# Misrouting, Not Misperception: Locating GUI-Grounding Failures Inside VLMs

Anonymous ACL submission


## Abstract

Vision–language models (VLMs) now serve as the visual grounding component of GUI agents, yet click failures persist: it remains unclear whether they are perception failures, where no target information is encoded, or routing failures, where the target is encoded but never reaches the click output. Ridge probes on mid-layer hidden states recover the target under the click-hit criterion for up to 73% of failure cases across Qwen2.5-VL-7B, Qwen2.5-VL-3B, and OS-Atlas-7B, showing that failed clicks carry linearly decodable target information. However, this decodability is not causal: no probe-direction intervention moves the click beyond an equal-norm random baseline; instead, Distributed Alignment Search (DAS; Geiger et al., 2024) at the same layer identifies a subspace distinct from the probe direction, and rank-2 amnesic erasure of this subspace collapses the click hit rate up to $5.8\times$ more than erasing the probe's rank-2 subspace at matched rank. A learned routing map from the full $k{=}32$ DAS projection of the readout-layer hidden state into this subspace corrects more held-out failures than a rank-2 oracle that injects the ground-truth coordinate into the same subspace, on every model—up to 64% on Qwen2.5-VL-7B and Qwen2.5-VL-3B, although it uses no test-time ground truth: the $(x, y)$ value alone is not the limiting signal. Together, these results recast many such click failures as routing failures rather than perception failures: improving click grounding may depend less on perceiving the target than on routing what is already perceived.


## 1 Introduction

Vision–language models (VLMs) are increasingly used as the visual grounding component of GUI agents: given a screenshot and a natural-language instruction, they must identify the pixel coordinates of the intended click target (Wu et al., 2024; Qin et al., 2025; Cheng et al., 2024; Nguyen et al., 2025).

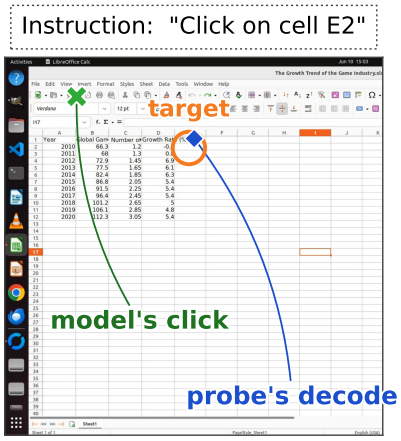


Figure 1: **A representative AgentNet failure with a linearly decodable target.** Ridge-probe prediction (blue ⋄), ground-truth target (orange ring, radius = click-hit threshold), and model click (green ×).

Yet click grounding remains a bottleneck on high-resolution professional screenshots: on ScreenSpot-Pro, a benchmark of professional interfaces with small targets, GPT-4o emits a correct click on fewer than 1% of items without agent scaffolding (Li et al., 2025). This gap raises two mechanistic possibilities: either the model never perceives the target, or it perceives the target but fails to route it into the click output. The distinction matters because current improvement efforts—larger grounding and computer-use datasets (Wu et al., 2024; Wang et al., 2025), dynamic-resolution visual processing (Wang et al., 2024), and end-to-end action training (Qin et al., 2025)—improve the data, visual encoding, or policy around coordinate prediction, but do not determine which failure mode occurs inside a trained model. No causal test inside the trained model has settled which reading is correct.

Following the mechanistic-interpretability convention of separating what a representation *con-*

*tains* from what it causally *drives* (Belinkov, 2022; Elazar et al., 2021; Hase et al., 2023), we first ask whether failed clicks lack target information. A ridge probe trained on mid-layer hidden states recovers the target under the click-hit criterion for 73%, 72%, and 65% of failure cases on Qwen2.5-VL-7B, Qwen2.5-VL-3B, and OS-Atlas-7B, respectively—the hidden state contains an answer the model does not emit (Figure 1).

If a target coordinate is linearly decodable, steering the hidden state along the probe direction toward the ground-truth coordinate should improve clicks, but it does not. At each model's probe-best layer (L14/L21/L19), additive steering toward the ground truth, steering along a random direction of equal norm, and amnesic erasure of the probe's rank-2 row space all leave the click hit rate within 5 points of the no-intervention baseline on every model. The probe direction is decodable but not causal: the probe reads a real signal that the native coordinate-emission pathway does not use.

To locate a mid-layer direction that does drive the click, we apply Distributed Alignment Search (DAS; Geiger et al., 2024) at each model's probe-best layer and stress-test the resulting $k$-dimensional subspace with two complementary causal interventions: donor interchange lifts coordinate transfer above the no-patch reference, and amnesic erasure collapses the click hit rate on every model. To rule out a rank-driven artifact, we contrast the two subspaces at matched rank: erasing the rank-2 DAS subspace produces $5.8\times$, $2.8\times$, and $1.7\times$ larger click-hit degradations than erasing the probe's rank-2 row space on Qwen2.5-VL-7B, Qwen2.5-VL-3B, and OS-Atlas-7B, respectively. The probe direction and the causal handle share a layer but not an axis.

Head-level ablations point in the same direction: probe-writing heads break the readout but not the click on all three models, consistent with a distributed click pathway. Routing the full $k{=}32$ readout-layer DAS projection into this subspace corrects 56% of held-out Qwen2.5-VL-7B failures (and 64% on Qwen2.5-VL-3B) with no ground truth at test time—exceeding, on every model, a *rank-2 oracle* that injects the ground-truth coordinate through the same map (0.44 and 0.37, respectively). The click pathway depends on click-relevant information that the $(x, y)$ coordinate alone does not carry, and that the rank-2 probe readout discards.

Our central contribution is to show that many GUI-grounding click failures are internal-routing failures rather than failures of visual perception: the target coordinate is often readable from one mid-layer representation, controllable through a distinct mid-layer subspace, and still disconnected from native coordinate emission. Most strikingly, routing a wider readout-layer signal into this subspace corrects more held-out failures than an oracle that injects the ground-truth coordinate through the same map—the $(x, y)$ value itself is not the limiting signal. Improving GUI click grounding may therefore depend less on perceiving the target than on routing what is already perceived.

## 2 Related Work

**Probing, decodability, and hidden-state readout.** Prior probing work warns that decodability does not imply use. Under the linear representation hypothesis, properties read by linear probes correspond to directions in representation space (Park et al., 2024), yet a direction that linearly encodes a concept need not be the one the network relies on (Bolukbasi et al., 2021). Amnesic probing (Elazar et al., 2021), linear concept erasure (Belrose et al., 2023b), and causal-mediation analysis (Vig et al., 2020) therefore ask how information is used rather than what is encoded. The logit lens and tuned lens (Belrose et al., 2023a) push the same question into hidden-state decoding, with a related body of work showing that models can retain recoverable latent knowledge even when their overt outputs are wrong (Mallen and Belrose, 2023).

**Causal abstraction and where to intervene.** Causal-abstraction methods provide tools for finding the representations a model actually uses. Interchange interventions (Geiger et al., 2022) replace a subspace component of one input's hidden state with another's, and Distributed Alignment Search (DAS) (Geiger et al., 2024) learns by gradient descent a distributed subspace carrying a target causal variable. We apply DAS to a continuous 2D coordinate emitted as digit tokens, rather than to validate a known symbolic algorithm. An alternative interventional family, additive steering vectors (Turner et al., 2023; Rimsky et al., 2024; Li et al., 2023; Zou et al., 2023), shifts hidden states along a single learned direction. A parallel line localizes computation to components—neurons, heads, or paths—via activation patching (Meng et al., 2022) and its fast attribution-patching approximation (Kramár et al., 2024). Most relevant to us, Hase et al. (2023) find

that where a fact is localized is not where editing it is effective; we extend this localization-versus-intervention gap to GUI coordinate grounding.

**GUI grounding VLMs and the coordinate-to-token bottleneck.** Prior GUI-grounding work identifies the coordinate-to-token interface as a bottleneck, but not its internal route. Grounding-specialized VLMs such as SeeClick (Cheng et al., 2024), OS-Atlas (Wu et al., 2024), and UI-TARS (Qin et al., 2025) emit click coordinates as digit tokens; benchmarks such as ScreenSpot-Pro (Li et al., 2025) and large-scale GUI corpora such as AgentNet (Wang et al., 2025) stress small targets on high-resolution interfaces. Inside such VLMs, causal-tracing tools localize object- and instruction-relevant content to specific layers and tokens (Palit et al., 2023), and Kang et al. (2025) show that a few frozen attention heads already localize the click target; we find that the target-attending heads are not the heads that drive the click. Wu et al. (2025) attribute the generation gap to weak spatial–semantic alignment between dense coordinates and coarse visual patches and propose coordinate-free emission, and Lin et al. (2026) locate a complementary bottleneck in the prefill stage; we instead characterize this bottleneck's internal route.

## 3 Setup

A GUI-grounding VLM maps a screenshot and a natural-language instruction to a single click coordinate $(x, y) \in [0, 1]^2$. The model emits this coordinate autoregressively as a string of digit tokens for an integer point in $\{0, \ldots, 999\}^2$ (e.g. `(190, 104)` for $(x, y) = (0.190, 0.104)$), which we divide by 1000 for evaluation.

**Models.** We study three open GUI-grounding VLMs from the Qwen-VL family: Qwen2.5-VL-7B and -3B (Bai et al., 2025), and OS-Atlas-7B (Wu et al., 2024) (built on Qwen2-VL). Mechanistic analysis centers on Qwen2.5-VL-7B and is replicated on the other two to separate scale from backbone generation. The cross-model decodability check of §4.1 additionally evaluates two further open GUI-grounding VLMs, UI-TARS-1.5 and Holo1.5.

**Data.** The primary evaluation set is 1,000 desktop GUI grounding tasks sampled from AgentNet (Wang et al., 2025) by water-filling across its four mega-domains under a fixed seed; sampling details and per-model usable-trace counts are in Appendix C. We also evaluate on ScreenSpot-Pro (Li et al., 2025), a professional high-resolution grounding benchmark, as a second distribution for the decodability gap (§4.1).

**Click accuracy.** A click is a *hit* when its normalized Euclidean distance to the ground-truth coordinate is at most $\tau = 3/24 = 0.125$—three cells of the $24 \times 24$ patch grid the vision encoder lays over a square crop, slightly tighter than AndroidInTheWild's 14% tap criterion (Rawles et al., 2023). The *hit rate* is the fraction of hits; *generation error* is the distance from the model's generated coordinate to the ground truth, and we call the examples the model misses its *failures*.

## 4 The Coordinate Is Decodable but Not Causal

A failed click admits a simple explanation: the model never located the target, so the coordinate is absent from its hidden state. The model's failure subset, however, remains decodable by a linear probe, and the paired gap between generation error and probe error is positive across most model×benchmark conditions. Yet steering or erasing along the probe direction leaves the click near the no-intervention baseline of every model we test. The probe reads a real signal, but this probe direction is not a causal handle under these interventions.

All three measurements share one instrument. At layer $\ell$ a linear probe reads a coordinate from a hidden state $h \in \mathbb{R}^d$ as $p(h) = Wh + b$ with $W \in \mathbb{R}^{2\times d}$. We fit $W$ by 5-fold ridge regression ($\lambda = 1$) on hidden states averaged over the coordinate-digit token positions, and report out-of-fold predictions. Unless noted, we use the layer maximizing out-of-fold $R^2$ on $x$ and $y$. The probe's output lies in the normalized $[0, 1]^2$ coordinate space; we write *probe error* for its Euclidean distance to the ground truth and call a coordinate *decodable* at layer $\ell$ when probe error is at most the click-hit threshold $\tau{=}3/24$. *Failure-conditional* probing restricts these metrics to the model's failures.

### 4.1 Generation-vs-probe gap

The probe-decoded coordinate usually lies closer to the target than the generated one. To check that the gap is not specific to the three Qwen-VL-family models of §3, we additionally evaluate UI-TARS-1.5 (Qin et al., 2025) and Holo1.5 (H Company,

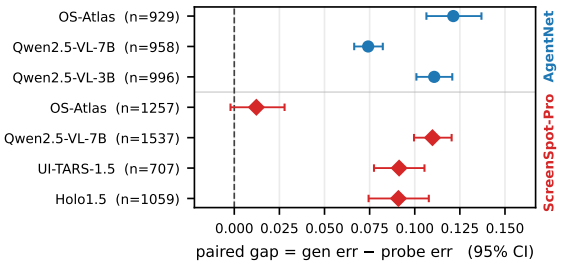


Figure 2: **Paired gap across model × benchmark conditions.** Gap is generation error minus probe error; positive values mean the probe recovers a closer coordinate than the model emits. Markers show bootstrap mean and 95% CI ($B$=10,000); circles are AgentNet, diamonds are ScreenSpot-Pro.

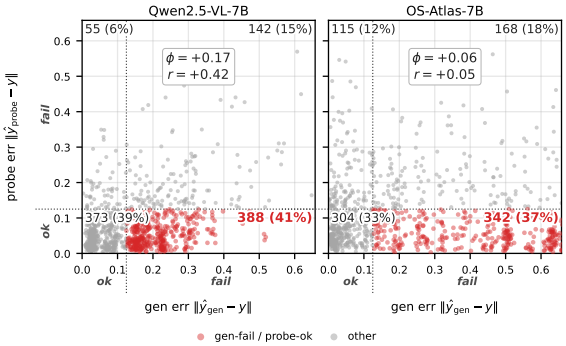


Figure 3: **Per-sample generation versus probe error at the best probe layer.** Dotted crosshair $\tau$=3/24 (click-hit threshold) splits each axis into ok / fail (labelled outside the axes); each cell is annotated with its count and % of parse-ok samples. Red markers are the gen-fail / probe-ok cell. Insets: $\phi$ and Pearson $r$.

2025) on ScreenSpot-Pro (a 200-task AgentNet pilot for both is reported in Appendix D), and compare the paired error gap gen err − probe err across seven conditions: three AgentNet models with the 1k collection and four ScreenSpot-Pro models (Figure 2). The bootstrap 95% CI excludes zero in six of the seven, with the largest AgentNet gap reaching gen = 0.235 vs. probe = 0.114 on OS-Atlas (layer 19) and gen = 0.160 vs. probe = 0.086 on Qwen2.5-VL-7B (layer 14). Only OS-Atlas on ScreenSpot-Pro remains near zero. Thus, a wrong output is not evidence of a missing representation.

### 4.2 Failure-conditional decodability

The coordinate survives the model's own failures. On the 550 of 978 AgentNet examples that Qwen2.5-VL-7B misclicks, a probe on its layer-14 (of 28) hidden state recovers a coordinate within $\tau$ of the target on 72.7% of them; Qwen2.5-VL-3B

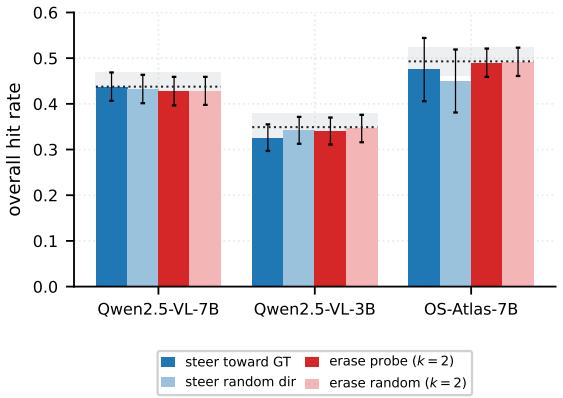


Figure 4: **Probe-direction steering and erasure.** Four interventions at each model's probe-best layer (L14 / L21 / L19 for 7B / 3B / OS-Atlas): additive steering at $\alpha$=10 *toward GT* vs. an equal-norm *random dir*; amnesic projection of the probe's rank-2 row space (*erase probe* $k$=2) vs. a size-matched random rank-2 (*erase random* $k$=2). Dotted line + band: per-model no-intervention generation baseline, 95% bootstrap CI. Bar error bars: Wald 95% ($n$=978 on 7B, 1000 on 3B and OS-Atlas).

and OS-Atlas-7B reach 72.2% and 64.5% at their own probe-best layers (layer 21 of 36 and layer 19 of 28; per-model counts in Table 2). A per-fold mean-target baseline reaches at most 1.8%, and a randomized-center crop control—which decouples the target's position from any image-independent location prior—leaves decodability within a few points of the centered run on every model, arguing against a simple image-independent location-prior explanation (Appendix B; Table 2). Per-sample, generation success and probe success are at most modestly correlated—$\phi \leq 0.17$ and Pearson $r \leq 0.42$ across the three models (Figure 3 insets; Qwen2.5-VL-3B replication in Appendix B)—so the gen-fail / probe-ok cell of the (gen err, probe err) plane is the largest off-diagonal cell.

### 4.3 Probe-direction steering and erasure

The coordinate the probe reads off a hidden state is present in failures but does not, by that alone, identify the direction the model uses to write the click. To test whether it does, we edit the hidden state along the probe's own direction in two complementary ways—additive steering and amnesic erasure—and ask whether either moves the output. Neither does.

We edit a hidden state $h$ so that the readout $p(h)$ equals a chosen coordinate, scaled by a strength

$\alpha$ (minimum-norm in the probe row space; Appendix E): at $\alpha$=1 the readout equals it exactly. We steer toward the ground truth and, as a matched control, along a random direction of equal norm.

Forcing the probe readout to the correct coordinate does not repair the click. We intervene at each model's probe-best layer of §4.2—layer 14 for Qwen2.5-VL-7B, 21 for Qwen2.5-VL-3B, 19 for OS-Atlas-7B—so that the perturbation acts on the same hidden state the probe reads. At the top of an $\alpha$ sweep (Appendix E), steering toward the ground-truth coordinate, along a random direction of equal norm, and toward a deliberately wrong coordinate all leave the hit rate within 5 points of each model's no-intervention baseline on every model (Figure 4; per-cell values in Table 3). On Qwen2.5-VL-7B (L14), for instance, the four $\alpha$=10 cells span 0.426–0.438 against a 0.430 baseline. This pattern is consistent with magnitude-driven disruption rather than direction-specific repair.

Erasing the probe direction does not measurably change the click rate either. We project the probe's rank-2 row space out of the hidden state during generation—an amnesic intervention in the sense of Elazar et al. (2021) (removing the linearly-readable signal)—and on every model the resulting hit rate matches a size-matched random rank-2 projection and falls inside the no-intervention baseline's bootstrap interval (Figure 4; Table 3). On Qwen2.5-VL-7B (L14), probe-amnesic and random rank-2 both land at 0.428 against the 0.430 baseline.

## 5 A Causal Subspace, Wider Than the Probe

The probe direction is decoded but not driven. At each model's probe-best layer we find a $k$-dimensional DAS subspace that does drive the click, and the $(x, y)$ value the rank-2 probe reads is recoverable inside it as a strict sub-signal.

### 5.1 Causal subspace via DAS

An *interchange* (Geiger et al., 2022) replaces a $k$-dimensional component of a *recipient* hidden state with a *donor*'s, $h' = h_{\text{rec}} + PP^{\top}(h_{\text{don}} - h_{\text{rec}})$ for an orthonormal basis $P \in \mathbb{R}^{d \times k}$; if $P$ carries the coordinate causally, the patched recipient emits the donor's coordinate. We score it by the *donor-match rate*—the fraction of (recipient, donor) example pairs whose patched output lands within $\tau$ of the donor coordinate—and report alongside the

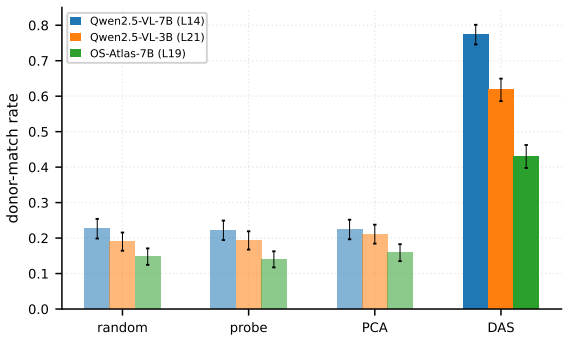


Figure 5: **Interchange success across subspace conditions.** Donor-match rate for four subspace types at the interchange layer ($k$=32 except the probe's rank-2 row space). Error bars are Wald 95%.

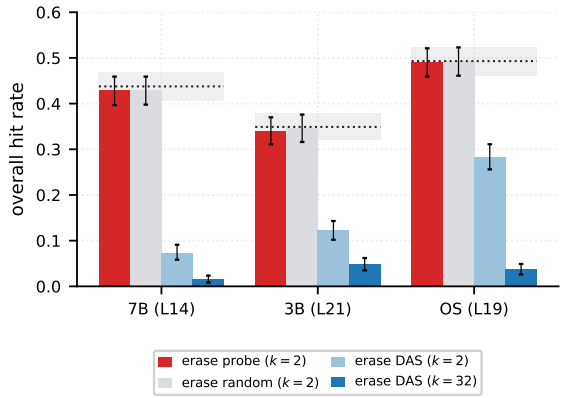


Figure 6: **Subspace erasure: probe vs. DAS at matched rank and at $k$=32.** Hit rate after amnesic projection of the legend subspace at each model's probe-best layer (L14 for 7B, L21 for 3B, L19 for OS-Atlas); *erase probe* ($k$=2) repeats the rank-2 null of Figure 4. Dotted line + band: per-model no-intervention generation baseline, 95% bootstrap CI. Bar error bars: Wald 95% ($n$=978 on 7B, 1000 on 3B and OS-Atlas).

*no-patch reference* (*chance*): the donor-match rate with the recipient left unpatched, a per-pair baseline set by the coordinate distribution and $\tau$ rather than by the intervention.

DAS (Geiger et al., 2024) learns $P$ by gradient descent with the model frozen, and we evaluate at test time by free generation. Same-layer controls—a random orthonormal subspace, the probe's rank-2 row space, and the top-$k$ PCA subspace—separate direction from magnitude. We intervene at the residual stream of a single layer (Appendix F).

DAS identifies a subspace whose interchange moves the output toward the donor coordinate. At each model's probe-best layer (§4.2), the $k$=32 DAS subspace lifts paired interchange success by +0.548 on Qwen2.5-VL-7B (L14), +0.428 on

Qwen2.5-VL-3B (L21), and +0.282 on OS-Atlas-7B (L19) over a same-layer random-orthonormal subspace (paired bootstrap 95% CI excludes zero on every model). In a coarse layer sweep, the same layer that maximizes probe out-of-fold $R^2$ also maximizes DAS donor-match on each model (Appendix F); the same pattern appears across the backbone generation, training objective, and output-format change introduced by OS-Atlas.

The lift is specific to the direction DAS selects. A same-layer random orthonormal subspace, the probe's rank-2 row space, and a top-$k$ PCA subspace all leave the transfer at each model's no-patch reference on every model (Figure 5; per-cell numbers in Table 4). Under this interchange test, the probe row space shows no measurable donor-coordinate transfer. A shuffled-label DAS control (Appendix F) rules out an optimization-procedure artifact.

The same direction-specificity appears under erasure (Figure 6). Amnesic projection of the $k$=32 DAS subspace at the probe-best layer collapses the hit rate to near-zero on every model (e.g., 0.438 → 0.015 on Qwen2.5-VL-7B L14; full numbers in Table 5), while a size-matched random $k$=32 projection at the same layer leaves each baseline unmoved. This comparison suggests that the probe-direction null of §4.3 is specific to the probe direction rather than to amnesic projection itself: applied to the DAS-learned subspace, the same intervention drives the click toward chance on every model.

The gap is not explained by rank alone. A rank-matched DAS subspace ($k$=2) already exceeds the probe in both directions of intervention on every model: $k$=2 DAS donor-match outpaces probe rank-2 under interchange, and $k$=2 DAS erasure drops the hit rate while probe-rank-2 erasure stays at baseline—a 5.8×, 2.8×, and 1.7× erasure asymmetry on Qwen2.5-VL-7B (L14), -3B (L21), and OS-Atlas-7B (L19) at the same rank under the same intervention (Figure 6; Tables 4, 5). These results support a wider-than-rank-2 handle (donor-match plateau onset $k \approx 8$, peak $k$=32; Appendix G). The rank-2 contrast is sharpest on 7B—the model with the steepest rank curve—and $k$=32 is the smallest $k$ at which DAS-subspace erasure brings the hit rate to chance on every model; we fix $k$=32 across erasure and routing for this reason rather than because it maximizes donor-match.

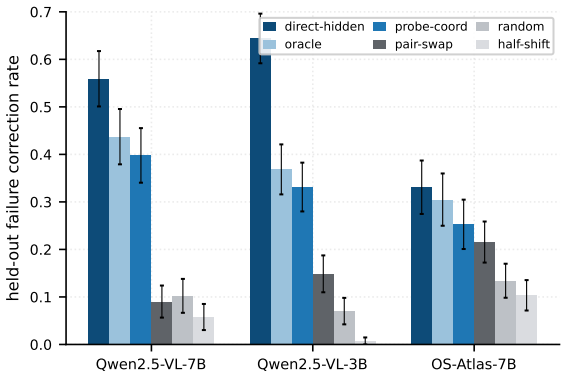


Figure 7: **Held-out failure correction by injection condition** at each model's probe-best layer (L14 / L21 / L19). *wide-routed* maps the readout-layer $k$=32 DAS projection into the probe-best-layer DAS subspace through a learned ridge map; *probe-coord* routes only the rank-2 probe readout through the analogous map; *rank-2 oracle* injects the ground-truth coordinate through the same rank-2 map. *pair-swap*, *random*, and *half-shift* are corrupted-coordinate controls. Error bars: Wald 95% over held-out failures.

### 5.2 Heads that write the probe break the readout, not the click

The rank-2 dissociation extends to the heads that build the probe-readable representation, and the readout-vs-click asymmetry reproduces on all three models. Per-head attribution patching against a probe-direction loss (Kramár et al., 2024) returns a five-head *probe-writing* cluster at each model's *probe-write layer*—the layer where the probe is re-fit, a few layers downstream of the probe-best layer (L18 on 7B, L24 on 3B, L21 on OS-Atlas; Appendix I): on Qwen2.5-VL-7B at L13–18 (L18H7, L18H9, L18H8, L14H14, L13H25); on Qwen2.5-VL-3B at L21–24 (L21H12, L24H15, L23H10, L21H9, L24H2); on OS-Atlas-7B at L18–21 (L19H20, L21H6, L19H17, L18H7, L19H15). On Qwen2.5-VL-7B the L18 cluster sits within a small band of *localizer* heads in L18–20 that attend sharply to the target patch, consistent with Kang et al. (2025).

We zero each head's contribution to the residual stream with an `o_proj` forward hook that sets the corresponding head slice to zero at all token positions, then measure the probe error at the probe-write layer and the click hit rate (Table 1). The readout breaks on every model: probe-writing top-5 ablation raises probe error by +63% on Qwen2.5-VL-7B (0.068 → 0.111), +515% on 3B (0.037 → 0.226), and +85% on OS-Atlas

| ablated heads | Qwen-7B (L18) | | Qwen-3B (L24) | | OS-Atlas (L21) | |
|---|---|---|---|---|---|---|
| | hit | err.↑ | hit | err.↑ | hit | err.↑ |
| none (baseline) | 0.428 | 0.068 | 0.346 | 0.037 | 0.493 | 0.102 |
| probe-writing top-1 | 0.423 | 0.080 | 0.343 | 0.093 | 0.474 | 0.131 |
| probe-writing top-3 | 0.489 | 0.088 | 0.344 | 0.154 | 0.484 | 0.166 |
| probe-writing top-5 | 0.486 | **0.111** | 0.318 | **0.226** | 0.409 | **0.189** |
| digit-logit top-5 | 0.410 | 0.078 | 0.345 | 0.358 | 0.490 | 0.178 |
| random-5 | 0.430 | 0.074 | 0.345 | 0.061 | 0.488 | 0.105 |

Table 1: **Cumulative head ablation across three models: probe error and click hit rate** ($n$=1000 per model, evaluated at each model's probe-write layer). The arrow marks the direction of readout breakage. *random-5* is a seeded random-head control; *digit-logit top-5* is the head set ranked by digit-logit attribution.

($0.102 \rightarrow 0.189$), while a seeded random-5 control stays within $\pm 0.025$ of baseline on each model. Under the same ablation the click hit rate moves by $+0.058$ on 7B, $-0.028$ on 3B, and $-0.084$ on OS-Atlas (Wald 95% half-width $\approx 0.031$ at $n$=1000); the 7B and 3B shifts are within noise of baseline, the OS-Atlas drop is the largest single shift, but the readout-vs-click asymmetry holds in every case (probe error multiplies by 1.6–6.1×, against at most a 17% relative drop in the click). The pattern is consistent with a distributed click pathway in which the localizer/probe-writing heads' contribution is partially recoverable downstream—qualitatively opposite to the DAS-subspace erasure at each model's probe-best layer (§5.1), which collapses the hit rate to near-zero on every model. A second control that ablates the five top heads by *digit-logit* attribution acts the same as random-5 on the click on every model; on 3B and OS-Atlas the digit-logit set overlaps with the probe-writing one (3/5 and 2/5 heads shared; Appendix I), so its probe-error rise on those models is not an independent perturbation.

### 5.3 Probe-to-handle routing

A learned map can route readout-layer information into the causal handle, and routing a wider slice of the hidden state—rather than only a rank-2 signal (the probe readout or the ground-truth coordinate)—corrects more failures on every model. All three injection conditions write into the same $k$=32 DAS subspace at the probe-best layer and differ only in what is mapped in (Appendix H). *rank-2 oracle* sends the ground-truth $(x, y)$ through a ridge $g : \mathbb{R}^2 \rightarrow \mathbb{R}^K$ fit on the training split; *probe-coord* substitutes the probe's $(x, y)$ readout into the same $g$; *wide-routed* bypasses the 2-d coordinate and uses a separate ridge $g' : \mathbb{R}^K \rightarrow \mathbb{R}^K$ from the *readout-layer* DAS projection—the source layer of the routing map, layer 25 for the Qwen models and 18 for OS-Atlas—to the probe-best-layer DAS-projection target. On held-out failures, wide-routed has the highest correction rate on every model, exceeding both probe-coord and rank-2 oracle (Figure 7; per-cell numbers in Table 6). The rank-2 oracle is an upper bound only conditional on the rank-2 coordinate being the available signal—its injected target lies in $g(\mathbb{R}^2)$, a rank-2 slice of the $K$-dim subspace—so wide-routed's edge over it isolates click-relevant information that the coordinate alone does not carry but the $k$=32 readout-layer DAS projection does ($k$-sweep plateau at $k \approx 8$; $k$=32 held fixed with §5.1; Appendix G).

The corrupted-coordinate controls argue against a generic injection-mechanism artifact. Random, half-shift, and pair-swap (Figure 7) correct at most 0.22 of held-out failures (OS-Atlas pair-swap; lower on the Qwen models; Table 6). Both maps' paired bootstrap 95% intervals exclude zero against every control on the Qwen models; on OS-Atlas, wide-routed still exceeds the strongest control by $+0.11$, while probe-coord is within range of pair-swap.

The rank-2 probe readout is a partial bottleneck for the handle. Wide-routed corrects a wider failure subset than probe-coord on every model, and the overall hit rate moves accordingly (Table 6). The bottleneck is sharpest on Qwen2.5-VL-3B at L21, where the wider $k$=32 injection nearly doubles the corrected fraction (probe-coord $0.33 \rightarrow$ wide-routed $0.64$); the wide-routed $>$ probe-coord ordering is preserved on every model. The same wide-routed intervention preserves a majority of baseline-OK clicks on every model (0.54–0.63; Table 6), so the correction does not simply trade the OK set for a disjoint set of repaired failures. Together, these results suggest that coordinate-relevant information in the readout-layer DAS projection can drive the click when routed through the causal subspace, while the rank-2 probe readout discards information useful for that routing.

## 6 Discussion and Conclusion

A GUI-grounding VLM that misses a click has often not failed to perceive the target. Across Qwen2.5-VL and OS-Atlas, the correct coordinate is linearly present in the hidden states of a large share of failures (§4). What fails is internal: the coordinate appears in three representations that do

not coincide. A rank-2 readout direction is recovered by a linear probe (§4). A distributed mid-layer causal subspace ($k$=32, with transfer saturating by $k \approx 8$) drives the click under interchange but is not aligned with the probe direction (§5). A late-layer emission pathway writes the digits without bottlenecking on the probe-writing heads—ablating them breaks the readout (probe error $\times$1.6–6.1) but moves the click by at most a $17\%$ relative drop on every model (§5.2), and a per-head search on Qwen2.5-VL-7B isolates no concentrated driver (Appendix J). Many clicks land wrong because these representations are misaligned, not because the target is unseen.

The routing account changes what a repair should target. The standard remedy—more grounding data, sharper localization—targets perception, but the click is emitted from a late-layer pathway whose probe-writing heads do not concentrate the click signal on any of the three models (§5.2), and a per-head search on Qwen2.5-VL-7B isolates no concentrated driver that a perceptual remedy could target (Appendix J). The coordinate is known; it does not reach the output. The leverage point is instead the path from the readout to the emission, and that path is realizable: routing the $k$=32 DAS-projection of the late-layer hidden state into the causal subspace corrects $56\%$ of held-out Qwen2.5-VL-7B failures and $64\%$ on 3B, while routing only the rank-2 probe readout reaches $40\%$ and $33\%$ respectively, so the probe itself is a partial bottleneck within the handle (§5.3). Wide-routed lifts the overall hit rate on both Qwen models (net $+0.16$ on 7B, $+0.28$ on 3B) and leaves it near-neutral on OS-Atlas ($-0.01$), but it still requires the full $k$=32 DAS-projection of the hidden state at test time; a deployable correction would identify the minimal subspace that drives this lift. What information beyond the $(x, y)$ coordinate the readout-layer $k$=32 projection carries—and which the rank-2 probe discards—is itself an open mechanistic question. The natural modelling targets are training objectives that align the probe-readable readout with the causal subspace, or inference-time routers that bridge the two as our correction does.

For interpretability, the result is a cautionary case in a practical grounding task. A linear probe that decodes a quantity accurately is routinely read as evidence that the model represents and uses that quantity. Here the probe decodes the click coordinate from a model's own failures, yet the direction it reads is causally inert: steering along it moves the output no more than a random direction does. The contrast is sharpest at matched rank in the same hidden state: a rank-2 direction the model uses (DAS) and a rank-2 direction that reads the coordinate best (the probe) are not the same direction. The two notions can nonetheless be separated by training objective: in this task, a causal objective (DAS) recovers the direction the model uses, while a correlational objective (the probe) does not. Probes remain useful as detectors of what information is present, but do not by themselves identify what the model acts on.

**Conclusion.** For a large share of the failures we examined, GUI-grounding click failure is misrouted representation, not absent perception. The readout a probe decodes, the subspace that drives the output, and the pathway that emits the digits do not coincide. This diagnostic picture holds across model scale and generation, and the routing correction recovers more failures than the rank-2 probe readout on every model, though it lifts the overall hit rate net-positive only on the two Qwen2.5-VL checkpoints. The narrower claim our intervention establishes is that the gap between knowing and emitting is bridgeable. Closing it through a more localized handle than the $k \approx 32$ subspace, or through structured multi-token coordinate decoding, is a concrete target for future work on GUI grounding.

## Limitations

**Model family and data.** The mechanistic analysis covers the Qwen-VL family—Qwen2.5-VL at 7B and 3B, and OS-Atlas, built on Qwen2-VL—and, for everything beyond the gap forest of Figure 2, a single corpus. The decodability gap is checked on ScreenSpot-Pro and on UI-TARS-1.5 and Holo1.5, but the failure-conditional probe (§4.2), the steering and erasure null (§4.3), the DAS subspace (§5.1), the head ablation (§5.2), and the routing correction (§5.3) are all evaluated on AgentNet only. Whether the readout–handle–emission split holds on a second distribution—ScreenSpot-Pro in particular, where the gap forest already shows the pattern attenuating on OS-Atlas—or on non-Qwen architectures—InternVL-based models in particular—is open. The central structure does reproduce across two scales and two model generations (§5), so it is not an artifact of a single checkpoint, but we claim it only for Qwen-VL-family GUI-grounding VLMs on AgentNet.

**OS-Atlas is the weakest of the three replications.** The readout–handle dissociation and the routing-as-repair story attenuate on OS-Atlas across every measurement. The matched-rank erasure asymmetry is $1.7\times$ versus $5.8\times$ and $2.8\times$ on Qwen2.5-VL-7B and -3B (§5.1); wide-routed leaves the overall hit rate effectively unchanged ($-0.01$ vs. $+0.16$ on 7B and $+0.28$ on 3B) and probe-coord moves it down by $-0.14$, the largest negative shift among the three models (§5.3); and the probe-writing head ablation produces a $-0.084$ click drop, the only one that breaks the readout-vs-click asymmetry by more than a noise margin (§5.2). The qualitative picture—wider-than-rank-2 handle, decoded probe direction that is not driven, distributed emission—still holds on every model, but the routing correction is a positive overall lift only on the two Qwen2.5-VL checkpoints; OS-Atlas's bbox prompt and its different post-training are the obvious candidate confounds, and separating them is left to future work.

**Layer selection.** The probe, the DAS subspace, and the routing target all share the same per-model layer (L14 / L21 / L19), chosen as the layer maximising probe out-of-fold $R^2$. A coarse DAS layer sweep finds the same layer also maximises donor-match (Appendix F), but selecting the layer by one analysis and then running a second analysis at that layer is a methodological dependency; whether a layer chosen by a different criterion would yield a different routing story is open.

**The correction is not a method.** The routing intervention is informative but not a deployable correction. Wide-routed lifts the overall hit rate by $+0.16$ / $+0.28$ / $-0.01$ on 7B / 3B / OS-Atlas (§5.3)—the first cell with a positive net effect on the two Qwen models—but it requires the full $k$=32 DAS-projection of the late-layer hidden state at test time, not a self-contained intervention. A deployable correction would identify the minimal subspace that drives this lift and a per-sample condition under which it is safe to apply.

**The handle-to-emission circuit.** Path patching shows that the causal handle's effect propagates through the residual stream without bottlenecking on any small head set (Appendix J), but it establishes this by freezing attention-head outputs; the residual-stream and MLP routing that carries the effect is not dissected, and the corroborating per-head emission search (Appendix J) is run on Qwen2.5-VL-7B only. The complete circuit from the mid-layer handle through the digit-by-digit emission is not mapped, and we leave it to future work.

## Ethics Statement

This work analyzes the internal representations of existing, publicly released vision–language models. It introduces no new model, training method, or capability, and it does not increase the autonomy of GUI agents; the contribution is diagnostic. The screenshots and instructions used are of standard desktop applications performing routine document and browser tasks, and were used solely for analysis. GUI-grounding models carry the general dual-use considerations of software automation, and a clearer mechanistic account of when their clicks fail can inform both more reliable agents and their safer evaluation. All models are used in accordance with their respective licenses.

**Use of AI assistants.** AI coding assistants were used for parts of the analysis-code implementation and for parts of the manuscript drafting. All code was reviewed and tested by the authors, and all text was reviewed and revised by the authors.

**Code and artifacts.** Upon acceptance, we will release the analysis pipeline (probe fitting, DAS training, interchange/erasure interventions, routing maps, and head-level ablations), together with the trained ridge-probe weights, DAS orthonormal bases $P$, and routing maps $g$ / $g'$ for all three analyzed models, plus the AgentNet sample-id list defining the $1{,}000$-task evaluation split. Code is released under the MIT license; weights and intermediates derived from Apache-2.0 checkpoints (Qwen2.5-VL-7B-Instruct, OS-Atlas-Base-7B) are released under MIT. Artifacts derived from Qwen2.5-VL-3B-Instruct inherit the Qwen Research License (non-commercial research use only) and are released under the same terms.

## References


Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Comput. Linguistics*, 48(1):207–219.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023a. Eliciting latent predictions from transformers with the tuned lens. *CoRR*, abs/2303.08112.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023b. LEACE: perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda B. Viégas, and Martin Wattenberg. 2021. An interpretability illusion for BERT. *CoRR*, abs/2104.07143.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. 2024. Seeclick: Harnessing GUI grounding for advanced visual GUI agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 9313–9332. Association for Computational Linguistics.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160â€“175.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah D. Goodman, and Christopher Potts. 2022. Inducing causal structure for interpretable neural networks. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, Proceedings of Machine Learning Research, pages 7324–7338. PMLR.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. 2024. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning, 1-3 April 2024, Los Angeles, California, USA*, Proceedings of Machine Learning Research, pages 160–187. PMLR.

H Company. 2025. Holo1.5-7B. `https://huggingface.co/Hcompany/Holo1.5-7B`. Hugging Face model card.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. Your large vision-language model only needs A few attention heads for visual grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 9339–9350. Computer Vision Foundation / IEEE.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. 2024. Atp*: An efficient and scalable method for localizing LLM behaviour to components. *CoRR*, abs/2403.00745.

Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. 2025. Screenspot-pro: GUI grounding for professional high-resolution computer use. In *Proceedings of the 33rd ACM International Conference on Multimedia, MM 2025, Dublin, Ireland, October 27-31, 2025*, pages 8778–8786. ACM.

Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Jiaping Lin, Fei Shen, Junzhe Li, Ping Nie, Fei Yu, Ming Li, and Haizhou Li. 2026. What happens before decoding? Prefill determines GUI grounding in VLMs. *CoRR*, abs/2605.12549.

Alex Mallen and Nora Belrose. 2023. Eliciting latent knowledge from quirky language models. *CoRR*, abs/2312.01037.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Dang Nguyen, Jian Chen, Yu Wang, Gang Wu, Namyong Park, Zhengmian Hu, Hanjia Lyu, Junda Wu, Ryan Aponte, Yu Xia, Xintong Li, Jing Shi, Hongjie Chen, Viet Dac Lai, Zhouhang Xie, Sungchul Kim, Ruiyi Zhang, Tong Yu, Md. Mehrab Tanjim, and 11 others. 2025. GUI agents: A survey. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 22522–22538. Association for Computational Linguistics.

Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. 2023. Towards vision-language mechanistic interpretability: A causal tracing tool for BLIP. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*, pages 2848–2853. IEEE.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Forty-first International Conference on Machine Learning, ICML*

*2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 39643–39666. PMLR / OpenReview.net.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, Wanjun Zhong, Kuanye Li, Jiale Yang, Yu Miao, Woyu Lin, Longxiang Liu, Xu Jiang, Qianli Ma, Jingyu Li, and 16 others. 2025. UI-TARS: pioneering automated GUI interaction with native agents. *CoRR*, abs/2501.12326.

Christopher Rawles, Alice Li, Daniel Rodriguez, Oriana Riva, and Timothy P. Lillicrap. 2023. Androidinthewild: A large-scale dataset for android device control. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15504–15522. Association for Computational Linguistics.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *CoRR*, abs/2308.10248.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. 2020. Causal mediation analysis for interpreting neural NLP: the case of gender bias. *CoRR*, abs/2004.12265.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191.

Xinyuan Wang, Bowen Wang, Dunjie Lu, Junlin Yang, Tianbao Xie, Junli Wang, Jiaqi Deng, Xiaole Guo, Yiheng Xu, Chen Henry Wu, Zhennan Shen, Zhuokai Li, Ryan Li, Xiaochuan Li, Junda Chen, Boyuan Zheng, Peihang Li, Fangyu Lei, Ruisheng Cao, and 23 others. 2025. Opencua: Open foundations for computer-use agents. *CoRR*, abs/2508.09123.

Qianhui Wu, Kanzhi Cheng, Rui Yang, Chaoyun Zhang, Jianwei Yang, Huiqiang Jiang, Jian Mu, Baolin Peng, Bo Qiao, Reuben Tan, Si Qin, Lars Liden, Qingwei Lin, Huan Zhang, Tong Zhang, Jianbing Zhang, Dongmei Zhang, and Jianfeng Gao. 2025. Gui-actor: Coordinate-free visual grounding for GUI agents. *CoRR*, abs/2506.03143.

Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, and Yu Qiao. 2024. OS-ATLAS: A foundation action model for generalist GUI agents. *CoRR*, abs/2410.23218.

Andy Zou, Long Phan, Sarah Li Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to AI transparency. *CoRR*, abs/2310.01405.

## A Models and Checkpoints

All experiments load HuggingFace checkpoints with `dtype bfloat16` and sdpa attention (PyTorch 2.4.0, transformers 4.57.x). Mechanistic analysis centers on Qwen2.5-VL-7B-Instruct (`Qwen/Qwen2.5-VL-7B-Instruct`, 28 decoder layers, hidden width $d = 3584$, 28 attention heads of dimension 128); the snapshot revision is `cc594898137f460bfe9f0759e9844b3ce807cfb5`. The central causal results are replicated on Qwen2.5-VL-3B-Instruct (`Qwen/Qwen2.5-VL-3B-Instruct`, 36 decoder layers, hidden width 2048) and on OS-Atlas-Base-7B (`OS-Copilot/OS-Atlas-Base-7B`, built on Qwen2-VL). The decodability gap is additionally measured on UI-TARS-1.5 (`ByteDance-Seed/UI-TARS-1.5-7B`) and Holo1.5 (`Hcompany/Holo1.5-7B`).

**Licenses.** All checkpoints are loaded under their published licenses: Apache 2.0 for Qwen2.5-VL-7B-Instruct, OS-Atlas-Base-7B, UI-TARS-1.5-7B, and Holo1.5-7B; and the Qwen Research License Agreement (non-commercial research use only) for Qwen2.5-VL-3B-Instruct. All uses in this work are non-commercial research analysis, consistent with each license.

Inputs are square-cropped around the target (click-aware crop) and resized so the vision encoder lays a $24 \times 24$ grid over the crop with an effective patch size of 28 px (`min_pixels` = `max_pixels` = $(24 \times 28)^2$). Greedy decoding (`do_sample=False`) with at most 64 new tokens is used throughout.

**Prompts.** On AgentNet, the Qwen-family models receive a point-format prompt:

```
Look at this UI screenshot. Find the
element described below and output
ONLY the click coordinates as a single
point (x, y), where x and y are
```

```
integers from 0 to 1000 normalised to
the image. Do not include any other
text. Example output: (354, 678)

Target element: {action}
```

where {action} is the natural-language instruction from the AgentNet trajectory's first action; the emitted 0–1000 integers are divided by 1,000 to give a coordinate in $[0,1]^2$. OS-Atlas on AgentNet instead receives a bounding-box prompt, "In this UI screenshot, what is the position of the element corresponding to the command "{action}" (with bbox)?", and the four emitted integers $(x_1, y_1, x_2, y_2)$ are reduced to a single click by box center, $((x_1{+}x_2)/2{,}000,\ (y_1{+}y_2)/2{,}000)$. On ScreenSpot-Pro, the four conditions use: the OS-Atlas bbox prompt unchanged (parsed by box center); a Qwen2.5-VL bbox variant that adds the line `Example: (354,678),(374,698)` after the format spec, also parsed by box center; a UI-TARS-1.5 GUI-agent action-space prompt of the form `click(point='<point>x1 y1</point>')`, with the two 0–1000 integers divided by 1,000; and a Holo1.5 JSON-schema prompt requesting an absolute-pixel click, with its emitted x, y integers divided by the 672-pixel resized side (the $24{\times}28$ effective grid).

## B Failure-Conditional Probe

The probe $p(h) = Wh + b$ with $W \in \mathbb{R}^{2\times d}$ is ridge regression with $\alpha = 1$. The hidden state $h$ is taken at every digit-emission generation step (the last query position) and averaged over those steps, per layer. Layer indices are 0-based: layer $\ell$ denotes the residual after the $(\ell{+}1)$-th transformer block, so layer 14 of Qwen2.5-VL-7B is the output of its 15th block. Out-of-fold predictions use 5-fold cross-validation with `KFold(shuffle=True, random_state=42)`; predictions are clipped to $[0,1]^2$. A coordinate is decodable when the probe error is within $\tau = 3/24$. The probe layer reported is the one maximizing overall out-of-fold $R^2$. Failure-conditional metrics restrict these quantities to the model's own failures. The Qwen2.5-VL-7B run uses 978 usable AgentNet samples; the Qwen2.5-VL-3B and OS-Atlas runs use 1,000.

**Crop randomization (D1 control).** The click-aware crop centers a square window on the target before resizing. To check that decodability does not depend on this centering, a randomized-crop variant places the target at a position drawn uniformly inside the window (subject to a two-patch margin from any edge, with a centered fallback when the margin cannot be satisfied; the random offset is seeded by `task_id` for reproducibility). The two-patch margin is the one used at data-collection time and we reclassify hits at the current $\tau{=}3/24$ on the same samples; a larger margin would shrink the random-offset set further but does not affect the direction of the comparison. Per-model rand-crop decodability sits within a few points of the centered values reported in §4 (Table 2, *rand-crop* column). The effective input change is narrow—only 189 of 1,000 targets sit far enough from an edge to actually move—but it removes the centering as a confound.

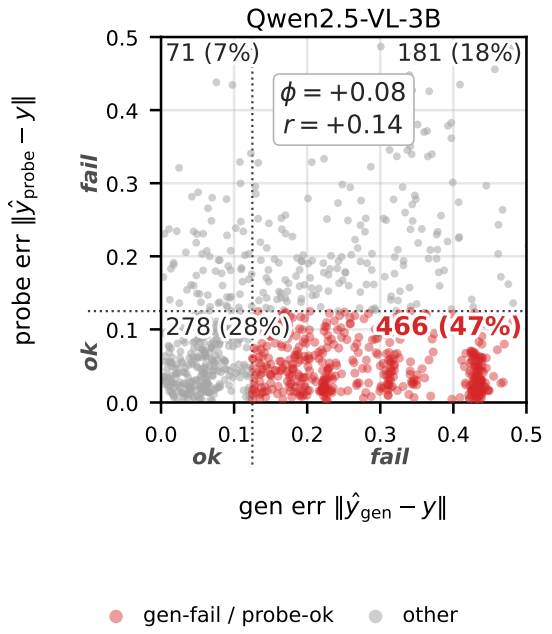


Figure 8: **Per-sample generation versus probe error at the best probe layer.** Companion panel to Figure 3; same axes, crosshair, and annotation conventions.

**Marginal coordinate baseline (D2 control).** A coordinate-only predictor with no access to hidden states predicts the per-fold training-mean coordinate (same 5-fold split as the probe) and is scored under the same $\tau$. Centered-crop hit rates appear in Table 2 (*mean-tgt* column, 0.016–0.018 across models); the randomized crop raises this baseline to 0.051–0.062. The marginal distribution of targets is too diffuse to explain the failure-conditional decodability the probe reports.

| Model | layer (of $L$) | $n_{\text{fail}}/n$ | fail $\leq \tau$ | rand-crop fail $\leq \tau$ | mean-tgt | $\phi$ / $r$ |
|---|---|---|---|---|---|---|
| Qwen2.5-VL-7B | 14 (of 28) | 550 / 978 | 0.727 | 0.720 | 0.016 | 0.171 / 0.415 |
| Qwen2.5-VL-3B | 21 (of 36) | 651 / 1000 | 0.722 | 0.694 | 0.018 | 0.084 / 0.143 |
| OS-Atlas-7B | 19 (of 28) | 581 / 1000 | 0.645 | 0.582 | 0.018 | 0.059 / 0.053 |

Table 2: **Failure-conditional decodability at the probe-best layer.** *fail* $\leq \tau$ is the fraction of the model's own click-failures (gen err $> \tau$) where the probe error is $\leq \tau$=3/24 (centered crop). *rand-crop fail* $\leq \tau$ is the same metric under the randomized-center crop. *mean-tgt* is the per-fold mean-target predictor on the full $n$ (centered crop). $\phi$ is the binary ok/fail correlation; $r$ is the Pearson correlation on continuous probe-vs-gen error. Numerical companion to Figures 2, 3.

## C AgentNet Sampling and Usable Traces

The 1,000 desktop GUI tasks from AgentNet are drawn under a fixed seed by water-filling across the four mega-domains—allocating quota equally up to per-domain availability—yielding 262 Daily, 262 System, 262 Work, and the 214 available Professional tasks. The tasks cover Ubuntu applications such as LibreOffice and Chrome. The 1,000-task set is a strict superset of the 200-task pilot (50 per mega-domain) of Appendix D. The hidden-state collection yields usable digit-emission traces for 978 Qwen2.5-VL-7B examples—the missing 22 are dropped because the 7B model produced coordinate strings whose digit positions could not be aligned to the hidden-state trace—and for all 1,000 Qwen2.5-VL-3B and OS-Atlas examples; unparseable generated coordinates within these traces are counted as failures.

**Source, language, and PII.** AgentNet (`xlangai/AgentNet`, MIT license) consists of human-annotated desktop trajectories collected via a screen-recording annotation tool on Windows, macOS, and Ubuntu; we sample its Ubuntu subset. ScreenSpot-Pro (`likaixin/ScreenSpot-Pro`, MIT license) consists of high-resolution professional-software screenshots. Both datasets are English-language. Neither documents an automated personally-identifiable-information (PII) filter; manual inspection of the 1,000-task AgentNet sample we use revealed no personally identifiable content beyond standard application UI, and the ScreenSpot-Pro screenshots are of generic professional-software interfaces.

## D AgentNet 200-Pilot Conditions

UI-TARS-1.5 and Holo1.5 lack a 1k AgentNet hidden-state collection and are excluded from the main-text Figure 2 for that reason; only their 200-task pilot is available. On the pilot we recover, at each model's best probe layer, gen $= 0.146$ vs. probe $= 0.094$ for UI-TARS-1.5 (layer 18, $n_{\text{parse}} = 151$, gap $+0.053$, CI $[+0.035, +0.070]$) and gen $= 0.133$ vs. probe $= 0.137$ for Holo1.5 (layer 21, $n_{\text{parse}} = 155$, gap $-0.004$, CI $[-0.034, +0.028]$). The UI-TARS gap is positive and CI-excludes zero on the pilot; Holo1.5's pilot gap is near-zero on AgentNet, in contrast to its $+0.091$ gap on the harder ScreenSpot-Pro distribution (Figure 2).

## E Probe-Direction Steering

The probe is refit on all samples at each model's probe-best layer of §4.2 (L14 / L21 / L19 for Qwen2.5-VL-7B / -3B / OS-Atlas). During generation a forward hook edits the last-position residual: with $M = W^\top(WW^\top)^{-1}$ mapping a readout delta back into the residual stream, the edit is $\Delta h = \alpha\, M(\,\text{target} - p(h)\,)$, the minimum-norm vector in the probe's rank-2 row space. The strength grid is $\alpha \in \{0.5, 1, 2, 5, 10\}$. Conditions: steer toward the ground-truth coordinate; toward the probe's own out-of-fold prediction; toward a deliberately wrong coordinate (each axis half-shifted, $(x + 0.5) \bmod 1$); and a matched random control along a fixed random unit direction (seed $1000 + i$ for sample $i$) scaled to the per-step norm of the ground-truth edit. The amnesic erasure run additionally evaluates three projection conditions at the steering layer and at the corresponding interchange layer of Appendix F: an *amnesic* projection that removes the probe's rank-2 row space at the chosen layer, an *amnesiccum* projection that removes each layer's probe rank-2 row space from the chosen layer through the final layer, and a size-matched random projection (*rand2*); bootstrap confidence intervals use $B = 10{,}000$ resamples.

## F DAS and Interchange Interventions

DAS learns an orthonormal $P \in \mathbb{R}^{d\times k}$ at each model's probe-best layer (§4.2). $P$ is parameter-

| Model (layer) | baseline | $\alpha$=1 GT | $\alpha$=10 GT | $\alpha$=10 rand | $\alpha$=10 wrong | erase probe | erase rand |
|---|---|---|---|---|---|---|---|
| Qwen2.5-VL-7B (L14) | 0.430 | 0.429 | 0.438 | 0.433 | 0.426 | 0.428 | 0.428 |
| Qwen2.5-VL-3B (L21) | 0.346 | 0.340 | 0.326 | 0.342 | 0.352 | 0.340 | 0.346 |
| OS-Atlas-7B (L19) | 0.493 | 0.495 | 0.475 | 0.450 | 0.482 | 0.490 | 0.492 |

Table 3: **Probe-direction steering and erasure at each model's probe-best layer.** Hit rate after each intervention against the no-intervention baseline. Steering uses the minimum-norm vector in the probe's rank-2 row space ($\alpha \in \{1, 10\}$); *wrong* half-shifts each axis. *erase probe* projects out the probe's rank-2 row space; *erase rand* is a size-matched random rank-2 projection. Numerical companion to Figure 4.

ized as the QR factor of a free matrix initialized $\mathcal{N}(0, 0.02^2)$ and optimized with Adam, learning rate $5 \times 10^{-3}$, for 3 epochs. The training objective is the teacher-forced cross-entropy of the donor's coordinate digits when the interchange $h' = h_{\text{rec}} + PP^\top(h_{\text{don}} - h_{\text{rec}})$ is applied over the donor's coordinate token positions; the model is frozen. The data are split 70/30 train/test with seed 0; DAS trains on the train split and is evaluated by free autoregressive generation on (recipient, donor) pairs drawn from the test split, 3 donors per recipient. An interchange contributes to the donor-match rate when the patched output lands within $\tau$ of the donor coordinate. Subspace dimensions $k \in \{2, 8, 16, 32\}$ are evaluated. Control subspaces at the same layer: a random orthonormal subspace (seed 7); the probe's rank-2 row space (ridge $\alpha = 1$, QR-orthonormalized); and the top-$k$ PCA subspace of the training-split hidden states. A no-patch cell gives the reference.

**Interchange layer selection.** A coarse DAS layer sweep on OS-Atlas over $\{14, 17, 19, 21, 24\}$ places the donor-match peak at L19 (+0.282 over a same-layer random subspace at $k$=32), the same layer that maximises probe out-of-fold $R^2$. Qwen2.5-VL-7B's L14 and Qwen2.5-VL-3B's L21 similarly pair the probe-readout best layer with the DAS donor-match peak in a coarse sweep; the older paper-interchange layers (L19 for 7B, L24 for 3B, L21 for OS-Atlas) yield smaller DAS lifts than the probe-best choice (+0.535 vs +0.548 on 7B, +0.377 vs +0.428 on 3B, +0.202 vs +0.282 on OS-Atlas at $k$=32).

**Shuffled-label control.** To rule out that the DAS lift is an artifact of orthonormal $k$=32 optimization regardless of signal, we train DAS with donor hiddens $h_B$ and the teacher-forced coordinate labels decoupled via independent shuffles: the donor hidden is always $h_B$ but the supervised coordinate is drawn from an independent permutation, so the (hidden, coordinate) pairs carry no consistent mapping. This control reaches a donor-match rate of 0.313 on Qwen2.5-VL-7B (L14), 0.288 on -3B (L21), and 0.246 on OS-Atlas (L19)—above the no-patch reference, reflecting a bias from the training procedure itself, but well below the true DAS subspace. The DAS-minus-placebo gap is +0.461, +0.330, and +0.184 respectively; the paired-bootstrap CI excludes zero on every model.

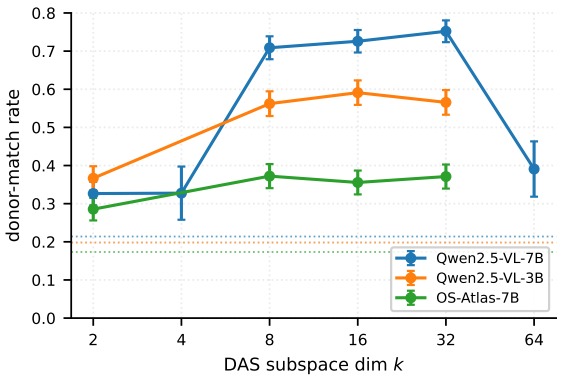


Figure 9: **Interchange success against DAS subspace dimension.** Dotted lines: per-model no-patch reference (chance). Error bars are Wald 95%.

## G Distributed Handle: $k$-Sweep

The causal subspace is not a low-dimensional coordinate axis. The full $k$-sweep $\{2, 4, 8, 16, 32, 64\}$ is available at the original DAS layer per model (L19 for Qwen2.5-VL-7B, L24 for -3B, L21 for OS-Atlas); at the probe-best layers reported in the main text we only have $k \in \{2, 32\}$. Shrinking the DAS subspace to $k = 2$ collapses the transfer to a donor-match rate of 0.327 on Qwen2.5-VL-7B at L19—above the no-patch reference, but well below the plateau the subspace reaches by $k = 8$ (0.709, growing only marginally to 0.752 at $k = 32$; Figure 9). The same shape appears on Qwen2.5-VL-3B at L24 ($k$=2 donor-match 0.367, peak at $k = 16$) and on OS-Atlas-7B at L21 ($k$=2 donor-match 0.286, plateau already at $k = 8$). The probe's rank-2 row space at the same layer reaches

| Model (layer) | DAS $k$=32 | DAS $k$=2 | probe rank-2 | random $k$=32 | PCA $k$=32 | no-patch |
|---|---|---|---|---|---|---|
| Qwen2.5-VL-7B (L14) | 0.78 | 0.35 | 0.22 | 0.23 | 0.22 | 0.22 |
| Qwen2.5-VL-3B (L21) | 0.62 | 0.38 | 0.19 | 0.19 | 0.21 | 0.19 |
| OS-Atlas-7B (L19) | 0.43 | 0.33 | 0.14 | 0.15 | 0.16 | 0.14 |

Table 4: **Interchange donor-match rate at each model's probe-best layer.** DAS subspaces and same-layer controls; the no-patch reference is the donor-match rate without intervention. Numerical companion to Figure 5.

| Model (layer) | baseline | DAS $k$=32 | DAS $k$=2 | probe rank-2 |
|---|---|---|---|---|
| Qwen2.5-VL-7B (L14) | 0.438 | 0.015 | 0.074 | 0.428 |
| Qwen2.5-VL-3B (L21) | 0.349 | 0.048 | 0.122 | 0.340 |
| OS-Atlas-7B (L19) | 0.493 | 0.037 | 0.283 | 0.490 |

Table 5: **Hit rate after amnesic projection of the named subspace at each model's probe-best layer.** Numerical companion to Figure 6.

only 0.218 / 0.190 / 0.171, well below DAS at the same rank. The causal coordinate signal therefore requires $k \approx 8$ dimensions, the same onset across model scale and generation.

**Erasure k-sweep (subtractive side).** The same dimensionality onset appears when the DAS subspace is erased rather than interchanged. On Qwen2.5-VL-7B at layer 19, amnesically projecting out a DAS subspace of dimension $k \in \{2, 8, 16, 32\}$ moves the hit rate from a 0.438 baseline to 0.341 / 0.036 / 0.013 / 0.222 respectively; a size-matched random $k$-dimensional projection at the same layer stays at the baseline at every $k$ (0.428 / 0.431 / 0.428 / 0.427). The collapse is partial at $k$=2, near-complete at $k$=8–$k$=16, and partially recovers at $k$=32 as the larger basis admits less-critical directions that dilute the per-dimension impact. The interchange and erasure sides therefore agree on the same $k \approx 8$ onset.

## H Probe-to-Handle Routing

The routing map $g$ is ridge regression ($\alpha = 1$) from coordinate to subspace value, fit on a training split (50/50, seed 0). The DAS subspace $P$ at each model's probe-best layer ($k = 32$) is from Appendix F; the seed-0 $P$ is reused across all routing cells per model. At test time the subspace component is set with the hook $h' = h + P(\text{target} - P^\top h)$. The *rank-2 oracle* condition injects $g(\text{ ground truth })$; the *probe-coord* condition reads the coordinate with the 5-fold out-of-fold ridge probe at the readout layer (layer 25 for Qwen2.5-VL-7B and -3B, layer 18 for OS-Atlas), sends it through $g$, and injects $g(\text{ probe readout })$, using no ground truth at test time. The *wide-routed* condition uses a second ridge map $g' \colon P^\top h^{(\text{readout})} \mapsto P^\top h^{(\text{interchange})}$ ($\alpha = 1$), fit on the same training split: it reads the full $k$=32 DAS-projection of the readout-layer hidden state and predicts the corresponding probe-best-layer projection without going through the rank-2 probe at any point. The test-time hook becomes $h' = h + P\big(g'(P^\top h^{(\text{readout})}) - P^\top h\big)$. Three corrupted-coordinate controls share the same injection pipeline as $g$: *halfshift* adds 0.5 to each axis modulo 1, *random* draws a uniform coordinate, and *donor* substitutes the activation from another (recipient, donor) pair. On the 1,000-task evaluation set (978 usable for Qwen2.5-VL-7B), the routing failure count is 557 / 654 / 561 for Qwen2.5-VL-7B / -3B / OS-Atlas at the probe-best layer.

## I Head-Level Analysis

**Attention and write scores.** For §5.2 we score every attention head on two axes. The READ score is iou@9: the intersection-over-union between a head's 9 most-attended image patches and the $3 \times 3$-patch region around the target, averaged over digit-emission steps. The WRITE score is the mean magnitude of a head's per-head residual write projected onto the probe's rank-2 row space, obtained by reshaping the pre-`o_proj` activation to (heads, head_dim), projecting through the corresponding slices of `o_proj`, and dotting with the probe basis (probe fit at layer 18, ridge $\alpha = 1$); attn@target is the raw attention mass on the target patch. Attention statistics are computed over 966 samples and write statistics over 47.

**Attribution patching (AtP*).** The signed AtP* score (Kramár et al., 2024) of head $(L, h)$ is $\sum_t a_{L,h,t} \cdot \partial\mathcal{L}/\partial a_{L,h,t}$, where $a_{L,h,t}$ is the per-head

| Model (layer) | wide-routed | probe-coord | rank-2 oracle | max ctrl | Δhit (wr) | OK kept (wr) |
|---|---|---|---|---|---|---|
| Qwen2.5-VL-7B (L14) | 0.56 | 0.40 | 0.44 | 0.10 | +0.16 | 0.63 |
| Qwen2.5-VL-3B (L21) | 0.64 | 0.33 | 0.37 | 0.15 | +0.28 | 0.57 |
| OS-Atlas-7B (L19) | 0.33 | 0.25 | 0.31 | 0.22 | −0.01 | 0.54 |

Table 6: **Routing held-out failure correction rate at each model's probe-best layer.** *max ctrl*: maximum across corrupted-coordinate controls {random, half-shift, pair-swap}. *Δhit (wr)*: change in overall hit rate under wide-routed vs. no-intervention baseline. *OK kept (wr)*: fraction of baseline-OK clicks that remain hits under wide-routed. Numerical companion to Figure 7.

pre-`o_proj` input slice at position $t$ and $\mathcal{L}$ is a per-task loss under teacher forcing. We use two losses, each computed per model. The *digit-logit* variant takes $\mathcal{L}$ as the cross-entropy of the next-digit logit at each digit-emission position against the ground-truth digit, summed over positions. On Qwen2.5-VL-7B (1,000 AgentNet samples, 16,054 digit positions) the top by absolute score forms a negative-signed cluster in L27 (H10, H11, H12, H13 at ranks 4, 5, 11, 12) and a positive-signed cluster across L18–23 led by L21H11, L18H8, L21H12—the signed ranking referenced in §J; the per-model digit-logit top-5 sets used as controls in Table 1 are listed in the head-ablation paragraph below. The *probe-direction* variant takes $\mathcal{L}$ as the squared error between a probe readout (ridge $\alpha$=1, fit at each model's probe-write layer: L18 on Qwen2.5-VL-7B, L24 on 3B, L21 on OS-Atlas-7B) and the target coordinate. The probe-write layer is the layer at which the probe error of Table 1 is measured and sits a few layers downstream of each model's probe-best layer (§4); the resulting top-5 probe-writing heads are listed in §5.2.

**Head ablation.** An ablation zeroes a head's contribution to the residual stream by an `o_proj` forward-pre-hook that sets the head's input slice $x[:, :, h \cdot d : (h{+}1) \cdot d]$ to zero at all token positions ($d$=128 on all three models). The probe-writing analysis of Table 1 is run per model with $n$=1000 AgentNet samples: cumulative ablation of the top-$\{1, 3, 5\}$ probe-writing heads plus two non-probe-writing controls (digit-logit top-5, seeded random-5). The digit-logit top-5 set is taken from each model's own AtP* digit-logit ranking: L21H11/L18H8/L21H12/L27H13/L27H10 on Qwen2.5-VL-7B (one head shared with probe-writing top-5); L35H4/L23H10/L21H12/L21H9/L35H3 on 3B (three shared: L21H12, L23H10, L21H9); L19H20/L19H0/L27H11/L0H10/L19H17 on OS-Atlas-7B (two shared: L19H20, L19H17). The random-5 control is drawn per model from a shared seed over the model's full head set (784 heads on Qwen-7B and OS-Atlas, 576 on 3B). A sanity ablation that zeroes all heads collapses the output on each model, confirming the hook is functional. The auxiliary head-ablation null of §J uses the same hook on an earlier $n$=200 Qwen2.5-VL-7B run, ablating the L27 and L18–23 clusters of Kramár et al. (2024) attribution and a seeded random-5 control.

## J Emission Has No Single Point of Control

§5.2 showed that ablating probe-writing heads breaks the readout but leaves the click intact. Here we ask the complementary question: which heads do drive the click. A search finds no concentrated set. The two AtP* clusters of Appendix I—a negative-signed L27 group and a positive-signed L18–23 group—are the natural candidates for a serial bottleneck on the digit logit, but neither moves the click under ablation, neither carries a digit-specific writing path under output patching, neither shifts under LoRA fine-tuning, and neither sits on a serial route through path patching.

Emission is therefore distributed across heads. Together with the $k \approx$ 8–16 plateau of the DAS interchange (Appendix G), the two analyses give the same picture: the handle is spread over a low-rank subspace whose head-level support has no concentrated driver. The coordinate pathway is distributed at every level we measured—in the readout (§5), in the handle (Appendix G), and, as shown here, in the route from the handle to the output.

**No head set redirects the emission by ablation.** Ablating the L27 and L18–23 clusters' top-$\{1, 3, 5\}$ heads cumulatively leaves the hit rate at 0.41–0.49 against a 0.45 baseline (bootstrap 95% CI roughly ±7 points) and the parse rate at 0.915–0.960. A seeded five-random-head ablation reaches 0.43, indistinguishable from the attribution-top con-

ditions. A damping variant that scales the same head slices by $0.5$ instead of zeroing them produces the same null.

**Head output patching does not steer the emitted digit.** Patching a head's output toward the unembedding direction of a target digit was meant to test whether the L27 cluster carries a digit-specific writing path. It does not. Across the four digit targets and four perturbation strengths we ran ($\alpha \in \{0.1, 0.3, 0.5, 1.0\}$), the patched digit's rate of appearance rises by at most $1.8$ points above its natural rate for the most frequent digits and is flat for the rare ones; the systematic effect of the patch is to drop the parse rate by 10–13 points. The patched output's main change is degradation rather than digit-specific redirection.

**LoRA fine-tuning sharpens the readout side, not the emission side.** After LoRA fine-tuning, the localizer heads of §5.2 attend more sharply to the target—iou@9 rises by up to $0.088$—while the attribution-top L27 heads' iou@9 changes by $0.003$ or less. Fine-tuning sharpens what the model sees on the readout side; the distributed support of the handle's emission, with no concentrated head set to retarget, is left in place.

**Path patching: the handle's effect is also distributed in route.** Path patching confirms that the handle's effect on the output is not routed through any small set of heads. Running a DAS interchange at layer 19 of Qwen2.5-VL-7B while freezing five L27 heads from the AtP*-top cluster (H7, H10, H11, H12, H13—four of which are among the top thirteen heads by absolute signed AtP* on the digit logit, §J) to their clean-run values still retains $98\%$ of the transfer (donor-match rate of $0.736$, against $0.751$ unfrozen and $0.092$ with no interchange; freezing five random L27 heads gives $0.747$, essentially unchanged; $n_{\text{pairs}} = 879$). The handle's influence propagates through the residual stream redundantly: no concentrated set of heads at the output side, and no serial path through them either.

## K Bootstrap Confidence Intervals

All $95\%$ confidence intervals resample the saved per-sample or per-pair records with replacement, $B = 10000$ resamples, seed $0$, taking the $2.5$ and $97.5$ percentiles. The statistic is the corresponding rate (hit rate, donor-match rate, correction rate, failure-conditional decodability). Paired differences between DAS and each control share the fixed interchange pairs and are resampled jointly. No GPU computation is involved; the probe readout is held fixed, so the interval reflects sampling of the evaluation set.

**Single-seed training and the role of cross-model replication.** Each trained component is fit from a single seed: the 5-fold probe split uses `KFold(shuffle=True, random_state=42)`, the DAS $70/30$ split and the routing $50/50$ split both use seed $0$, the random-direction steering control uses per-sample seed $1000+i$, and the random-subspace DAS control uses seed 7. The reported confidence intervals therefore quantify evaluation-set sampling variability rather than seed-to-seed training variability. A multi-seed sweep was not run because the compute budget was allocated to the cross-model replication of §5 and to the $k$-sweep of Appendix G; the qualitative agreement of every reported effect across Qwen2.5-VL-7B, Qwen2.5-VL-3B, and OS-Atlas-7B provides an independent robustness check that a single-model multi-seed sweep would not.

## L Hardware and Reproducibility

The mechanistic experiments of §4–§5 were run on rented $2\times$ or $4\times$H200 (141 GB) GPUs (Vast.ai). Total compute was approximately 160 H200-hours, dominated by the DAS $k$-sweep and the head-ablation sweep.