SEMANTIC STATE ATLAS: STATE-CONDITIONED COORDINATES FOR ACTIVATION INTERVENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation steering can alter model behavior without updating weights, but an edit calibrated at one hidden state need not retain the same effect at another state or representation stream. Local intervention directions can weaken, affect a different behavior, or alter unintended predictions as the anchor changes. We study this problem as *semantic persistence*. Semantic State Atlas represents an intervention with anchor-specific local coordinates and explicitly transfers those coordinates between charts. Each chart selects directions that respond to the requested target while limiting changes to held-out non-target predictions. Transition operators are trained so that a transported edit reproduces the source edit's measured target responses at the destination, with consistency imposed only across semantically equivalent routes. A denoising-based validity surrogate further regularizes edited states toward locally supported regions. On controlled semantic-binding tasks, Chart and Atlas perform almost identically at one hop (91.5% versus 91.6%), but the gap grows to 73.0% versus 85.1% at three hops and 53.2% versus 71.8% at six hops. A matched state-conditioned MLP reaches 76.7% and 59.3%, respectively, indicating that state conditioning alone does not account for the full gain. With paired stream maps fixed, mean cross-stream success rises from 63.8% to 67.5%. Single-step benchmark gains are smaller, while nested procedural edits and added inference cost remain limitations.

1 INTRODUCTION

Activation-level intervention can modify selected behaviors of a vision-language model without updating its parameters. Hidden-state control has been used for instruction following, visual grounding, and hallucination mitigation (Su et al., 2025; Sivakumar et al., 2025; Li et al., 2025). Many interventions are evaluated only at the state where they are constructed. Other settings require the same semantic change to be re-expressed after the internal representation has changed. If an object is temporarily assigned a new role, for example, later relation queries or judgments expressed through another representation stream should still reflect that assignment while unrelated predictions remain stable. The problem is therefore not whether the same activation vector can be reused, but whether the intended behavioral change can be re-expressed at a new internal state.

A shared steering direction makes this difficulty visible because its effect can depend on the hidden state around which it is applied. A direction that changes one spatial relation near its calibration state may weaken elsewhere, affect another behavior, or introduce collateral changes. Related failures have been reported in studies of steering specificity and reliability (Goyal & Daumé, 2026; Braun et al., 2025). Context-dependent and dynamic steering methods reduce this dependence by adapting the update to the current representation (Sivakumar et al., 2025; Yin et al., 2026). They determine how an edit should change with the current state, but do not by themselves define how two local updates should correspond when both are meant to realize the same intervention.

Local coordinates make this distinction explicit. Around an anchor, a local chart can select directions that affect a requested behavior while limiting changes to designated non-target predictions. Changing the anchor changes both the local basis and the model's response to its coordinates. The same semantic intent may therefore require different coordinates at two anchors, while reusing the same coordinates can produce different effects. Independently finding a successful local edit at each anchor can satisfy the local objective, but it provides no explicit correspondence between those real-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

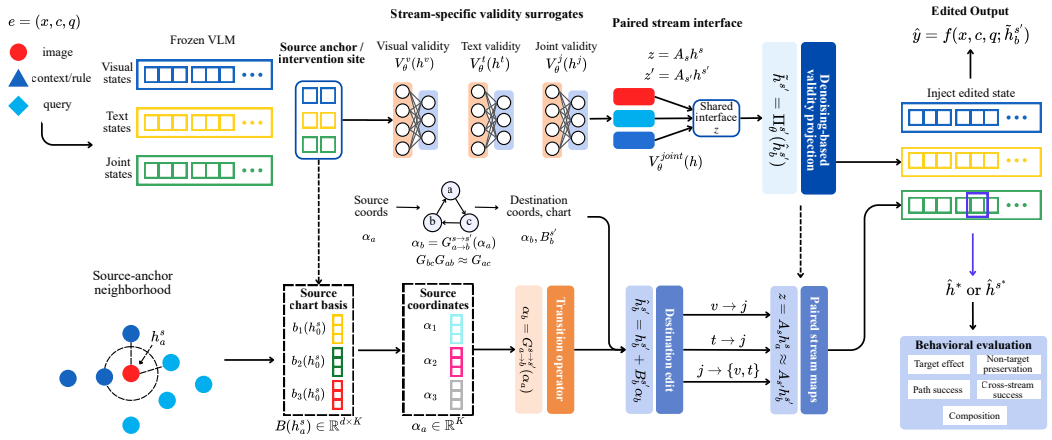

Figure 1: Overview of Semantic State Atlas. Local intervention coordinates are transitioned between anchor-specific charts, regularized by stream-specific validity surrogates, and reused across states or representation streams.

izations. Visual, joint, and textual streams introduce the same problem because their local coordinate systems need not agree.

We call the required property *semantic persistence*. An intervention is persistent along a sequence of anchors when the requested target effect remains present at the required evaluation points and designated non-target predictions remain within their preservation bounds. The coordinates themselves may change. This behavioral definition also separates semantic persistence from geometric consistency: a transition can be reversible or cycle-consistent without preserving the requested behavioral effect.

We introduce *Semantic State Atlas*, illustrated in Fig. 1, which represents an intervention with anchor-specific local coordinates and transforms those coordinates when the intervention is re-expressed at a new anchor. Each chart selects target-sensitive directions while penalizing non-target leakage. Given a destination anchor, a learned transition maps the incoming coordinates into the destination chart before the edit is applied there. Transitions are trained so that the transported edit reproduces the source edit's measured target responses at the destination while preserving non-target predictions. Consistency is imposed only on task-defined semantically equivalent routes, so order-sensitive operations are not forced to commute. A denoising-based validity surrogate further regularizes edited states toward locally supported representations (Luo et al., 2026). For cross-stream use, paired representation maps provide the interface between visual, joint, and textual states, while transition operators carry the intervention coordinates across the resulting charts. The formulation does not require a shared global coordinate system or a globally smooth activation manifold.

We evaluate Atlas on controlled semantic-binding tasks and on VSR, ChartQA, and HallusionBench. Qwen3-VL-8B-Instruct is used for the main mechanism analysis, with Gemma 4 and Qwen3.5 providing additional cross-backbone evaluation. The controlled tasks specify both the requested intervention and predictions that should remain unchanged, allowing us to vary anchor distance, follow paths of one to six transitions, compare equivalent cross-stream routes, and separate independent from order-sensitive compositions. Independent charts test whether repeated local reconstruction is sufficient. A matched state-conditioned MLP receives the same source and destination information and tests whether generic nonlinear state conditioning can account for the effect of explicit transitions. We also remove path consistency while retaining transition supervision and keep paired stream mappings fixed in cross-stream comparisons. Public benchmarks evaluate single-edit behavior outside the controlled persistence setting.

The results show that the advantage of explicit correspondence appears primarily as representation changes accumulate. At one hop, Chart and Atlas obtain nearly identical joint success, 91.5% and 91.6%. At three hops, the gap grows to 73.0% versus 85.1%, and at six hops to 53.2% versus 71.8%. The matched state-conditioned MLP reaches 76.7% and 59.3% at three and six hops, narrowing but

not removing this gap. With paired stream mappings fixed, mean cross-stream success increases from 63.8% for independent charts to 67.5% for Atlas, and the long-path trend also persists after refitting on Gemma 4. Single-step benchmark gains are substantially smaller and task dependent. Nested procedural edits remain difficult, and maintaining cross-chart transitions adds inference cost. These results identify repeated re-expression across changing internal states as the setting where explicit cross-chart correspondence provides its clearest benefit.

2 Related Work

Activation steering and state-dependent control. Activation steering modifies intermediate representations at inference time while leaving model parameters fixed. Recent work has moved beyond a single additive direction by introducing sparse feature spaces, nonlinear or rotational updates, and input-dependent intervention policies (Bayat et al., 2025; He et al., 2025; Hua et al., 2025; Oozeer et al., 2025; Vu & Nguyen, 2025; Wang et al., 2025; Hsu et al., 2026; Yu et al., 2026; Ye et al., 2026b). In vision-language models, related methods adapt interventions using visual evidence, token- or layer-specific signals, sparse features, or instance-specific control (Su et al., 2025; Li et al., 2025; Sivakumar et al., 2025; Yin et al., 2026; Zou et al., 2025; Wu et al., 2025; Liu et al., 2026). These approaches establish that the appropriate update can depend on the current representation. Semantic State Atlas studies the next question: once different states admit different local updates, how should an intervention specified at one state be re-expressed after the representation changes while remaining tied to the same behavioral intent?

Structured intervention spaces and state validity. A complementary line of work constrains where an intervention acts and which behavior it is allowed to change. Sparse, null-space, multi-subspace, and causal-subspace methods seek localized or protected directions (Bayat et al., 2025; He et al., 2025; Zhu et al., 2026; Jiang et al., 2025; Doan et al., 2026), while position-dependent and nonlinear steering relax the assumption that one intervention geometry applies uniformly (Yu et al., 2026; Oozeer et al., 2025). Behavioral studies further show that target success alone does not guarantee selectivity: steering effects can vary across related queries and examples and can alter refusal or jailbreak behavior (Goyal & Daumé, 2026; Braun et al., 2025; Xiong et al., 2026; Li et al., 2026). Generative activation models instead constrain edited states with learned denoising or generative priors (Luo et al., 2026). Our local charts use the same general principles of selectivity and state support around a single anchor. Atlas adds explicit correspondence between these local realizations when the anchor changes.

Multimodal editing and cross-stream consistency. Multimodal knowledge editing studies whether an edited fact or relation generalizes across visual and textual inputs while preserving unrelated behavior. Existing benchmarks and methods cover visual entities, fine-grained relations, modality consistency, multi-granularity knowledge, and reasoning-intensive edits (Du et al., 2025; Ma et al., 2025; Zhang et al., 2025; Zeng et al., 2025; Zhou et al., 2025; Qiu et al., 2026). Recent evidence also shows that an edit successful for paired image–text inputs can fail to propagate through modality-specific pathways (Fu et al., 2026). Our setting keeps model weights frozen and transports an activation-level intervention rather than edited model knowledge. Paired representation maps align visual, joint, and textual states, whereas Atlas transitions map the local coordinates of the intervention. State alignment therefore provides a cross-stream interface, while semantic persistence depends on how the intervention itself is re-expressed through that interface.

3 Problem Formulation

Let f be a frozen vision-language model and $e = (x, r, q, y^\star)$ an evaluation episode, where x is an image, r is optional context or a task rule, q is the primary query, and $y^\star$ is its reference answer. We consider visual, textual, and joint representation streams $s \in \mathcal{S} = \{v, t, j\}$, following recent work that intervenes directly on internal representations of vision-language models (Sivakumar et al., 2025). For stream s, let $\ell_s^\star$ denote the intervention site and

$$h_i^s(e) \in \mathbb{R}^{d_s}$$

the hidden state at anchor i. An anchor indexes a representation state at which an intervention can be realized; it need not denote a time step or imply a continuous trajectory through activation space.

An intervention is specified by a semantic intent κ, which remains fixed as its activation-level realization changes. This distinction is motivated by recent evidence that steering behavior depends on the current context, model block, and the activations from which the steering signal is constructed (Hsu et al., 2026; Ye et al., 2026a). For (e, κ), target probes $\mathcal{Q}_{\mathrm{tar}}(e, \kappa)$ measure the requested behavioral change, while non-target probes $\mathcal{Q}_{\mathrm{non}}(e, \kappa)$ specify predictions that should remain stable, reflecting the distinction between intervention efficacy and specificity emphasized in recent steering evaluations (Goyal & Daumé, 2026; Doan et al., 2026). Applying the intervention at node (i, s) produces an edited state $\tilde{h}_i^s$, which is injected at $\ell_s^\star$ before the model continues. We identify an intervention through its behavioral effect on these probes rather than through a particular activation vector or coordinate representation.

We represent changes in internal representation by a discrete path

$$\pi = \big((i_0, s_0), (i_1, s_1), \ldots, (i_m, s_m)\big).$$

The intervention is initialized at (i_0, s_0) with intent κ. At each subsequent node where the task requires the intervention, its current representation is re-expressed at that node and applied there. Consecutive nodes may differ in anchor, stream, or both, covering same-stream reuse, cross-stream transfer, and multi-hop intervention. The path therefore couples multiple local realizations of one fixed intent rather than evaluating unrelated edits constructed independently at each node.

Let $\mathcal{E}(\pi)$ denote the nodes on π at which the task evaluates the intervention. We say that the intervention is *semantically persistent* along π when, at every $(i, s) \in \mathcal{E}(\pi)$, the required target effect is present and the designated non-target predictions remain within their allowed change. Writing $\mathrm{TargetOK}(e, \kappa, i, s)$ and $\mathrm{Preserve}(e, \kappa, i, s)$ for these two conditions, we define

$$\mathrm{Pers}(e, \kappa, \pi) = \prod_{(i,s)\in\mathcal{E}(\pi)} \mathbf{1}[\mathrm{TargetOK}(e, \kappa, i, s) \wedge \mathrm{Preserve}(e, \kappa, i, s)]\,.$$

The target and preservation thresholds are fixed from calibration data and remain unchanged at evaluation time.

Semantic persistence does not require identical coordinates at different nodes. Geometric consistency is also insufficient on its own: a reversible but semantically incorrect transition can satisfy a cycle constraint, while a zero intervention can be geometrically consistent without producing the requested target effect. We therefore evaluate semantic identity behaviorally and treat state validity as a separate constraint on whether an edited representation remains locally supported.

4 Semantic State Atlas

For node $a = (i, s)$, write $h_a = h_i^s(e)$ and suppress the fixed intent κ. Atlas carries coordinates α_a in an intent-specific local basis B_a. A transition expresses these coordinates in a destination chart, where the resulting candidate state undergoes validity refinement before injection.

4.1 Local Semantic Charts

For each stream, we fit a denoiser $D_\theta^s(h, \sigma)$ on frozen-backbone fitting activations (Luo et al., 2026), where θ denotes its parameters and σ the noise level. A neighborhood of h_a combines nearby fitting states with denoising-consistent perturbations. Its covariance provides orthonormal candidate directions $u_k \in \mathbb{R}^{d_s}$, scored using labeled target and non-target probes at nearby fitting states. Evaluation answers are not used in chart construction.

An answer score is its mean teacher-forced token log-probability. For direction u_k, sensitivity S_k is the absolute centered finite difference of the mean target score. Leakage L_k is the root-mean-square centered finite difference across non-target scores, using the same perturbation magnitude. With weight $\tau \geq 0$, the K highest-ranked directions form

$$R_k = S_k - \tau L_k, \qquad B_a = [u_{k_1}, \ldots, u_{k_K}] \in \mathbb{R}^{d_s \times K}, \tag{1}$$

where $k_1, \ldots, k_K$ are the selected indices. This selection makes the subspace depend on the requested behavior. The dimension K and stream-specific trust-region radius $\rho_s > 0$ are fixed before evaluation. Coordinates $\alpha_a \in \mathbb{R}^K$ define

$$\hat{h}_a = h_a + B_a \alpha_a. \tag{2}$$

A single-factor source uses $\alpha_a = \gamma \mathbf{e}_p$, where $\mathbf{e}_p$ is the pth standard basis vector. The axis and signed strength γ are selected using fitting probes under the target and preservation criteria. Transport episodes specify the source node. Isolated queries retrieve a prepared source for the same intent and source stream, with the query state serving as the first destination. Subsequent nodes receive transported coordinates without reinitialization.

4.2 Learning Cross-Chart Transitions

We fit maps $A_s \in \mathbb{R}^{d_u \times d_s}$ by minimizing squared distances between paired fitting states across streams, with row-orthogonality regularization, and then freeze them. Here d_u is the aligned-space dimension. For each node, define $z_a = \mathrm{normalize}(A_s h_a)$ and $H_a = A_s B_a$, where $\mathrm{normalize}$ denotes ℓ_2 normalization with a fixed denominator floor. The maps provide a common representation interface. A separate learned correction determines how a source displacement is expressed in the destination chart.

For destination $b = (j, s')$, let η collect matrices $U, V \in \mathbb{R}^{d_u \times K}$ and the parameters of a shared network $g_\eta(z_a, z_b, s, s') \in \mathbb{R}^K$. With ridge coefficient $\varepsilon_G > 0$, the correction M_{ab} and coordinate map $G_{ab} \in \mathbb{R}^{K \times K}$ are

$$\begin{aligned} M_{ab} &= I_{d_u} + U \,\mathrm{diag}\big(g_\eta(z_a, z_b, s, s')\big) V^\top, \\ G_{ab} &= (H_b^\top H_b + \varepsilon_G I_K)^{-1} H_b^\top M_{ab} H_a. \end{aligned} \tag{3}$$

Here I_d is the d-dimensional identity and diag forms a diagonal matrix. The ridge solve expresses the corrected source displacement in the destination basis. Transport applies $\alpha_b = G_{ab}\alpha_a$, with $G_{aa} = I_K$ assigned explicitly for an unchanged chart and basis version. The operator depends on both endpoint states and is linear in the incoming coordinates. Shared parameters permit new chart pairs without edge-specific fitting. For fixed charts, orthogonal rebasing transforms their coordinates and G_{ab} consistently, leaving the candidate edit unchanged.

Training tuples $(a, b, \kappa, \alpha_a) \in \mathcal{C}$ pair the same intent across fitting states, using source edits that pass the calibration target and preservation criteria. Let $t_a(\alpha) \in \mathbb{R}^{n_t}$ contain changes in target answer scores on n_t matched probes, evaluated after the finite-step validity solver in Sec. 4.3. Changes are relative to unedited predictions, signed so that the requested effect is positive, and divided by fixed positive probe-family references shared by both endpoints. Corresponding entries test the same semantic condition. Source responses are precomputed and held fixed. Let $\ell_{\mathrm{non},b}(\alpha)$ be the mean token-level KL divergence from unedited to edited predictions on fitting non-target probes, also computed after refinement on fixed unedited continuations. With $\lambda_{\mathrm{non}} \geq 0$, we optimize

$$\mathcal{L}_{\mathrm{sem}} = \mathbb{E}_{\mathcal{C}} \left[\frac{\|t_b(G_{ab}\alpha_a) - t_a(\alpha_a)\|_2^2}{n_t} + \lambda_{\mathrm{non}} \ell_{\mathrm{non},b}(G_{ab}\alpha_a) \right]. \tag{4}$$

This loss trains the destination edit to reproduce the source's calibrated target responses while penalizing non-target changes. It imposes no equality between source and destination coordinates.

Let $\mathcal{T}$ contain same-intent triples (a, b, c) whose direct and two-edge routes have task-defined equivalent semantics, identical endpoint probes, and consistent chart versions. Reversing order-sensitive operations does not produce an eligible route. With Frobenius norm $\|\cdot\|_F$ and weights $\lambda_{\mathrm{path}}, \lambda_{\mathrm{reg}} \geq 0$, we use

$$\begin{aligned} \mathcal{L}_{\mathrm{path}} &= \mathbb{E}_{\mathcal{T}} \left[\frac{\|G_{bc}G_{ab} - G_{ac}\|_F^2}{K} \right], \\ \mathcal{L}_{\mathrm{trans}} &= \mathcal{L}_{\mathrm{sem}} + \lambda_{\mathrm{path}} \mathcal{L}_{\mathrm{path}} + \lambda_{\mathrm{reg}} \|\eta\|_2^2. \end{aligned} \tag{5}$$

The path term compares coordinate transports in the same endpoint bases. Only η is updated. Semantic gradients pass through the frozen backbone and unrolled destination solver, while denoisers, maps, charts, and calibration references remain fixed. Fitting excludes evaluation episodes and probes, and auxiliary components are fitted separately for each backbone.

4.3 VALIDITY-CONSTRAINED EXECUTION

At destination b, the candidate $\hat{h}_b = h_b + B_b\alpha_b$ is refined using a denoising-consistency energy at fixed noise level σ_{eval}. For proximity weight $\beta > 0$, define

$$E_\theta^{s'}(h) = \|D_\theta^{s'}(h, \sigma_{\text{eval}}) - h\|_2^2,$$
$$\Pi_b(\hat{h}_b) \in \underset{\|h-h_b\|_2 \le \rho_{s'}}{\arg\min} \left[E_\theta^{s'}(h) + \beta\|h - \hat{h}_b\|_2^2 \right]. \tag{6}$$

Lower residual indicates greater denoising consistency. The proximity term penalizes departures from the proposed edit, and the trust region bounds displacement from the destination anchor. We approximate the optimization target with T_P projected-gradient steps of size $\delta_P > 0$, initialized by projecting $\hat{h}_b$ onto the trust ball. The anchor and candidate remain fixed during this solve. The numerical result $\tilde{h}_b$ is injected at $\ell^\star_{s'}$.

Same-stream and cross-stream edits use this solver. Subsequent transitions carry the nominal α_b, without replacing it by $B_b^\top(\tilde{h}_b - h_b)$. The path penalty thus constrains nominal coordinate transport, while semantic training and evaluation assess states after validity refinement. Composed intents are applied in their specified order, each initialized at the state produced by the preceding edit. Cached charts are reused only for identical anchors and matching intent and component versions. The fast configuration shares learned parameters and cache policy but reduces T_P. Architecture, source retrieval, and solver and cache details appear in Appendix B.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We evaluate Semantic State Atlas on controlled semantic-binding tasks and on VSR, ChartQA, and HallusionBench. The controlled tasks specify both the requested intervention and predictions that should remain unchanged, allowing us to measure semantic persistence as anchor distance increases, across multi-hop paths, and between visual, joint, and textual streams. Qwen3-VL-8B-Instruct is the primary backbone; Gemma 4 and Qwen3.5 provide additional cross-backbone evaluation. Public benchmarks are used primarily to measure single-edit behavior, where long-range persistence is not explicitly required.

We compare Atlas with global steering, independent local charts, naive coordinate reuse, and a matched state-conditioned MLP that receives the same source and destination information and uses the same chart and validity interfaces. Ablations remove transition supervision, path consistency, validity refinement, or semantic axis selection as appropriate. Calibration data are grouped by scene or episode to avoid cross-split leakage, with at most 384 labeled examples per intent (288 for fitting and 96 for validation). Evaluation episodes are held out from chart construction and transition fitting. We report target performance together with non-target preservation, multi-hop joint success, cross-stream success, and intervention coverage. Results are summarized across completed seeds using mean and sample standard deviation; key paired method differences use 95% scene/episode-cluster bootstrap intervals. Additional implementation, calibration, and runtime details are provided in the appendix.

5.2 PERSISTENCE ACROSS STATES AND STREAMS

Table 1 first compares methods under the controlled semantic-binding setting. Independent local charts improve over global steering, but their advantage is limited when the same intervention must be re-expressed repeatedly. A matched state-conditioned MLP improves 3-hop joint success from 73.0% to 76.7%, showing that conditioning on the source and destination states already recovers part of the loss. Explicit transitions provide a larger gain: Atlas without path consistency reaches 81.5%, while the full model reaches 85.1%. These gains do not come from uniformly larger edits. Atlas improves ID task gain from 2.14 to 2.43 points over independent charts while maintaining similar non-target preservation (94.3% versus 94.0%) and collateral drift (6.2 versus 6.0).

The difference becomes clearer as the representation moves farther from the source. Figure 2 shows that all methods perform similarly at nearby anchors, but their behavior separates with increasing

Table 1: Main comparison on controlled semantic-binding tasks. ID Gain is measured relative to the frozen backbone in percentage points; NTP denotes non-target preservation and Drift denotes collateral drift. The final column reports 3-hop joint success.

Method	ID Gain ↑	OOD ↑	NTP ↑	Drift ↓	3-hop ↑
Global + validity	1.72	65.0	92.4	7.3	64.6
Independent Chart	2.14	67.0	94.0	**6.0**	73.0
Naive coordinate reuse	2.05	66.3	93.2	6.7	68.8
Matched MLP	2.18	67.1	93.4	6.6	76.7
Atlas w/o path consistency	2.30	67.5	94.1	6.1	81.5
Semantic State Atlas	**2.43**	**68.0**	**94.3**	6.2	**85.1**

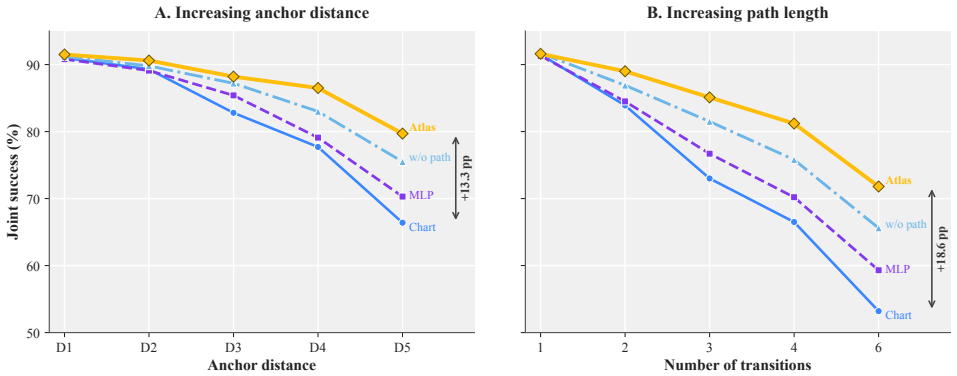

Figure 2: Semantic persistence under increasing representation change. The advantage of Atlas grows with anchor distance and transition length.

anchor distance and path length. At the closest distance, Chart and Atlas obtain 91.0% and 91.5% joint success; at the farthest distance, they fall to 66.4% and 79.7%, respectively. The same pattern appears under repeated transport. At one hop, Chart and Atlas are nearly indistinguishable (91.5% versus 91.6%), whereas at three hops the gap widens to 73.0% versus 85.1%, and at six hops to 53.2% versus 71.8%. The matched MLP reaches 76.7% and 59.3% at three and six hops, while removing path consistency reduces Atlas to 81.5% and 65.6%. The growing separation indicates that the main benefit of explicit cross-chart correspondence appears when an intervention must survive repeated changes of local representation rather than at a single destination.

Cross-stream results show the same distinction between representation alignment and intervention transport. As shown in Table 2, separate stream-specific charts obtain 58.6% mean destination success, and adding paired representation maps alone increases this only to 59.7%. Independent charts using the shared stream interface reach 63.8%, the matched MLP reaches 64.8%, and Atlas without path consistency reaches 65.5%. Full Atlas obtains 67.5%, with improvements in all four evaluated directions: $v \to j$, $t \to j$, $j \to v$, and $j \to t$. Shuffling the paired stream correspondences lowers mean success to 57.5%, confirming that the interface itself must retain meaningful cross-stream structure. At the same time, the gap between paired maps alone and Atlas shows that representation alignment does not by itself determine how an intervention should be re-expressed across streams.

The persistence pattern also reproduces beyond the primary backbone. Table 3 shows that on Gemma 4, Chart and Atlas remain nearly tied at one hop (92.8% versus 93.1%), while the gap widens to 78.2% versus 87.2% at three hops and 59.7% versus 74.0% at six hops. Public single-edit benchmarks show a different regime: Atlas yields modest gains on VSR and HallusionBench, while ChartQA is largely unchanged and occasionally favors the independent chart by 0.1–0.2 points. This contrast supports the interpretation that Atlas is most useful when the same intervention must be re-expressed across changing internal states, rather than as a general mechanism for improving isolated single-step predictions.

Table 2: Cross-stream destination success (%). The mean is the arithmetic average over the four evaluated routes.

Method	$v \to j$	$t \to j$	$j \to v$	$j \to t$	Mean ↑
Representation-interface controls					
Separate charts	63.5	60.7	54.5	55.8	58.6
Paired maps only	64.6	61.8	55.5	56.9	59.7
Shuffled paired maps	61.8	59.5	53.8	54.9	57.5
Intervention transport					
Independent Chart	69.8	66.5	58.7	60.2	63.8
Matched MLP	70.9	67.6	59.6	61.0	64.8
Atlas w/o path consistency	71.7	68.3	60.1	61.8	65.5
Atlas-Fast	72.8	69.3	61.3	63.0	66.6
Semantic State Atlas	**73.4**	**70.1**	**62.1**	**64.2**	**67.5**

Table 3: Generalization across backbones and evaluation settings. Controlled multi-hop experiments test whether the persistence pattern reproduces on an additional backbone, while public benchmarks measure single-edit behavior across three model families. More ablation see A6.

(a) Controlled multi-hop persistence (%)

Backbone	Method	1-hop	3-hop	6-hop
Qwen3-VL	Chart	91.5	73.0	53.2
	Matched MLP	91.3	76.7	59.3
	Atlas w/o path	**91.7**	81.5	65.6
	Atlas	91.6	**85.1**	**71.8**
Gemma 4	Chart	92.8	78.2	59.7
	Matched MLP	92.6	80.3	64.1
	Atlas w/o path	93.0	84.6	69.8
	Atlas	**93.1**	**87.2**	**74.0**

(b) Public single-edit benchmarks (%)

Backbone	Method	VSR	ChartQA	HallusionBench
Qwen3-VL	Frozen	74.5	89.6	61.1
	Chart	76.4	**89.9**	65.5
	Matched MLP	76.5	89.8	65.4
	Atlas	**76.8**	**89.9**	**65.8**
Gemma 4	Frozen	79.1	91.2	66.7
	Chart	80.1	**91.3**	67.7
	Matched MLP	80.2	91.2	67.6
	Atlas	**80.5**	91.2	**67.9**
Qwen3.5	Frozen	82.7	92.2	70.0
	Chart	83.4	**92.3**	**70.8**
	Matched MLP	83.5	92.2	70.7
	Atlas	**83.7**	92.1	**70.8**

5.3 Composition and Efficiency

Figure 3 first examines whether path consistency improves behavior along semantically equivalent routes. Averaged over the two evaluated cross-stream directions, independent charts make both routes succeed in 44.7% of cases and disagree in 21.7%. The matched MLP improves these rates to 48.0% and 18.9%, while Atlas without path consistency reaches 50.7% and 16.6%. Full Atlas further raises both-route success to 54.5% and lowers disagreement to 12.1%. Thus, the path objective contributes beyond pairwise transition supervision, although agreement alone is not used as evidence of semantic success.

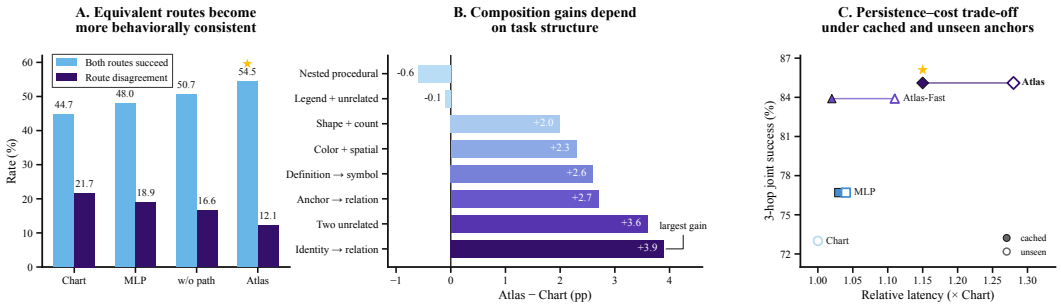

Figure 3: Behavioral consistency, composition, and efficiency. Atlas improves agreement across equivalent routes and most compositions, while additional persistence comes with increased inference cost.

The benefit is not uniform across compositions. Relative to independent charts, Atlas improves six of the eight evaluated composition settings by 2.0–3.9 points, including independent attribute combinations and several order-sensitive sequences. The largest gain occurs for identity-to-relation composition (+3.9 points), while two unrelated edits improve by 3.6 points. In contrast, combining a legend edit with an unrelated value edit is essentially unchanged (-0.1 points), and nested procedural edits decrease slightly (-0.6 points). These cases mark a current boundary of the method: transporting local intervention coordinates helps when successive edits retain identifiable local effects, but does not resolve compositions whose meaning depends on more deeply nested procedural state.

The additional structure also incurs inference cost. On complete 3-hop paths, Atlas improves joint success from 73.0% for independent charts to 85.1%, but its relative latency is 1.15× under exact-chart caching and 1.28× for unseen anchors. The fast configuration retains 83.9% success while reducing these ratios to 1.02× and 1.11×, respectively. The matched MLP is closer to the chart baseline in cost (1.03×/1.04×) but reaches only 76.7% 3-hop success. The main computational overhead therefore appears when new anchors require additional chart and validity processing; reducing the number of validity steps recovers much of this cost with a smaller loss in persistence.

6 Discussion and Conclusion

Semantic State Atlas treats intervention persistence as correspondence between local realizations of the same behavioral intent. Its advantage is small for nearby anchors and single-step edits, but grows as representation changes accumulate. The matched state-conditioned controller recovers part of this degradation, showing that endpoint conditioning is useful, while the remaining gap supports explicit mapping between local coordinate systems. Cross-stream results show a similar distinction: paired representation maps provide an interface between visual, textual, and joint states, but do not specify how an intervention should be re-expressed. Path consistency is imposed only on task-defined equivalent routes and complements behavioral supervision; geometrically consistent transport need not preserve the intended semantic effect.

Local charts use a fixed-dimensional linear approximation and become less reliable farther from the anchor neighborhood. The denoising residual is a local validity surrogate rather than a calibrated likelihood or exact manifold distance. Our visual, textual, and joint streams are pooled states from fixed decoder sites, leaving token-level and multi-layer persistence unexplored. Atlas requires separately fitted auxiliary components for each backbone and adds inference cost for unseen anchors. Nested procedural compositions show little benefit, suggesting that some interventions require richer, history-dependent state. Because the controlled tasks make semantic identity and equivalent routes explicit, the present evidence should not be interpreted as establishing unrestricted long-horizon persistence. These limitations motivate nonlinear or structured local coordinates, amortized chart construction, and intervention states retaining token-, layer-, or history-dependent information. More broadly, our results suggest that activation intervention need not preserve a fixed direction or coordinate vector. Semantic State Atlas instead maintains a behavioral intent through anchor-specific realizations, with its clearest benefit when that intent must be repeatedly re-expressed across changing internal states or representation streams.

AI USE STATEMENT

Generative AI tools were used during the preparation of this work for several supporting purposes. They assisted with language polishing and rewriting, literature retrieval and discovery of related work, and early-stage research ideation and experimental planning. They were also used to help draft and reorganize portions of the manuscript, including the presentation of methods, experiments, figures, and tables. AI tools were not used to generate synthetic datasets or to establish mathematical proofs. All methodological choices, experimental procedures, reported results, citations, and scientific claims were independently reviewed and verified by the authors. The authors take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

We provide the representation interfaces, intervention sites, calibration protocol, local-chart construction procedure, transition parameterization, validity projection, evaluation metrics, and statistical reporting protocol in the appendix. Backbone-specific settings and implementation details required to reproduce the main experiments are also documented there. Code and configuration files will be released with the supplementary material where permitted by the submission policy.

REFERENCES

Reza Bayat, Ali Rahimi-Kalahroudi, Mohammad Pezeshki, Sarath Chandar, and Pascal Vincent. Steering large language model activations in sparse spaces. *CoRR*, abs/2503.00177, 2025. doi: 10.48550/ARXIV.2503.00177. URL https://doi.org/10.48550/arXiv.2503.00177.

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (un)reliability of steering vectors in language models. *CoRR*, abs/2505.22637, 2025. doi: 10.48550/ARXIV.2505.22637. URL https://doi.org/10.48550/arXiv.2505.22637.

Toan Doan, Uyen Le, and Thin Nguyen. Causal activation steering via sparse mediation. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 1079–1097. Association for Computational Linguistics, 2026.

Yuntao Du, Kailin Jiang, Zhi Gao, Chenrui Shi, Zilong Zheng, Siyuan Qi, and Qing Li. MMKE-Bench: A multimodal editing benchmark for diverse visual knowledge. In *The Thirteenth International Conference on Learning Representations*, 2025.

Tingchao Fu, Wenkai Wang, Fanxiao Li, Huadong Zhang, Jinhong Zhang, Dayang Li, Yunyun Dong, Renyang Liu, and Wei Zhou. Correct when paired, wrong when split: Decoupling and editing modality-specific neurons in mllms. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10031–10048. Association for Computational Linguistics, 2026.

Navita Goyal and Hal Daumé, III. Steering safely or off a cliff? rethinking specificity and robustness in inference-time interventions. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5723–5738. Association for Computational Linguistics, 2026.

Zirui He, Mingyu Jin, Bo Shen, Ali Payani, Yongfeng Zhang, and Mengnan Du. SAE-SSV: Supervised steering in sparse representation spaces for reliable control of language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2207–2236. Association for Computational Linguistics, 2025.

Brandon Hsu, Daniel Beaglehole, Adityanarayanan Radhakrishnan, and Mikhail Belkin. Contextual linear activation steering of language models. *CoRR*, abs/2604.24693, 2026. doi: 10.48550/ARXIV.2604.24693. URL https://doi.org/10.48550/arXiv.2604.24693.

Zhenglin Hua, Jinghan He, Zijun Yao, Tianxu Han, Haiyun Guo, Yuheng Jia, and Junfeng Fang. Steering lvlms via sparse autoencoder for hallucination mitigation. In *Findings of the Association*

for Computational Linguistics: EMNLP 2025, pp. 10808–10828. Association for Computational Linguistics, 2025.

Xinyan Jiang, Lin Zhang, Jiayi Zhang, Qingsong Yang, Guimin Hu, Di Wang, and Lijie Hu. MSRS: Adaptive multi-subspace representation steering for attribute alignment in large language models. *CoRR*, abs/2508.10599, 2025. doi: 10.48550/ARXIV.2508.10599. URL https://doi.org/10.48550/arXiv.2508.10599.

Yuxiao Li, Alina Fastowski, Efstratios Zaradoukas, Bardh Prenkaj, and Gjergji Kasneci. Analysing the safety pitfalls of steering vectors. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 11182–11204. Association for Computational Linguistics, 2026.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35799–35819. PMLR, 2025.

Shuliang Liu, Songbo Yang, Dong Fang, Sihang Jia, Yuqi Tang, Lingfeng Su, Ruoshui Peng, Yibo Yan, Xin Zou, and Xuming Hu. Vision-language introspection: Mitigating overconfident hallucinations in mllms via interpretable bi-causal steering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38518–38543. Association for Computational Linguistics, 2026.

Grace Luo, Jiahai Feng, Trevor Darrell, Alec Radford, and Jacob Steinhardt. Learning a generative meta-model of LLM activations. *CoRR*, abs/2602.06964, 2026. doi: 10.48550/ARXIV.2602.06964. URL https://doi.org/10.48550/arXiv.2602.06964.

Yaohui Ma, Xiaopeng Hong, Shizhou Zhang, Huiyun Li, Zhilin Zhu, Wei Luo, and Zhiheng Ma. Comprehendedit: A comprehensive dataset and evaluation framework for multimodal knowledge editing. In *Thirty-Ninth AAAI Conference on Artificial Intelligence*, pp. 19323–19331, 2025.

Narmeen Fatimah Oozeer, Luke Marks, Fazl Barez, and Amir Abdullah. Beyond linear steering: Unified multi-attribute control for language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23513–23557. Association for Computational Linguistics, 2025.

Jiaxing Qiu, Kaihua Hou, Roxana Daneshjou, Ahmed Alaa, and Thomas Hartvigsen. Reasonedit: Editing vision-language models using human reasoning. *CoRR*, abs/2602.02408, 2026. doi: 10.48550/ARXIV.2602.02408. URL https://doi.org/10.48550/arXiv.2602.02408.

Anushka Sivakumar, Andrew Zhang, Zaber Ibn Abdul Hakim, and Christopher Thomas. SteerVLM: Robust model control through lightweight activation steering for vision language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23640–23665. Association for Computational Linguistics, 2025.

Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Li Qing, and Zhaoxiang Zhang. Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12964–12974. Association for Computational Linguistics, 2025.

Minh Hieu Vu and Tan M. Nguyen. Angular steering: Behavior control via rotation in activation space. In *Advances in Neural Information Processing Systems 38*, 2025.

Weixuan Wang, Jingyuan Yang, and Wei Peng. Semantics-adaptive activation intervention for LLMs via dynamic steering vectors. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=c4d26a95fd83f8e590f81c54ae670b5d.

Junfei Wu, Yue Ding, Guofan Liu, Tianze Xia, Ziyue Huang, Dianbo Sui, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. SHARP: Steering hallucination in lvlms via representation engineering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14346–14361. Association for Computational Linguistics, 2025.

Chen Xiong, Zhiyuan He, Pin-Yu Chen, Ching-Yun Ko, and Tsung-Yi Ho. Steering externalities: Benign activation steering unintentionally increases jailbreak risk for large language models. *CoRR*, abs/2602.04896, 2026. doi: 10.48550/ARXIV.2602.04896. URL https://doi.org/10.48550/arXiv.2602.04896.

Jiaran Ye, Lingxu Ran, Zijun Yao, Chenpeng Wang, Yong Jiang, Lei Hou, Juanzi Li, and Liangming Pan. Where steering signals come from: Activation source selection in activation steering. *CoRR*, abs/2607.25270, 2026a. doi: 10.48550/ARXIV.2607.25270. URL https://doi.org/10.48550/arXiv.2607.25270.

Wencheng Ye, Xiaoyang Yuan, Yi Bin, Hengyu Jin, Liang Peng, Pengpeng Zeng, and Heng Tao Shen. RISER: Orchestrating latent reasoning skills for adaptive activation steering. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 4627–4644. Association for Computational Linguistics, 2026b.

Jianghao Yin, Qin Chen, Kedi Chen, Jie Zhou, Xingjiao Wu, and Liang He. Dynamic multimodal activation steering for hallucination mitigation in large vision-language models. *CoRR*, abs/2602.21704, 2026. doi: 10.48550/ARXIV.2602.21704. URL https://doi.org/10.48550/arXiv.2602.21704.

Manjiang Yu, Hongji Li, Priyanka Singh, Xue Li, Di Wang, and Lijie Hu. PIXEL: Adaptive steering via position-wise injection with exact estimated levels under a subspace calibration. In *Proceedings of the ACM Web Conference 2026*, pp. 1574–1585, 2026.

Zhen Zeng, Leijiang Gu, Xun Yang, Zhangling Duan, Zenglin Shi, and Meng Wang. Visual-oriented fine-grained knowledge editing for multimodal large language models. In *IEEE/CVF International Conference on Computer Vision*, pp. 2491–2500, 2025.

Junzhe Zhang, Huixuan Zhang, Xunjian Yin, Baizhou Huang, Xu Zhang, Xinyu Hu, and Xiaojun Wan. MC-MKE: A fine-grained multimodal knowledge editing benchmark emphasizing modality consistency. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 17430–17445. Association for Computational Linguistics, 2025.

Yang Zhou, Pengfei Cao, Yubo Chen, Qingbin Liu, Dianbo Sui, Xi Chen, Kang Liu, and Jun Zhao. M2Edit: Locate and edit multi-granularity knowledge in multimodal large language model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29029–29042. Association for Computational Linguistics, 2025.

Xingyu Zhu, Beier Zhu, Shuo Wang, Junfeng Fang, Kesen Zhao, Hanwang Zhang, and Xiangnan He. Principled steering via null-space projection for jailbreak defense in vision-language models. *CoRR*, abs/2603.22094, 2026. doi: 10.48550/ARXIV.2603.22094. URL https://doi.org/10.48550/arXiv.2603.22094.

Zhengtao Zou, Ya Gao, Jiarui Guan, Bin Li, and Pekka Marttinen. Adaptive residual-update steering for low-overhead hallucination mitigation in large vision language models. *CoRR*, abs/2511.10292, 2025. doi: 10.48550/ARXIV.2511.10292. URL https://doi.org/10.48550/arXiv.2511.10292.

APPENDIX

A EXPERIMENTAL PROTOCOL AND REPRESENTATION INTERFACE

This section specifies the supervision protocol, evaluation structure, and backbone-specific representation interfaces used throughout the experiments. All backbone parameters remain frozen. Auxiliary components are fitted separately for each backbone, so cross-backbone experiments test reproducibility of the intervention mechanism rather than direct parameter transfer.

A.1 DATA PARTITIONING AND SUPERVISION ACCOUNTING

The calibration budget is capped at 384 labeled examples per semantic intent. One labeled example is a distinct input–context–query–required-answer specification. Every externally specified target or non-target probe, including a derived probe, counts toward this budget. Extracting visual, textual, and joint activations from the same labeled example does not consume additional supervision. Repeated model evaluations introduce computation but no additional labels.

Scene or episode groups determine split membership rather than budget units. Groups are assigned once to fitting and validation partitions in a 3:1 ratio using split seed 0. All views, questions, and derived probes from one group remain in the same partition, and group assignment is shared across semantic intents. We retain at most 288 fitting and 96 validation examples per intent. A matched probe bundle is used only when the complete task-required bundle fits within the corresponding cap. If fewer eligible examples are available, we use the available pool and do not supplement it with held-out evaluation examples.

Denoiser fitting, paired-map fitting, chart construction, source calibration, and transition training use fitting data only. Validation data are used for checkpoint and hyperparameter selection. We record labeled-example counts, independent scene or episode counts, and cross-intent reuse separately.

Frozen-model continuations used for preservation scoring introduce no external labels, although their generation and scoring are counted as computation. Controlled comparisons use identical data partitions and supervision budgets unless explicitly stated otherwise.

A.2 EVALUATION TASKS AND PATHS

The controlled semantic-binding suite contains interventions involving local rebinding, temporary definitions, compositional relations, reverse editing, and transport between visual, textual, and joint representation streams. Each episode specifies the requested semantic intent, target probes that should change, and non-target probes whose predictions should remain stable.

The controlled out-of-distribution split holds out combinations of templates, relations, and task rules rather than individual questions from otherwise identical episode groups. This split is fixed before method selection.

Anchor-distance analysis groups eligible source–destination pairs into five equally populated ordered bins, from near to far. These bins are quantiles of the fixed paired-anchor distance score used by the evaluation manifest; they should be interpreted as ordinal representation-change levels rather than physical Euclidean distances in activation space.

Multi-hop evaluation uses paths of

$$1,\ 2,\ 3,\ 4,\ \text{and}\ 6$$

transitions. All methods are evaluated on the same path identities. Every required path node is evaluated under the same target and preservation protocol.

Equivalent-route experiments use only route pairs whose semantics are declared equivalent by the task specification. In particular, a direct route

$$a \to c$$

may be compared with

$$a \to b \to c$$

Table A1: Backbone-specific representation streams and intervention sites.

Backbone	Stream	Decoder block	Hook target	State definition / d_s
Qwen3-VL-8B-Instruct	Visual	18 / 36	Layer-output residual	Mean visual-token state / 4096
	Text	18 / 36	Layer-output residual	Mean text/query-token state / 4096
	Joint	18 / 36	Layer-output residual	Mean multimodal-token state / 4096
Gemma 4	Visual	15 / 30	Layer-output residual	Mean projected visual-token state / 2304
	Text	15 / 30	Layer-output residual	Mean text/query-token state / 2304
	Joint	15 / 30	Layer-output residual	Mean multimodal-token state / 2304
Qwen3.5-9B	Visual	16 / 32	Layer-output residual	Mean visual-token state / 4096
	Text	16 / 32	Layer-output residual	Mean text/query-token state / 4096
	Joint	16 / 32	Layer-output residual	Mean multimodal-token state / 4096

only when both represent the same requested operation and share the same endpoint tests. Reversing an order-sensitive sequence does not create an equivalent route.

Public single-edit evaluation uses VSR, ChartQA, and HallusionBench. These benchmarks complement the controlled persistence setting but do not themselves provide long-path intervention semantics.

A.3 Backbones, Representation Streams, and Intervention Sites

All three representation streams are extracted from the output residual of a decoder block near the midpoint of each backbone. Using comparable relative depth reduces architectural confounding across model families. Within a backbone, visual, textual, and joint states are taken from the same type of residual representation; stream identity changes only the token support used to construct the state.

Let

$$X^{\ell}(e) \in \mathbb{R}^{N \times d_s}$$

denote the residual output of the selected decoder block for episode e. We define token-index sets $\mathcal{I}_v$, $\mathcal{I}_t$, and $\mathcal{I}_j$ for visual, textual, and joint states. The three pooled states are

$$\begin{aligned} h_i^v &= \frac{1}{|\mathcal{I}_v|} \sum_{n \in \mathcal{I}_v} X_n^{\ell}, \\ h_i^t &= \frac{1}{|\mathcal{I}_t|} \sum_{n \in \mathcal{I}_t} X_n^{\ell}, \\ h_i^j &= \frac{1}{|\mathcal{I}_j|} \sum_{n \in \mathcal{I}_j} X_n^{\ell}, \qquad \mathcal{I}_j = \mathcal{I}_v \cup \mathcal{I}_t. \end{aligned} \tag{7}$$

For Gemma 4, $\mathcal{I}_v$ denotes visual-token positions after projection into the decoder hidden dimension.

Here "layer-output residual" denotes the residual tensor produced by the selected decoder block before it is passed to the next decoder block. Intervention sites remain fixed across intents, methods, and evaluation episodes.

State reinsertion. Because the stream representation is mean-pooled, the edited pooled state is mapped back to the token sequence through its displacement. For stream s, define

$$\Delta h_i^s = \tilde{h}_i^s - h_i^s.$$

We apply this displacement uniformly to the corresponding token support:

$$\tilde{X}_n^\ell = \begin{cases} X_n^\ell + \Delta h_i^s, & n \in \mathcal{I}_s, \\ X_n^\ell, & n \notin \mathcal{I}_s. \end{cases} \tag{8}$$

The mean state over the edited support is therefore exactly $\tilde{h}_i^s$. For the joint stream, the displacement is applied to all positions in $\mathcal{I}_j$. All methods compared within a stream use the same extraction and reinsertion rule.

Since all three backbones have $d_s > 256$, the common aligned-space dimension is

$$d_u = 256.$$

B Atlas and Control Implementation

This section specifies chart construction, source initialization, cross-chart transition fitting, validity refinement, and the matched state-conditioned control used in the main comparisons.

B.1 Denoiser and Local Chart Construction

For stream s, let $\mathcal{H}_s$ denote the fitting activations. We define

$$\mu_s = \frac{1}{|\mathcal{H}_s|} \sum_{h \in \mathcal{H}_s} h, \qquad \omega_s = \max\left\{ \left(\frac{1}{|\mathcal{H}_s|} \sum_{h \in \mathcal{H}_s} \|h - \mu_s\|_2^2 \right)^{1/2}, 10^{-6} \right\}. \tag{9}$$

The denoiser operates on normalized activations

$$\bar{h} = \frac{h - \mu_s}{\omega_s},$$

while chart coordinates, trust regions, and validity calculations outside the denoiser remain in native activation units.

Noise levels follow

$$\log \sigma \sim \mathcal{U}(\log 0.01, \log 0.20), \qquad \epsilon \sim \mathcal{N}(0, I_{d_s}),$$

with scalar noise embedding

$$\ell(\sigma) = \frac{2 \log(\sigma / 0.01)}{\log 20} - 1. \tag{10}$$

Let f_θ^s be a two-layer MLP. We define the residual normalized predictor

$$F_\theta^s(u, \ell) = u + f_\theta^s([u; \ell]) \tag{11}$$

and minimize

$$\mathcal{L}_D^s = \mathbb{E}_{h,\sigma,\epsilon} \left\| F_\theta^s \left(\bar{h} + \frac{\sigma}{\sqrt{d_s}} \epsilon, \ell(\sigma) \right) - \bar{h} \right\|_2^2. \tag{12}$$

For a native-unit activation y,

$$D_\theta^s(y, \sigma) = \mu_s + \omega_s F_\theta^s \left(\frac{y - \mu_s}{\omega_s}, \ell(\sigma) \right). \tag{13}$$

The validity energy used during execution is

$$E_\theta^s(h) = \| D_\theta^s(h, \sigma_{\text{eval}}) - h \|_2^2, \qquad \sigma_{\text{eval}} = 0.05. \tag{14}$$

We apply no additional normalization by d_s or ω_s^2.

Chart neighborhood. For anchor $a = (i, s)$, we retrieve up to 64 nearest fitting states in the same stream under Euclidean distance, excluding the anchor's episode group. We additionally draw 256 local perturbations

$$p_r = h_a + \frac{0.10\,\omega_s}{\sqrt{d_s}}\epsilon_r, \qquad \epsilon_r \sim \mathcal{N}(0, I_{d_s}). \tag{15}$$

Proposal ranking uses

$$E_{\text{prop}}(p) = \|D_\theta^s(p, 0.10) - p\|_2^2.$$

A perturbation is eligible only if

$$\|p - h_a\|_2 \le \rho_s, \qquad \rho_s = 0.20\omega_s,$$

and its evaluation-time validity energy is no larger than the 90th-percentile fitting-state energy for that stream. Among eligible perturbations, we retain at most 64 in increasing E_{prop} order. We do not sample additional proposals to fill the quota.

The covariance neighborhood contains the retrieved fitting states and the retained original perturbations p_r. Denoising is used to test local support rather than to replace a retained perturbation with its denoised output.

Centered SVD of the equally weighted neighborhood supplies at most 32 candidate directions. We retain eigenvectors whose covariance eigenvalues exceed 10^{-6} times the largest eigenvalue and denote the resulting count by m_a. Basis construction fails when $m_a < K = 16$ or the required fitting probes are unavailable. When $16 \le m_a < 32$, all retained candidates are scored and exactly $K = 16$ directions are selected.

Candidate signs are canonicalized by requiring the largest-magnitude coordinate to be positive, with the lowest coordinate index breaking ties. Proposal seeds, candidate ordering, and chart basis versions are stored for reproducibility.

Behavioral axis selection. Four nearest eligible fitting episodes are used for chart scoring. Each contributes one target and one non-target probe for the semantic intent. For probe ξ with reference answer $y_\xi = (y_{\xi,1}, \ldots, y_{\xi,L_\xi})$, define

$$\psi_\xi(h) = \frac{1}{L_\xi}\sum_{t=1}^{L_\xi} \log p_\xi\left(y_{\xi,t} \mid y_{\xi,<t}; h\right). \tag{16}$$

For candidate direction u_k,

$$d_{\xi k} = \frac{\psi_\xi(h_\xi + \delta_{\text{ch},s}u_k) - \psi_\xi(h_\xi - \delta_{\text{ch},s}u_k)}{2\delta_{\text{ch},s}}, \qquad \delta_{\text{ch},s} = 0.02\omega_s. \tag{17}$$

Target sensitivity S_k is the absolute mean response over target probes. Leakage L_k is the root-mean-square response over non-target probes. Both signs use identical answer tokens and teacher-forced prefixes. Candidate axes are ranked by

$$R_k = S_k - \tau L_k, \qquad \tau = 0.30, \tag{18}$$

and the top K directions form B_a.

B.2 Source Calibration

The calibration source bank is prepared offline from fitting data. For every eligible source chart, all $K = 16$ axes are tested at

$$\frac{\gamma}{\rho_s} \in \{\pm 0.10, \pm 0.25, \pm 0.50, \pm 1.00\}. \tag{19}$$

Each candidate is evaluated on four target and four non-target probes using the Full validity solver.

A preliminary source candidate must move every target probe in the requested direction, have mean non-target token-level KL no greater than 0.02, and preserve all four unedited greedy non-target continuations. Among candidates satisfying this gate, we select the largest mean signed target response, breaking ties by smaller $|\gamma|$ and then by axis index.

Table A2: Atlas implementation defaults.

Component	Setting
Chart dimension	$K = 16$
Trust radius	$\rho_s = 0.20\omega_s$
Candidate pool	At most 32 covariance directions
Chart score	$R_k = S_k - 0.30L_k$
Denoiser	Two 1024-unit GELU layers with residual normalized output
Denoiser initialization	Xavier-uniform hidden weights; zero biases and output layer; no dropout or normalization layers
Denoiser noise	Log-uniform $\sigma \in [0.01, 0.20]$
Validity noise	$\sigma_{\text{eval}} = 0.05$
Aligned dimension	$d_u = 256$
Paired-map initialization	Transpose of a thin orthonormal factor of a Gaussian matrix
Transition network	Input $[z_a; z_b; z_b - z_a; \text{onehot}(s); \text{onehot}(s')]$; one $4K$-unit GELU layer; K-dimensional $\tanh$ output
Transition initialization	Xavier-uniform hidden weights; zero hidden and output biases; U, V entries sampled from $\mathcal{N}(0, 0.02^2)$
Transition ridge	$\varepsilon_G = 10^{-3}$
Aligned-state norm floor	10^{-8}
Projection	$\beta = 1$, $\delta_P = 0.05$
Projection steps	Full: 8; Fast: 4

For semantic probe family φ, let $\mathcal{R}_\varphi$ contain positive signed target changes from preliminary fitting sources. We define

$$\nu_\varphi = \max\left\{10^{-3}, \text{median}\,\mathcal{R}_\varphi\right\}. \tag{20}$$

This reference is shared across streams for the same semantic test.

For matched target probe j,

$$[t_a(\alpha)]_j = \frac{\chi_j\left[\psi^{\text{edit}}_{a,j}(\alpha) - \psi^0_{a,j}\right]}{\nu_{\varphi(j)}}, \tag{21}$$

where $\chi_j \in \{-1, 1\}$ specifies the requested direction and $\varphi(j)$ identifies the semantic probe family.

Final calibration sources require

$$[t_a(\alpha)]_j \geq 0.8 \qquad \forall j$$

and must satisfy the same preservation criteria. References are not recomputed after filtering. Intents without a valid calibration source remain unsupported and are retained in coverage accounting.

These gates certify source interventions for initialization. They do not by themselves define path-level semantic persistence or joint path success.

The non-target transition term uses

$$\text{KL}(p^0 \| p^\alpha)$$

over the full vocabulary, averaged first over tokens and then over non-target probes. Both distributions are evaluated on prefixes from the fixed unedited greedy continuation, terminated at EOS or after 32 tokens. Source responses and normalization references are precomputed and detached. Destination responses use the Full eight-step validity solver during transition fitting.

Each valid source pairs with up to four fitting destinations per allowed stream. Source and destination share the semantic intent and matched semantic probe slots, belong to distinct episode groups, and are ranked by cosine similarity in the aligned representation interface. Validation uses fitting sources and fixed references with validation destinations.

B.3 Paired Stream Maps and Cross-Chart Transitions

The paired stream maps A_s align representation states across streams; they do not transport intervention coordinates. Coordinate transport is performed by G_{ab} after the endpoint chart bases are mapped into the aligned interface.

For paired states from the same episode, let $\mathcal{P}$ denote the eligible cross-stream pairs. We fit

$$\mathcal{L}_A = \mathbb{E}_{\mathcal{P}}\left[\|A_s h^s - A_{s'} h^{s'}\|_2^2\right] + \lambda_A \sum_s \|A_s A_s^\top - I_{d_u}\|_F^2, \qquad \lambda_A = \max_s \omega_s^2. \tag{22}$$

The maps are frozen before transition fitting.

For node a, define

$$z_a = \mathrm{normalize}(A_s h_a), \qquad H_a = A_s B_a, \tag{23}$$

where normalization uses denominator floor 10^{-8}. The aligned basis H_a itself is not normalized.

The transition network receives

$$[z_a; z_b; z_b - z_a; \mathrm{onehot}(s); \mathrm{onehot}(s')]$$

and contains one $4K$-unit GELU hidden layer followed by a K-dimensional $\tanh$ output. Let

$$g_{ab} = g_\eta(z_a, z_b, s, s').$$

The coordinate transition can be evaluated without explicitly materializing M_{ab}:

$$\begin{aligned} C_b &= H_b^\top H_b + \varepsilon_G I_K, \\ R_{ab} &= H_b^\top H_a + (H_b^\top U)\,\mathrm{diag}(g_{ab})(V^\top H_a), \\ C_b G_{ab} &= R_{ab}. \end{aligned} \tag{24}$$

We solve this positive-definite ridge system by Cholesky factorization. An unchanged chart and basis version receives the exact override

$$G_{aa} = I_K.$$

Orthogonal rebasing. For fixed selected chart subspaces, suppose

$$B'_a = B_a O_a, \qquad B'_b = B_b O_b$$

for orthogonal O_a, O_b. Then

$$G'_{ab} = O_b^\top G_{ab} O_a. \tag{25}$$

With $\alpha'_a = O_a^\top \alpha_a$,

$$B'_b G'_{ab} \alpha'_a = B_b G_{ab} \alpha_a.$$

For an equivalent path,

$$\|G'_{bc} G'_{ab} - G'_{ac}\|_F = \|G_{bc} G_{ab} - G_{ac}\|_F.$$

This property concerns different orthonormal coordinates for the same selected subspaces; it does not imply invariance to replacing the subspaces themselves.

Equivalent-route supervision. Path triples (a, b, c) are included only when the task declares the direct route $a \to c$ and the composed route $a \to b \to c$ semantically equivalent, with identical endpoint probes and compatible chart versions. Reversed order-sensitive operations are excluded.

Transition fitting uses

$$\lambda_{\mathrm{non}} = 1, \qquad \lambda_{\mathrm{path}} = 0.2, \qquad \lambda_{\mathrm{reg}} = 10^{-5}.$$

If an intent has no eligible equivalent triple, its path term is zero and the absence is recorded.

B.4 Matched State-Conditioned MLP

The Matched MLP controls for generic nonlinear state conditioning without Atlas's explicit low-rank coordinate operator. It receives the same endpoint information:

$$z_a, \quad z_b, \quad z_b - z_a, \quad \mathrm{vec}(H_a), \quad \mathrm{vec}(H_b), \quad \alpha_a,$$

together with one-hot source and destination stream identities.

The concatenated input is processed by two $8K$-unit GELU hidden layers and a linear K-dimensional output:

$$\alpha_b^{\mathrm{MLP}} = F_\phi(z_a, z_b, H_a, H_b, s, s', \alpha_a). \tag{26}$$

Table A3: Optimization settings; batch sizes are effective batch sizes.

Component	Updates	Batch	Learning rate	Weight decay
Denoiser	5000	256	10^{-4}	10^{-4}
Paired maps	2000	256	10^{-3}	0
Atlas transitions	2000	8	10^{-4}	0
Matched MLP	2000	8	10^{-4}	0

The control directly predicts destination coordinates and does not use Atlas's low-rank correction $U\,\mathrm{diag}(g)V^\top$ or its ridge coordinate solve.

Matched MLP shares the same calibration examples, calibrated sources, paired maps, chart construction, source–destination behavioral pairs, non-target probes, destination validity solver, and validation budget as Atlas.

For an eligible equivalent triple, its consistency objective is defined through function composition:

$$\mathcal{L}_{\mathrm{path}}^{\mathrm{MLP}} = \mathbb{E}\left[\frac{\|F_{\phi,bc}(F_{\phi,ab}(\alpha_a)) - F_{\phi,ac}(\alpha_a)\|_2^2}{K}\right]. \tag{27}$$

This provides the same task-level equivalent-route supervision but is not numerically identical to Atlas's matrix Frobenius objective. Parameter count and inference cost are therefore reported separately.

B.5 Optimization

All auxiliary models use AdamW with moments $(0.9, 0.999)$, numerical epsilon 10^{-8}, constant learning rates, and global gradient-norm clipping at 1.0.

Each Atlas transition update averages eight source–destination pairs and eight eligible path triples using microbatches of one and gradient accumulation. Only transition parameters η are updated during transition fitting. The frozen backbone, denoisers, paired maps, chart bases, source responses, and normalization references remain fixed.

Gradients from semantic supervision pass through destination backbone computation and all unrolled validity-refinement steps.

Checkpoints are evaluated every 100 updates. Denoisers use validation reconstruction loss under fixed noise draws, paired maps use $\mathcal{L}_A$, and transition models use their validation semantic loss. Ties retain the earlier checkpoint.

The main Qwen3-VL experiments use five seeds,

$$\{0, 1, 2, 3, 4\},$$

while additional-backbone experiments use

$$\{0, 1, 2\}.$$

The no-path-consistency Atlas ablation sets

$$\lambda_{\mathrm{path}} = 0$$

with all remaining settings unchanged.

B.6 Validity Projection and Carried Coordinates

Let $\mathcal{P}_b$ denote Euclidean projection onto the radius-$\rho_{s'}$ ball centered at destination anchor h_b:

$$\mathcal{P}_b(w) = h_b + \frac{w - h_b}{\max\{1, \|w - h_b\|_2/\rho_{s'}\}}. \tag{28}$$

For candidate

$$\hat{h}_b = h_b + B_b\alpha_b,$$

execution begins from

$$h_b^{(0)} = \mathcal{P}_b(\hat{h}_b)$$

and performs

$$h_b^{(t+1)} = \mathcal{P}_b\left(h_b^{(t)} - \delta_P\left[\nabla_h E_\theta^{s'}(h_b^{(t)}) + 2\beta\big(h_b^{(t)} - \hat{h}_b\big)\right]\right). \tag{29}$$

The anchor and candidate remain fixed throughout the solve. We use

$$\beta = 1, \qquad \delta_P = 0.05.$$

Full execution uses $T_P = 8$ iterations and Fast uses $T_P = 4$.

After T_P iterations,

$$\tilde{h}_b = h_b^{(T_P)}$$

is injected at $\ell_{s'}^\star$. This finite-step procedure approximates Eq. 6; we do not claim that it reaches the exact constrained minimizer.

Subsequent transitions receive the nominal coordinate state

$$\alpha_b = G_{ab}\alpha_a$$

rather than coordinates recovered from

$$B_b^\top(\tilde{h}_b - h_b).$$

Thus, $\mathcal{L}_{\text{path}}$ constrains composition of nominal coordinate transports, whereas semantic supervision and behavioral evaluation are computed after validity refinement. The projection may alter the realized hidden-state displacement and is not interpreted as a new semantic coordinate vector.

For multiple semantic intents, interventions are executed in the task-specified order. Each subsequent edit is initialized from the representation state produced by the preceding edit.

Implementation failures. Basis-construction failure, insufficient fitting probes, absence of a valid calibration source, and non-finite transition or projection outputs are treated as failed interventions. Such cases do not trigger an alternative source or a post-hoc intervention construction.

C Computation, Caching, and Runtime

This section separates offline preparation from online intervention cost and defines the cache regimes used in the efficiency analysis.

C.1 Preparation, Caching, and Workload Accounting

Offline preparation fits the denoisers and paired maps, constructs the calibration source bank, searches source axes and strengths, and caches calibrated source responses.

An isolated evaluation query retrieves one prepared source with the same semantic intent and source stream and treats the query state as the first destination. A fresh path root absent from the source bank requires online chart construction and source initialization from fitting data; this work is charged to that evaluation path. Downstream nodes do not repeat source search.

For m_a chart candidates, behavioral chart scoring requires

$$2m_a(4+4) = 16m_a$$

perturbation–probe scoring instances, with a maximum of 512 when $m_a = 32$.

Source search tests

$$16 \times 8 = 128$$

axis–strength candidates. Each candidate is evaluated on four target and four non-target probes, yielding 1024 candidate–probe combinations before additional greedy-preservation decoding.

These quantities are workload counts rather than measured latency and need not correspond one-to-one with full-backbone forward passes. Denoising, validity iterations, unedited reference evaluations, rejected source candidates, autoregressive preservation checks, and cache construction contribute additional computation.

Chart-cache reuse requires an exact match of

$$\{\text{backbone revision}, \text{hook}, \text{anchor state}, \text{stream}, \text{intent}, \text{basis version}\}.$$

Chart bases are not reused across distinct anchors.

Numerical transition reuse additionally requires identical source and destination states, endpoint basis versions, and transition-parameter version. Full and Fast execution share learned parameters and cache contents; Fast changes only the number of validity-refinement steps and therefore does not reduce cold chart construction or source-search workload.

C.2 Runtime Profiling

Runtime profiling uses one NVIDIA A100 80GB accelerator at batch size one. The frozen backbone runs in BF16; auxiliary networks, linear algebra, validity refinement, log-probability calculations, and KL reductions use FP32. Twenty untimed warm-up paths precede 100 synchronized timed paths for each condition.

We distinguish two timing regimes.

Exact-chart-cached. All evaluation-chart and transition entries required by the timed path are prepopulated. Cache-construction cost is measured separately and is not included in online cached latency.

Unseen-anchor. Model weights, fitting-state banks, and the prepared calibration source bank remain resident, but no evaluation-chart or transition entry is available at the beginning of the path. Fresh chart construction and any required path-root initialization are therefore included.

End-to-end timing includes state extraction, source retrieval, all uncached chart work, coordinate transport, validity refinement, and continued decoding. For complete three-hop paths we report wall-time median and 90th percentile, peak memory, intervention coverage, and offline preparation cost.

Relative latency in Fig. 3 is normalized to the Independent Chart median wall time within the same cache regime:

$$\mathrm{RelLatency}(m) = \frac{\mathrm{p50WallTime}(m)}{\mathrm{p50WallTime}(\mathrm{Chart})}. \tag{30}$$

No cache-hit fraction is assumed, and latency is not inferred from denoiser or backbone evaluation counts.

D Evaluation Metrics and Statistical Reporting

This section defines the behavioral quantities reported in the main tables and figures and separates path-level persistence from geometric or numerical consistency.

D.1 Task Gain, OOD Performance, and Preservation

Let $\mathrm{Acc}_{\mathrm{edit}}$ and $\mathrm{Acc}_{\mathrm{frozen}}$ denote task accuracy under the edited and frozen models on the same evaluation set. Task gain is reported in percentage points:

$$\Delta_{\mathrm{task}} = 100\,(\mathrm{Acc}_{\mathrm{edit}} - \mathrm{Acc}_{\mathrm{frozen}})\,. \tag{31}$$

The OOD column in the controlled experiments reports task accuracy on the held-out controlled OOD split described in Sec. A.2; it is not computed from the public benchmarks.

For a designated non-target probe, preservation is evaluated using its predeclared task-specific criterion. Non-target preservation (NTP) is the percentage of designated probes whose required behavior remains unchanged after intervention.

Collateral drift. NTP is a discrete preservation event and can miss distributional movement that does not flip the final answer. We therefore additionally measure full-distribution drift on non-target probes. For $q' \in \mathcal{Q}_{\text{non}}(e)$, let

$$p_{q'}(y) = p_f(y \mid e, q'; h), \qquad \tilde{p}_{q'}(y) = p_f(y \mid e, q'; \tilde{h}).$$

We use symmetric KL divergence,

$$D_{\text{SKL}}(p, \tilde{p}) = \frac{1}{2} D_{\text{KL}}(p \| \tilde{p}) + \frac{1}{2} D_{\text{KL}}(\tilde{p} \| p), \tag{32}$$

where

$$D_{\text{KL}}(p \| q) = \sum_y p(y) \log \frac{p(y)}{q(y) + \epsilon}$$

with a fixed numerical floor $\epsilon > 0$. Collateral drift is

$$\text{CD} = 100 \cdot \frac{1}{|\mathcal{D}|} \sum_{e \in \mathcal{D}} \frac{1}{|\mathcal{Q}_{\text{non}}(e)|} \sum_{q' \in \mathcal{Q}_{\text{non}}(e)} D_{\text{SKL}}\left(p_{q'}, \tilde{p}_{q'}\right). \tag{33}$$

Lower collateral drift is better.

D.2 Semantic Persistence

For controlled response experiments, a target probe satisfies its calibrated response condition when its normalized signed response from Eq. 21 is at least 0.8. Preservation is evaluated using the corresponding predeclared non-target criterion. These thresholds are fixed from calibration data and are not retuned on evaluation examples.

Let $\mathcal{E}(\pi)$ denote the required evaluation nodes of path π. At node $u = (i, s)$, define

$$\text{NodePers}(e, \kappa, u) = \mathbf{1}\left[\text{TargetOK}(e, \kappa, u) \wedge \text{Preserve}(e, \kappa, u)\right]. \tag{34}$$

Path-level semantic persistence is

$$\text{Pers}(e, \kappa, \pi) = \prod_{u \in \mathcal{E}(\pi)} \text{NodePers}(e, \kappa, u). \tag{35}$$

Construction failures remain in the denominator and therefore cannot increase persistence through abstention.

D.3 Joint Path Success

The multi-hop curves use joint path success, which is intentionally distinct from normalized-response persistence. Every required node must produce the task-defined correct target answer and satisfy its preservation condition:

$$\text{JointSuccess}(e, \kappa, \pi) = \prod_{u \in \mathcal{E}(\pi)} \mathbf{1}\left[\text{AnswerCorrect}(e, \kappa, u) \wedge \text{Preserve}(e, \kappa, u)\right]. \tag{36}$$

The 0.8 normalized-response threshold is not added to this answer-based metric.

D.4 Cross-Stream and Equivalent-Route Success

For route $s \to s'$, cross-stream destination success is the percentage of evaluation episodes for which the destination target task succeeds and the corresponding preservation requirement holds.

The reported mean is the arithmetic mean over

$$v \to j, \qquad t \to j, \qquad j \to v, \qquad j \to t.$$

For semantically equivalent direct and indirect routes, *both-route success* requires both route-level endpoint evaluations to succeed. Let C_{dir} and C_{via} denote their binary correctness events. Route disagreement is

$$\text{Disagree} = \Pr\left[C_{\text{dir}} \neq C_{\text{via}}\right]. \tag{37}$$

This is disagreement between behavioral success events rather than a distance between hidden representations or output distributions.

Table A4: Representative boundary cases in which Atlas provides little or no improvement over an independent local chart. Negative differences indicate lower performance for Atlas.

Setting	Metric	Char	Atlas	Δ	Observed boundary
Composition					
Nested procedural	Joint success (%)	42.4	41.8	−0.6	History-dependent structure is not captured by the carried coordinates.
Legend + unrelated	Joint success (%)	80.2	80.1	−0.1	Transport adds no measurable benefit for this weakly coupled composition.
Preservation stress test					
Adversarial non-target set	NTP (%)	88.5	88.2	−0.3	Additional transport does not improve preservation under adversarial probes.

D.5 Composition

Composition success uses the task-defined final behavioral criterion after all requested edits have been executed. Independent and order-sensitive compositions retain their declared execution order; no commutativity assumption is imposed.

For two intents κ_1 and κ_2, an ordered task evaluates

$$\mathcal{R}_{\kappa_2}\left(\mathcal{R}_{\kappa_1}(h)\right)$$

under that order. The reversed sequence is treated as a different task unless the task specification explicitly declares the two orders semantically equivalent.

Nested procedural examples are evaluated with the same final task-defined success criterion and are retained even when they expose a failure mode.

D.6 Boundary Cases and Negative Results

Atlas does not improve every intervention setting. Table A4 summarizes representative cases in which its additional transport structure provides little or no benefit. These cases are reported without filtering and help delimit the regime in which semantic correspondence across local charts is useful.

The failures are structured rather than uniform. Atlas offers little benefit for the weakly coupled legend composition and slightly underperforms Chart on the nested procedural case, where the required intervention depends on execution history rather than only the current local coordinate state. The adversarial preservation test shows a similarly small degradation in non-target preservation. These results support the narrower interpretation that Atlas is most useful when a stable behavioral intent must be repeatedly re-expressed across changing representation states; it is not uniformly superior for isolated, history-dependent, or preservation-dominated edits.

D.7 Public Benchmarks and Intervention Coverage

VSR, ChartQA, and HallusionBench use their benchmark-specific scoring protocols. These benchmarks evaluate single-edit behavior and are not treated as direct measurements of long-path semantic persistence.

Implementation failures remain in the relevant intervention-success denominator. On a public benchmark, a failed intervention returns the frozen model's unedited continuation, which is scored under the benchmark's normal evaluation rule. Intervention coverage is reported separately, so task performance cannot improve by selectively abstaining from difficult examples.

Table A5: Statistical uncertainty for the principal persistence comparisons. Method values are mean ± sample standard deviation across completed training seeds. Δ denotes Atlas minus the corresponding baseline. Confidence intervals are pointwise 95% paired scene/episode-cluster bootstrap intervals on Δ using 10,000 resamples.

Setting	Atlas	Baseline	Δ	95% paired CI
Qwen3-VL multi-hop persistence				
3-hop, vs. Chart	$\mathbf{85.1 \pm 0.6}$	73.0 ± 0.7	$\mathbf{+12.1}$	$[\mathbf{10.8}, \mathbf{13.6}]$
6-hop, vs. Chart	$\mathbf{71.8 \pm 0.9}$	53.2 ± 1.1	$\mathbf{+18.6}$	$[\mathbf{16.2}, \mathbf{20.9}]$
3-hop, vs. Matched MLP	$\mathbf{85.1 \pm 0.6}$	76.7 ± 0.8	$\mathbf{+8.4}$	$[\mathbf{6.9}, \mathbf{9.9}]$
6-hop, vs. Matched MLP	$\mathbf{71.8 \pm 0.9}$	59.3 ± 1.0	$\mathbf{+12.5}$	$[\mathbf{10.3}, \mathbf{14.8}]$
Cross-stream transport				
Mean, vs. Chart	$\mathbf{67.5 \pm 0.5}$	63.8 ± 0.6	$\mathbf{+3.7}$	$[\mathbf{2.5}, \mathbf{4.9}]$
Mean, vs. Matched MLP	$\mathbf{67.5 \pm 0.5}$	64.8 ± 0.7	$\mathbf{+2.7}$	$[\mathbf{1.4}, \mathbf{4.0}]$
Gemma 4 replication				
3-hop, vs. Chart	$\mathbf{87.2 \pm 0.7}$	78.2 ± 0.9	$\mathbf{+9.0}$	$[\mathbf{7.1}, \mathbf{10.8}]$
6-hop, vs. Chart	$\mathbf{74.0 \pm 1.0}$	59.7 ± 1.2	$\mathbf{+14.3}$	$[\mathbf{11.7}, \mathbf{16.9}]$

D.8 Aggregation and Statistical Uncertainty

Metrics are first computed independently for each completed training seed. Tables report the arithmetic mean and sample standard deviation across seeds.

For the principal controlled comparisons, we additionally report pointwise 95% paired bootstrap intervals on method differences using 10,000 resamples. The independent scene or episode cluster is the resampling unit. Questions, paths, and routes derived from the same cluster remain together, and identical sampled clusters are applied synchronously to all compared methods and completed seeds.

The bootstrap interval therefore measures evaluation-cluster variation conditional on the completed trained runs. Training randomness is reported separately through the sample standard deviation across seeds. The intervals are pointwise and are not interpreted as simultaneous multiple-comparison guarantees.

Table A5 reports uncertainty for the principal comparisons supporting the persistence claim. Across the five Qwen3-VL seeds, Atlas exceeds Independent Chart by 12.1 points at three hops and 18.6 points at six hops, with paired 95% confidence intervals of $[10.8, 13.6]$ and $[16.2, 20.9]$, respectively. The corresponding advantages over the matched state-conditioned MLP are 8.4 and 12.5 points, and both intervals remain separated from zero. The cross-stream mean shows smaller but consistent improvements over both Chart and Matched MLP. The same long-path pattern reproduces on Gemma 4 across three seeds, where Atlas exceeds Chart by 9.0 points at three hops and 14.3 points at six hops.

E Additional Analyses

E.1 Component Ablations

We isolate the roles of behavioral transition supervision, semantic axis selection, validity refinement, and path consistency in Table A6. We additionally include an independent-chart baseline and a geometric-transition control to separate the effect of explicit transport structure from semantic supervision. All Atlas ablations retain the remaining training and evaluation protocol unchanged.

Semantic transition supervision provides the largest contribution to persistent transport. Removing it reduces 3-hop and 6-hop success from 85.1% and 71.8% to 69.1% and 49.7%, respectively, and lowers mean cross-stream success from 67.5% to 59.4%. A geometric transition alone recovers part of the multi-hop degradation relative to Independent Chart, reaching 76.2% at three hops and 57.6% at six hops, but does not improve cross-stream success. These results indicate that transport structure alone is insufficient: the transition must be tied to the behavioral effect that defines the intervention.

Table A6: Component ablations and transition controls on the controlled semantic-binding tasks. The 1-, 3-, and 6-hop columns report joint path success (%); NTP denotes non-target preservation, Drift denotes collateral drift, and Cross-stream is mean destination success across the four evaluated stream directions.

Variant	1-hop ↑	3-hop ↑	6-hop ↑	NTP ↑	Drift ↓	Cross-stream ↑
Local and geometric controls						
Independent Chart	91.5	73.0	53.2	94.0	**6.0**	63.8
Geometric transition only	89.8	76.2	57.6	93.7	6.5	62.8
Atlas component ablations						
Atlas w/o semantic transition supervision	85.6	69.1	49.7	93.6	6.4	59.4
Atlas w/o semantic axis selection	89.4	78.3	60.7	90.9	8.0	63.2
Atlas w/o validity refinement	90.9	82.6	67.1	91.8	7.5	66.0
Atlas w/o path consistency	**91.7**	81.5	65.6	94.1	6.1	65.5
Semantic State Atlas	91.6	**85.1**	**71.8**	**94.3**	6.2	**67.5**

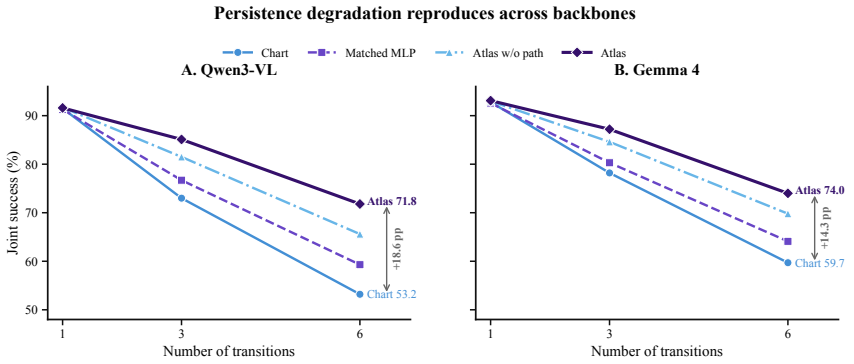

Figure A1: Multi-hop semantic persistence across backbones. Atlas is close to the local-chart baselines at one hop, while its advantage increases as transitions accumulate on both Qwen3-VL and Gemma 4.

The remaining components affect complementary aspects of the method. Removing semantic axis selection increases collateral drift from 6.2 to 8.0 and lowers non-target preservation from 94.3% to 90.9%, while also reducing 6-hop success to 60.7%. Removing validity refinement produces a similar preservation penalty and lowers 6-hop success to 67.1%. Path consistency has little effect at one hop (91.7% without it versus 91.6% for the full model), but its contribution grows with repeated transport: the gap reaches 3.6 points at three hops, 6.2 points at six hops, and 2.0 points in mean cross-stream success. This pattern is consistent with the intended roles of the components: semantic axis selection limits off-target effects, validity refinement stabilizes realized edits, and path consistency primarily constrains accumulated transport rather than single-step intervention.

E.2 Cross-Backbone Persistence

Figure A1 visualizes the cross-backbone trend reported in Table 3. All methods are similar after a single transition, but their degradation rates differ as the path grows. On Qwen3-VL, the gap between Atlas and Chart increases from 0.1 points at one hop to 18.6 points at six hops; on Gemma 4, the corresponding six-hop gap is 14.3 points. The repeated pattern suggests that the benefit is associated with accumulated representation change rather than a backbone-specific single-step gain.

The advantage of explicit coordinate transport is not uniform across intervention regimes. Figure A2 visualizes representative cases in which Atlas provides little or no benefit over an independent local chart.

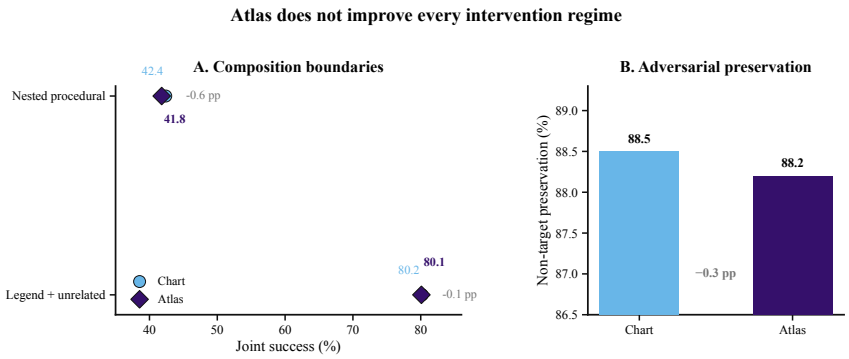

Figure A2: Representative boundary cases. Atlas slightly underperforms the independent chart on nested procedural and weakly coupled compositions, and does not improve preservation under an adversarial non-target stress test.

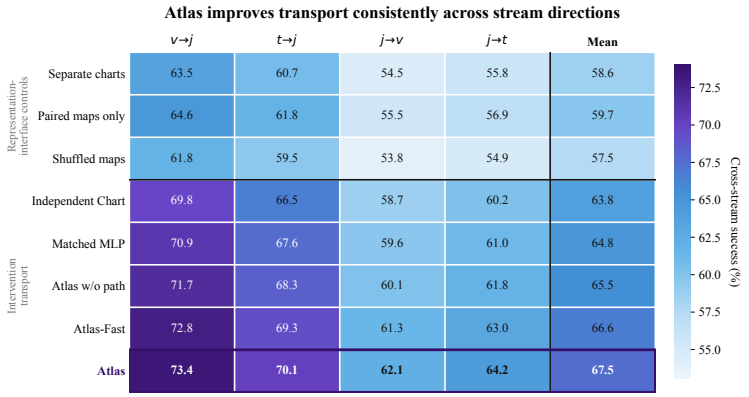

Figure A3: Route-level cross-stream intervention success. Atlas improves all four evaluated transfer directions, while paired representation maps alone provide only limited gains and shuffled maps degrade performance. Joint-to-visual and joint-to-text transfer remain the more difficult directions across methods.

E.3 Boundary Cases and Negative Results

The failures are small in magnitude but structured. The nested procedural case requires information about prior execution history that is not represented by the carried local coordinates, while the weakly coupled legend composition leaves little correspondence for transport to exploit. Under adversarial non-target probes, Atlas also provides no preservation advantage. These cases delimit the claim: the method is most useful when a behavioral intent must be repeatedly re-expressed across changing representation states, rather than for every isolated or history-dependent edit.

E.4 Route-Level Cross-Stream Transport

Figure A3 decomposes the mean cross-stream result by transfer direction. The gain of Atlas is not concentrated in a single route: it improves over Independent Chart on all four evaluated directions. The joint-to-visual and joint-to-text routes remain consistently harder, while shuffled paired maps reduce performance below the corresponding aligned-map controls. This pattern suggests that representation alignment provides an interface for cross-stream transfer, whereas explicit coordinate transport is needed to preserve the intervention across that interface.