     | [-.0120 | [-.0400,.0164]]  | 1.000    | 64.15 | [-.0083 | [-.0342,.0182]]             | 1.000 | 52.21 | [-.0110 | [-.0384,.0175]]  | 1.000 |
|              | CM    |       |      | 57.33     | [-.0057 | [-.0337,.0224]]  | 1.000    | 65.38 | [-.0182 | [-.0439,.0075]]             | 1.000 | 57.51 | [-.0079 | [-.0351,.0197]]  | 1.000 |
| OmniVLA      | C     | 7,228 | 87.3 | 51.70     | [-.0257 | [-.0138,.0634]]  | .3345    | 13.77 | [-.0199 | [-.0491,.0089]]             | .3345 | 51.44 | [-.0215 | [-.0175,.0598]]  | .3345 |
| Uni-NaVid    | C     |       |      | 54.77     | [-.0338 | [-.0583,-.0072]] | .0294    | 39.64 | [-.0159 | [-.0087,.0408]]             | .5100 | 55.11 | [-.0291 | [-.0539,-.0031]] | .0622 |
|              | M     | 7,228 | 52.4 | 47.07     | [-.0393 | [-.0659,-.0127]] | .0079    | 34.35 | [-.0144 | [-.0094,.0389]]             | .5100 | 47.66 | [-.0296 | [-.0554,-.0035]] | .0622 |
|              | CM    |       |      | 54.70     | [-.0351 | [-.0595,-.0091]] | .0230    | 40.33 | [-.0092 | [-.0150,.0348]]             | .5100 | 54.79 | [-.0330 | [-.0584,-.0067]] | .0314 |
| FantasyVLN   | C     |       |      | 56.45     | [-.0049 | [-.0307,.0207]]  | 1.000    | 35.78 | [-.0241 | [-.0001,.0479]]             | .2955 | 53.10 | [-.0227 | [-.0488,.0028]]  | .4162 |
|              | M     | 7,222 | 42.4 | 54.57     | [-.0029 | [-.0235,.0286]]  | 1.000    | 49.18 | [-.0085 | [-.0162,.0337]]             | 1.000 | 50.93 | [-.0320 | [-.0571,-.0058]] | .0688 |
|              | CM    |       |      | 58.76     | [-.0002 | [-.0251,.0244]]  | 1.000    | 50.57 | [-.0168 | [-.0088,.0422]]             | .8005 | 55.34 | [-.0191 | [-.0443,.0065]]  | .6518 |
| StreamVLN    | C     |       |      | 40.19     | [-.0154 | [-.0107,.0411]]  | .9914    | 38.95 | [-.0460 | [-.0207,.0710]]             | .0009 | 40.02 | [-.0169 | [-.0093,.0427]]  | .9493 |
|              | M     | 7,228 | 62.2 | 45.49     | [-.0099 | [-.0159,.0354]]  | 1.000    | 55.41 | [-.0275 | [-.0007,.0532]]             | .1627 | 44.26 | [-.0066 | [-.0189,.0323]]  | 1.000 |
|              | CM    |       |      | 45.42     | [-.0059 | [-.0199,.0319]]  | 1.000    | 56.00 | [-.0406 | [-.0143,.0669]]             | .0065 | 44.66 | [-.0007 | [-.0247,.0266]]  | 1.000 |
| VAMOS        | C     | 7,228 | 37.0 | 28.64     | [-.0058 | [-.0184,.0295]]  | 1.000    | 21.14 | [-.0060 | [-.0146,.0272]]             | 1.000 | 30.80 | [-.0065 | [-.0307,.0178]]  | 1.000 |
| TIC-VLA      | C     |       |      | 14.65     | [-.0042 | [-.0166,.0254]]  | 1.000    | 12.40 | [-.0035 | [-.0160,.0228]]             | 1.000 | 11.41 | [-.0085 | [-.0106,.0274]]  | 1.000 |
|              | M     | 7,228 | 30.5 | 52.38     | [-.0486 | [-.0775,-.0199]] | .0013    | 41.91 | [-.0282 | [-.0559,-.0015]]            | .2084 | 38.24 | [-.0224 | [-.0499,.0042]]  | .5180 |
|              | CM    |       |      | 19.66     | [-.0085 | [-.0321,.0153]]  | 1.000    | 18.47 | [-.0103 | [-.0323,.0126]]             | 1.000 | 13.43 | [-.0024 | [-.0228,.0178]]  | 1.000 |

**S5.2. Broad Visual Response**

864

**Full-frame edits establish broad visual response.** Full-frame edits yield a positive mean response in all 23 configurations. Magnitudes use model-native scales. Table S8 reports valid sample counts and native-scale means for every configuration. InternVLA-N1 retains branch-comparable clean-edited pairs.

865

866

867

Table S8. Complete results for the full-frame visual intervention.  $N_F$  is the native-comparable sample count, and  $\overline{D^F}$  is the mean response after editing all visual inputs in the reported scope. Values use the model and branch-specific measures in Table S5. Dashes indicate scopes that are not applicable.

| Model/branch         | Scope | $N_F$ | $\overline{D^F}$ |
|----------------------|-------|-------|------------------|
| OneVLA               | C     | 7,228 | 1.2558           |
|                      | M     | 7,228 | 1.2524           |
|                      | CM    | 7,228 | 2.3358           |
| InternVLA-N1 (cont.) | C     | 2,217 | .1105            |
|                      | M     | 2,815 | .1709            |
|                      | CM    | 31    | .3409            |
| InternVLA-N1 (disc.) | C     | 2,677 | .4856            |
|                      | M     | 2,606 | .5211            |
|                      | CM    | 2,783 | .6093            |
| OmniVLA              | C     | 7,228 | .05986           |
| Uni-NaVid            | C     | 7,228 | 3.3812           |
|                      | M     | 7,228 | .3985            |
|                      | CM    | 7,228 | 5.5151           |
| FantasyVLN           | C     | 7,225 | 7.7829           |
|                      | M     | 7,225 | 7.6620           |
|                      | CM    | 7,225 | 15.8745          |
| StreamVLN            | C     | 7,228 | 1.3062           |
|                      | M     | 7,228 | .3053            |
|                      | CM    | 7,228 | .8570            |
| VAMOS                | C     | 7,228 | .19340           |
| TIC-VLA              | C     | 7,228 | .13271           |
|                      | M     | 7,228 | .46347           |
|                      | CM    | 7,228 | .49705           |

### S5.3. Matched Visual Intervention Results

**Landmark-specific response.** Six of 23 configurations favor the landmark over its control. Ten intervals lie below zero and seven include zero. Table S9 supplements main Sec. 6.2 and Table 4 with all estimates and tests.

Table S9. Complete matched-intervention results for main Table 4.  $\bar{E}_s$  is the sample mean under the native response measure in Table S5.  $E_{i,s} > 0$  is the target-favored fraction, and  $p_H$  is the within-native-family Holm-adjusted one-sided randomization value.

| Model/branch         | Scope | $N$   | $\bar{E}_s$ | 95% CI               | $E_{i,s} > 0$ (%) | $p_H$ |
|----------------------|-------|-------|-------------|----------------------|-------------------|-------|
| OneVLA               | C     | 7,228 | -.01975     | [-.03119, -.00763]   | 50.9              | 1.000 |
|                      | M     | 7,228 | +.01757     | [.00545, .02973]     | 54.3              | .0087 |
|                      | CM    | 7,228 | -.09455     | [-.10936, -.07955]   | 43.0              | 1.000 |
| InternVLA-N1 (cont.) | C     | 4,171 | +.001807    | [-.003446, .007057]  | 53.5              | .2651 |
|                      | M     | 4,150 | +.005948    | [-.000535, .012465]  | 55.5              | .0846 |
|                      | CM    | 3,995 | +.009383    | [.002147, .016774]   | 55.2              | .0222 |
| InternVLA-N1 (disc.) | C     | 2,622 | -.01306     | [-.02378, -.00201]   | 6.4               | 1.000 |
|                      | M     | 2,545 | -.02338     | [-.03592, -.01093]   | 8.1               | 1.000 |
|                      | CM    | 2,480 | -.03690     | [-.05041, -.02342]   | 8.8               | 1.000 |
| OmniVLA              | C     | 7,228 | +.001910    | [.001626, .002202]   | 55.0              | .0001 |
| Uni-NaVid            | C     | 7,228 | +.01003     | [-.00145, .02134]    | 51.3              | .1422 |
|                      | M     | 7,228 | -.01643     | [-.02334, -.00945]   | 46.3              | 1.000 |
|                      | CM    | 7,228 | -.00497     | [-.01879, .00821]    | 49.5              | 1.000 |
| FantasyVLN           | C     | 7,225 | -.0373      | [-.2321, .1577]      | 49.6              | 1.000 |
|                      | M     | 7,225 | +.1068      | [-.0438, .2578]      | 49.2              | .2544 |
|                      | CM    | 7,225 | -.5450      | [-.8053, -.2894]     | 47.4              | 1.000 |
| StreamVLN            | C     | 7,228 | -.06540     | [-.07331, -.05774]   | 44.6              | 1.000 |
|                      | M     | 7,228 | -.00407     | [-.00815, .00008]    | 49.0              | 1.000 |
|                      | CM    | 7,228 | -.07838     | [-.08564, -.07092]   | 39.1              | 1.000 |
| VAMOS                | C     | 7,228 | +.002786    | [.000068, .005325]   | 53.0              | .0193 |
| TIC-VLA              | C     | 7,228 | -.004544    | [-.005109, -.003968] | 42.7              | 1.000 |
|                      | M     | 7,228 | +.010748    | [.006386, .015070]   | 52.0              | .0003 |
|                      | CM    | 7,228 | +.009710    | [.005252, .014056]   | 51.4              | .0003 |

### S5.4. Attribution-Intervention Association

**Weak continuous association.** All 23 correlations are positive, ranging from .016 to .182, and 20 intervals lie above zero. Table S10 supplements main Sec. 6.4 with all estimates, jointly valid sample counts, and intervals.

Table S10. Spearman  $\rho$  between continuous landmark attribution and the matched target-control response. The  $N$  column gives jointly valid sample counts in C/M/CM order. Brackets give trajectory-group 95% CIs. Native output branches are reported separately, and dashes indicate scopes that are not applicable.

| Model/branch         | $N_C/N_M/N_{CM}$  | C $\rho$ [95% CI]  | M $\rho$ [95% CI]  | CM $\rho$ [95% CI] |
|----------------------|-------------------|--------------------|--------------------|--------------------|
| OneVLA               | 7,228/7,228/7,228 | .100 [.074, .126]  | .105 [.079, .133]  | .106 [.081, .131]  |
| InternVLA-N1 (cont.) | 4,171/4,150/3,995 | .087 [.054, .119]  | .119 [.086, .151]  | .144 [.111, .178]  |
| InternVLA-N1 (disc.) | 2,622/2,545/2,480 | .031 [-.010, .070] | .047 [.007, .087]  | .073 [.032, .115]  |
| OmniVLA              | 7,228/—/—         | .182 [.156, .209]  | —                  | —                  |
| Uni-NaVid            | 7,228/7,228/7,228 | .044 [.017, .070]  | .016 [-.011, .043] | .057 [.030, .083]  |
| FantasyVLN           | 7,219/7,219/7,219 | .035 [.011, .058]  | .034 [.010, .059]  | .046 [.022, .070]  |
| StreamVLN            | 7,228/7,228/7,228 | .045 [.017, .072]  | .065 [.039, .092]  | .033 [.006, .059]  |
| VAMOS                | 7,228/—/—         | .030 [.005, .054]  | —                  | —                  |
| TIC-VLA              | 7,228/7,228/7,228 | .118 [.093, .143]  | .020 [-.003, .044] | .024 [.001, .048]  |

## S6. Additional Analyses

### S6.1. Heterogeneity by Source Benchmark and Split

**Stratified analysis.** This subsection extends the dataset composition in Sec. S2.1 and the two analyses in main Secs. 6.1 and 6.2. We repeat the VxA and matched-intervention calculations separately for R2R and RxR, and for the train, val-seen, and

val-unseen splits. The calculation within each group is unchanged and uses the samples assigned to that group.

878

**VxA remains weak within each source and split.** Table S11 reports the 60-panel summary within each group. Mean landmark localization ranges from 41.44% to 45.32%. Mean RD is negative in all five groups, from -.12 percentage points for train to -1.77 for R2R. The number of positive RDs ranges from 20 of 60 for val-unseen to 34 of 60 for train.

879

880

881

Table S11. VxA results stratified by source benchmark and split.  $N$  is the dataset sample count. Landmark localization ( $V$ ) and RD are computed within each of the 60 model-method-scope panels and then averaged across panels. Positive RD counts panels with RD above zero.

| Group  | Value      | $N$   | $V$ (%) | RD (pp) | Positive RD |
|--------|------------|-------|---------|---------|-------------|
| Source | R2R        | 726   | 45.32   | -1.77   | 21/60       |
|        | RxR        | 6,502 | 41.86   | -.32    | 29/60       |
| Split  | Train      | 5,622 | 41.86   | -.12    | 34/60       |
|        | Val-seen   | 615   | 41.44   | -.31    | 29/60       |
|        | Val-unseen | 991   | 44.61   | -1.58   | 20/60       |

**Matched effects vary by configuration rather than one data source.** Table S12 gives each matched effect within the same five groups. Positive means occur in 15 of 23 configurations for R2R, 10 for RxR, 9 for train, 10 for val-seen, and 11 for val-unseen. Twelve configurations keep the same sign in every group. Every group contains both landmark-favoured and control-favoured means.

882

883

884

885

Table S12. Matched-intervention effects stratified by source benchmark and split. Each cell is the mean target-minus-control effect over eligible samples in that group. Positive values favour the landmark. Native response scales differ across rows, so comparisons are made within each row.

| Model/branch         | Scope | Full    | R2R     | RxR     | Train   | Val-seen | Val-unseen |
|----------------------|-------|---------|---------|---------|---------|----------|------------|
| OneVLA               | C     | -.01975 | +.03773 | -.02616 | -.02170 | -.02978  | -.00244    |
|                      | M     | +.01757 | +.07256 | +.01143 | +.02123 | -.00980  | +.01382    |
|                      | CM    | -.09455 | -.00652 | -.10438 | -.09442 | -.08515  | -.10110    |
| InternVLA-N1 (cont.) | C     | +.00181 | +.00292 | +.00165 | -.00015 | +.00540  | +.01086    |
|                      | M     | +.00595 | +.00986 | +.00542 | +.00384 | -.00020  | +.02238    |
|                      | CM    | +.00938 | +.01417 | +.00872 | +.00764 | +.00215  | +.02419    |
| InternVLA-N1 (disc.) | C     | -.01306 | -.02546 | -.01215 | -.01879 | +.01010  | +.00384    |
|                      | M     | -.02338 | -.01013 | -.02439 | -.02405 | -.04889  | -.00290    |
|                      | CM    | -.03690 | -.04762 | -.03608 | -.04021 | -.01852  | -.02988    |
| OmniVLA              | C     | +.00191 | +.00227 | +.00187 | +.00203 | +.00145  | +.00153    |
| Uni-NaVid            | C     | +.01003 | +.04352 | +.00629 | +.01356 | +.00232  | -.00524    |
|                      | M     | -.01643 | -.03208 | -.01468 | -.01914 | -.01731  | -.00053    |
|                      | CM    | -.00497 | +.01072 | -.00672 | -.00691 | -.00690  | +.00722    |
| FantasyVLN           | C     | -.03733 | +.03947 | -.04589 | -.00940 | -.07255  | -.17384    |
|                      | M     | +.10685 | +.47202 | +.06618 | +.08288 | +.33450  | +.10145    |
|                      | CM    | -.54504 | +.31705 | -.64105 | -.61772 | -.38559  | -.23191    |
| StreamVLN            | C     | -.06540 | -.07954 | -.06382 | -.06611 | -.06742  | -.06015    |
|                      | M     | -.00407 | +.00543 | -.00513 | -.00593 | +.00212  | +.00265    |
|                      | CM    | -.07838 | -.08033 | -.07816 | -.07738 | -.08276  | -.08133    |
| VAMOS                | C     | +.00279 | +.00192 | +.00288 | +.00295 | +.00921  | -.00212    |
| TIC-VLA              | C     | -.00454 | -.00425 | -.00458 | -.00460 | -.00443  | -.00432    |
|                      | M     | +.01075 | +.02658 | +.00898 | +.00914 | +.01846  | +.01511    |
|                      | CM    | +.00971 | +.02547 | +.00795 | +.00799 | +.01691  | +.01498    |

## S6.2. Heterogeneity by Landmark Apparent Size

886

**Apparent-size distribution.** This analysis supplements the dataset construction in main Sec. 3 and Table 1. Current landmark apparent size is the percentage of current-image pixels covered by the verified target mask. Across the 7,228 samples, it ranges from .107% to 14.958%, with a median of 4.628% and an interquartile range of 2.477%-7.766%. Figure S1 shows

887

888

889

890 one-percentage-point intervals. The distribution peaks at 2%-3%, with 907 samples (12.55%). Table S13 uses ten similarly  
891 sized groups for the comparisons below.

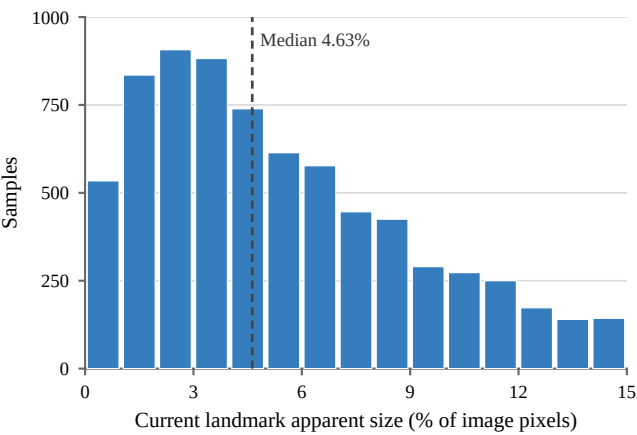


Figure S1. Distribution of current landmark apparent size. Apparent size is the percentage of current-image pixels inside the verified landmark mask. Bars use one-percentage-point intervals over all 7,228 samples. The dashed line marks the 4.628% median.

892 **Localization increases with apparent size.** This result supplements main Sec. 6.1 and Table 3. Mean localization rises  
893 from 28.54% in D1 to 45.39% in D10. It increases with log apparent size in 22 of 24 current-view model-method panels, with  
894 a median increase of 4.52 percentage points for each doubling of apparent size.

895 **VxA association remains weak across apparent size.** This result also supplements main Sec. 6.1 and Table 3. Mean RD is  
896 -.39 percentage points in D1 and -.90 in D10, with 10 of 24 panels positive in both groups. RD increases with log apparent size  
897 in 14 of 24 panel-wise fits, with a median change of only +.40 percentage points per doubling.

Table S13. Current-view VxA summaries by current landmark apparent-size decile. *N* gives the dataset sample count. *V* and RD are first computed within each of the 24 model-method panels and then averaged across panels. Positive RD counts panels with RD above zero.

| Decile | Mask area (%)   | <i>N</i> | <i>V</i> (%) | RD (pp) | Positive RD |
|--------|-----------------|----------|--------------|---------|-------------|
| D1     | [.107,1.241]    | 723      | 28.54        | -.39    | 10/24       |
| D2     | (1.241,2.079]   | 723      | 31.65        | -1.37   | 10/24       |
| D3     | (2.079,2.850]   | 724      | 35.56        | +1.12   | 14/24       |
| D4     | (2.850,3.696]   | 721      | 36.27        | +.64    | 14/24       |
| D5     | (3.696,4.628]   | 723      | 36.57        | +.43    | 10/24       |
| D6     | (4.628,5.662]   | 723      | 39.85        | +1.95   | 14/24       |
| D7     | (5.662,6.951]   | 722      | 39.27        | +.83    | 14/24       |
| D8     | (6.951,8.602]   | 723      | 44.22        | -.19    | 14/24       |
| D9     | (8.602,10.932]  | 723      | 44.20        | +.52    | 14/24       |
| D10    | (10.932,14.958] | 723      | 45.39        | -.90    | 10/24       |

**Matched specificity does not consistently increase with apparent size.** This result supplements main Sec. 6.2 and Table 4. Target and equal-area control responses both increase with size across all nine model/branch rows, with correlations of .098-.484 and .099-.584, respectively. Their matched difference has six negative and three positive correlations with size, ranging from -.113 to +.071. Table S14 reports the row-level values. 898 899 900 901

Table S14. Current-view matched effects by current landmark apparent size.  $\bar{E}_C$  is the full-sample matched effect,  $\rho_{S,E}$  is its Spearman correlation with apparent size, and D1 and D10 give the means in the smallest and largest deciles. Native response scales differ across rows.

| Model/branch         | $\bar{E}_C$ | $\rho_{S,E}$ | D1       | D10      |
|----------------------|-------------|--------------|----------|----------|
| OneVLA               | -.0197      | -.033        | +.00949  | -.04520  |
| InternVLA-N1 (cont.) | +.00181     | -.090        | +.00978  | -.01132  |
| InternVLA-N1 (disc.) | -.0131      | -.025        | -.01427  | -.02683  |
| OmniVLA              | +.00191     | -.007        | +.001123 | +.000848 |
| Uni-NaVid            | +.0100      | +.020        | +.00154  | -.02192  |
| FantasyVLN           | -.0373      | +.003        | +.2545   | -.0788   |
| StreamVLN            | -.0654      | -.061        | -.00908  | -.11232  |
| VAMOS                | +.00279     | +.071        | -.00495  | +.01309  |
| TIC-VLA              | -.00454     | -.113        | -.000173 | -.01225  |