WHAT FLOWS WHERE? CONTEXTUAL CROSS-JACOBIAN LENSES FOR DIFFUSION LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vocabulary lenses make intermediate language-model states readable. Most either decode a state directly or replace the downstream network with one map averaged over many prompts and positions. The latter is useful for asking what a state is generally poised to verbalize, but it obscures a different question: what route-conditioned information becomes readable when this source is locally transported toward this recipient in this execution? The distinction is especially visible in diffusion language models, where one visible token can influence several unresolved positions. We introduce the Contextual Cross-Jacobian (CCJ) lens, which linearizes the remaining computation from one source to one recipient in the current forward pass, transports the source state through that local map, and decodes the result with the model's own unembedding. On Dream and LLaDA, CCJ improves target-blind multihop recovery over globally averaged J- and R-fields when every method searches every eligible source layer. Holding the forward pass fixed while changing only the recipient changes the decoded vocabulary, and following the same route through denoising reveals latent concepts before they are highly ranked in the current output distribution. Finally, finite interventions follow CCJ-J's predicted scale, and layers selected without answer labels support activation-only two-way binding swaps. A matched autoregressive comparison finds the same advantage; diffusion provides the additional ability to trace unresolved recipients over time. These results establish CCJ as a training-free, route-conditioned diagnostic of execution-specific source-to-recipient transport that reusable global maps average away.

1 INTRODUCTION

Language models transform token sequences through many layers of hidden computation before producing an answer. Vocabulary lenses provide a rare direct view into this process by translating intermediate vectors into words (nostalgebraist, 2020; Belrose et al., 2025; Gurnee et al., 2026; Blank et al., 2026). A vocabulary readout, however, can answer two different questions. It can describe what is decodable from a state in general, or what becomes readable after transporting that state along a particular downstream route in the current execution.

Existing lenses primarily address the first question. They either decode an intermediate state directly or transport it through a reusable linearization of the downstream network. Reuse is valuable: one map can summarize what states at a layer are generally poised to verbalize. The same averaging, however, collapses variation among prompts, source positions, and recipients. It therefore cannot say which part of a source state is routed to which destination in one forward pass.

The distinction becomes concrete in a prompt with two masked facts: *The capital of France is [MASK A]. The capital of China is [MASK B].* At an intermediate layer, the state at *France* may contain many decodable associations. Yet "what is represented at *France*?" and "what will flow from *France* to *MASK A*?" are different questions. The second should produce a different answer if the recipient is changed to *MASK B*, even though the source state itself is unchanged.

We introduce the Contextual Cross-Jacobian (CCJ) lens to preserve that distinction. For prompt X, layer ℓ, source p, and recipient q, CCJ computes the local map from $h_{\ell,p}(X)$ to the final residual

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.

at q, applies this map to the source state, and decodes the transported vector with the model's own unembedding. The resulting vocabulary ranking is indexed by an ordered route (X, ℓ, p, q), rather than only by a layer and hidden state. In the running example, the same *France* state can therefore be read separately along its routes to the two masks.

Masked diffusion language models (DLMs) provide a natural setting in which to study such routes. Their denoising computation is bidirectional: visible and unresolved positions interact while several masks remain open (Sahoo et al., 2024; Lou et al., 2024; Nie et al., 2025; Ye et al., 2025). We can hold one recipient masked, restore the surrounding context in stages, and ask when a particular incoming route begins to verbalize a latent concept. This temporal view is especially natural in masked diffusion, although the underlying question of source-to-recipient conditioning applies to transformers more broadly.

The paper follows a single progression from readout to mechanism. We first ask whether retaining the current route improves target-blind vocabulary recovery over globally averaged transport. If it does, we then ask whether the gain actually follows the chosen recipient and how it develops as a diffusion state is denoised. Finally, we intervene on the identified routes to test whether their predicted effects appear in the model. Each stage addresses an alternative explanation left open by the preceding one.

Our contributions are:

- We define a training-free, full-vocabulary readout for an ordered source-to-recipient route. A conditional-risk decomposition identifies the prompt, source, and recipient variation that a reusable global field necessarily averages away.
- On Dream and LLaDA multihop, contextual transport improves target-blind concept recovery over global J/R fields when every method searches every eligible layer. Recipient controls and denoising trajectories show that the additional signal is conditional on where it is routed and can become readable before it is prominent in the current output distribution.
- Finite-dose calibration and activation-only binding swaps connect the vocabulary readout to causal effects on the computation. A matched autoregressive comparison shows that contextual conditioning is not unique to diffusion, while masked denoising makes its temporal structure directly observable.

Together, these results turn the broad question "what is represented here?" into the more precise question "what route-conditioned information is readable from this source toward that recipient?" They establish CCJ as a route-conditioned vocabulary lens for information that reusable layerwise maps average away.

2 Related Work

Vocabulary-space lenses provide the closest methodological precedent. The logit lens applies the model's unembedding to an intermediate residual, yielding a model-native ranking without fitting a probe (nostalgebraist, 2020). Its simplicity assumes that intermediate residuals are readable in the final-state basis. The tuned lens relaxes that assumption by learning a layer-specific affine translator (Belrose et al., 2025). Both methods ask what can be decoded from a state at a selected layer and position; neither isolates the downstream computation from that state to a particular recipient.

J-lens incorporates the remaining computation by linearizing the network downstream of an intermediate state and averaging those Jacobians over a calibration corpus (Gurnee et al., 2026). The resulting layerwise field is reusable across examples: multiplying it by a hidden state and applying the model's unembedding gives a training-free, full-vocabulary readout. This construction has exposed intermediate answers, semantic associations, and prompt-level abstractions that are not directly legible under the logit lens. R-lens keeps the same transport-and-unembed recipe while replacing the ordinary backward pass with activation-aware layer-wise relevance propagation (Blank et al., 2026). Thus, J-lens and R-lens differ in how transport is estimated, but both summarize that transport in a reusable layerwise map.

CCJ changes a different axis: the granularity at which the map is retained. CCJ-J uses the ordinary Jacobian and CCJ-R uses the relevance-style rule, but neither averages over the current prompt X, source p, or recipient q. The method therefore preserves the model-native vocabulary and the

transport-then-verbalize construction of J/R lenses while moving from a layerwise average to an ordered route (X, ℓ, p, q). Our global baselines instantiate the averaging design in the same masked-query states, and our multihop and association benchmarks come from the released J-lens evaluation suite. The comparison consequently asks whether execution-specific conditioning adds information beyond the reusable field, not whether one unembedding vocabulary is preferable to another.

Route-level readouts also sit between probing and intervention. Causal mediation, causal tracing, activation patching, and circuit analysis test whether internal components influence model behavior (Vig et al., 2020; Meng et al., 2022; Wang et al., 2022), but they do not usually assign a full-vocabulary interpretation to an ordered source-to-recipient map. CCJ supplies that interpretation, and our finite-dose and activation-swap experiments test whether it predicts the effects of changing the source.

Masked and discrete diffusion language models make this connection especially visible. They generate by iterative denoising rather than left-to-right factorization (Sahoo et al., 2024; Lou et al., 2024; Nie et al., 2025; Ye et al., 2025), so one visible position can simultaneously influence several unresolved tokens. Keeping a recipient masked while restoring its context turns a static route-level readout into a trajectory. This motivates the construction we define next.

3 Contextual Cross-Jacobian Lens

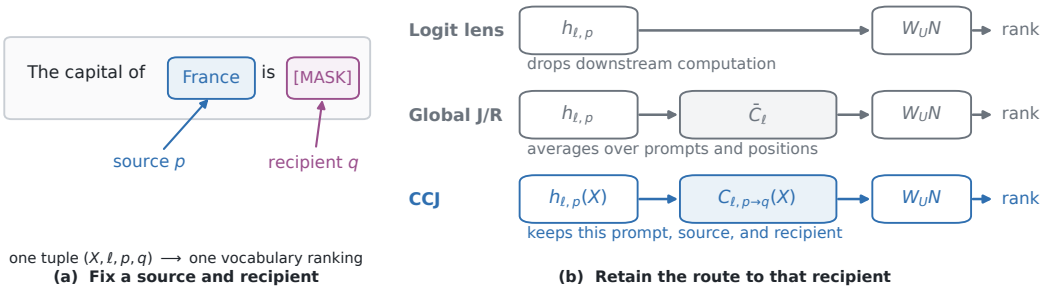

Figure 1: A route-conditioned vocabulary lens. (a) CCJ selects a source state and a final recipient. **(b)** The logit lens omits downstream computation; global J/R applies one reusable averaged route; CCJ retains the route instantiated by the current prompt, source, and recipient.

The construction has three operations. We choose a source state and a final recipient, linearize the remaining computation between them around the current forward pass, and verbalize the transported source with the model's own vocabulary head. Figure 1 contrasts this route-specific map with direct decoding and with a field averaged across executions.

3.1 Route-conditioned transport

Let a transformer process a sequence X of n token positions with residual width d. At layer ℓ, let $H_\ell(X) \in \mathbb{R}^{n\times d}$, and denote the residual at source position p by

$$h_{\ell,p}(X) = H_\ell(X)[p, :] \in \mathbb{R}^d.$$

Let $r_{L,q}(X) \in \mathbb{R}^d$ be the final residual at a selected recipient position q. In masked-language experiments, q is the masked reporter or answer position.

With this notation, the contextual cross-Jacobian (Figure 1) is

$$C_{\ell,p\to q}(X) = \frac{\partial r_{L,q}(X)}{\partial h_{\ell,p}(X)} \in \mathbb{R}^{d\times d}.$$

Each entry measures how one destination coordinate changes under an infinitesimal change to one source coordinate, around the current prompt and activation state. Unlike a reusable field, this operator retains the prompt X, source p, and recipient q.

We transport the source state through this local route:

$$v_{\ell,p\to q}(X) = C_{\ell,p\to q}(X) h_{\ell,p}(X).$$

The vector $v_{\ell,p\to q}$ is the source state after local first-order transport along this route. It is not an additive decomposition of the recipient residual: downstream nonlinearities prevent that interpretation in general.

For the running example, p may be the position of *France* and q the first mask. Replacing q with the second mask leaves the forward activations unchanged but produces a different operator $C_{\ell,p\to q}(X)$. This recipient dependence is the central difference from applying one global field to the *France* state.

3.2 Vocabulary readout and estimators

Let N be the model's final normalization and $W_U \in \mathbb{R}^{|\mathcal{V}|\times d}$ its unembedding. CCJ produces full-vocabulary scores

$$s_{\ell,p\to q}(X) = W_U N(v_{\ell,p\to q}(X)) .$$

Sorting s yields a standalone vocabulary ranking for one (X, ℓ, p, q) tuple. These scores are diagnostic readouts rather than the model's actual output logits at q. For a source perturbation δh, the corresponding first-order change in the model's actual logits is $W_U \, \mathrm{D}N(r_{L,q}) \, C_{\ell,p\to q}\delta h$, a distinct estimand evaluated separately in Figure 4.

We use two route-conditioned backward rules. CCJ-J uses the ordinary Jacobian and is our predeclared primary estimator. It is the operator analyzed in Section 4 and has the direct first-order intervention interpretation. CCJ-R keeps the same route indexing but substitutes the relevance-style backward rules of R-lens (Blank et al., 2026). This often yields cleaner vocabulary rankings, but it is no longer the exact local derivative. We therefore report CCJ-R as a secondary readout and do not choose J versus R separately for each model.

The comparison methods differ in how much downstream structure they retain. The logit-lens operator decodes the source directly:

$$s^{\text{logit}}_{\ell,p} = W_U N(h_{\ell,p}).$$

A global Jacobian field applies one averaged operator $\bar{C}_\ell$:

$$s^{\text{Global-J}}_{\ell,p} = W_U N(\bar{C}_\ell h_{\ell,p}), \qquad \bar{C}_\ell \approx \mathbb{E}_{X,p,q}[C_{\ell,p\to q}(X)].$$

The global relevance field is analogous. All comparisons therefore share the same unembedding; the distinction is whether transport is omitted, globally averaged, or conditioned on (X, p, q).

3.3 Source selection and computation

When a source span is not predefined, we evaluate every eligible source position and choose

$$p^*_\ell(X, q) = \arg\max_p \|v_{\ell,p\to q}(X)\|_2 .$$

This rule measures the magnitude of the routed signal without inspecting the gold vocabulary item. It does not guarantee semantic relevance and is treated as a target-independent heuristic rather than a mechanistic theorem.

The exact operator is recovered without finite differences. Batched vector–Jacobian products recover groups of destination coordinates with automatic differentiation. For width d and reverse batch size b, exact reconstruction needs $\lceil d/b \rceil$ reverse sweeps per recipient; one sweep covers all retained source positions. This explains both the "hundreds of backward passes" intuition and the substantial cost measured in Section 6.4. We next formalize what is gained by conditioning this operator on the current route rather than replacing it with an average.

4 Why Conditioning Can Help

The distinction between a global field and CCJ can be stated as a prediction problem. A global field must predict the local transport operator without seeing the current route, whereas CCJ observes the prompt, source, and recipient that determine it. Fix a model and layer, let (X, P, Q) vary under the evaluation distribution, and write $C = C_\ell(X, P, Q)$. For an information set $\mathcal{G}$, the minimum-risk estimator of C under squared Frobenius loss is $\mathbb{E}[C \mid \mathcal{G}]$, with risk

$$\mathcal{R}(\mathcal{G}) = \mathbb{E}\left[\|C - \mathbb{E}[C \mid \mathcal{G}]\|^2_F\right].$$

For nested information sets $\mathcal{G} \subseteq \mathcal{H}$, the conditional-expectation Pythagorean identity gives

$$\mathcal{R}(\mathcal{G}) - \mathcal{R}(\mathcal{H}) = \mathbb{E}\left[\|\mathbb{E}[C \mid \mathcal{H}] - \mathbb{E}[C \mid \mathcal{G}]\|_F^2\right] \geq 0.$$

Conditioning can therefore only reduce the best achievable squared operator error. Revealing Q, then P, and finally X decomposes the error of a global field into recipient variation, source-given-recipient variation, and the remaining prompt-context variation. The proof and full decomposition are in Appendix B.

This identity explains why contextual transport can preserve information that an average removes, but it does not guarantee a better vocabulary rank: normalization, unembedding, and sorting are nonlinear. A preregistered scalar summary of route heterogeneity is likewise insufficient to predict the ranking gain (Appendix F). The theory therefore motivates the comparison; the experiments must establish whether the operator advantage survives verbalization and whether the resulting routes are causally meaningful.

5 Experimental Setup

5.1 Models, benchmarks, and common protocol

We study Dream-v0-Instruct-7B (Ye et al., 2025) and LLaDA-8B-Instruct (Nie et al., 2025). Our primary benchmark is the released J-lens suite (Gurnee et al., 2026): 93 multihop items containing 103 intermediate-concept instances and 102 semantic-association passages. We retain concepts represented by one tokenizer token, leaving 94 multihop observations per model, 99 Dream association observations, and 94 LLaDA association observations. For association, we append *The implied concept is [MASK]* and restrict eligible sources to the original passage. A matched-scale LLaMA-3.1-8B-Instruct experiment tests whether the effect extends beyond diffusion (Grattafiori et al., 2024).

Concept strings, rather than individual records, are split into 40% development and 60% test partitions so repeated concepts cannot cross the split. At each candidate layer and for each method, the source is selected by transported-vector norm without target access. Development MRR selects one layer per model, task, method, and source rule; that layer is frozen before unseen test concepts are scored. A fixed-last eligible token is a pre-specified source control. We repeat selection over 100 deterministic, concept-disjoint splits. The common metrics are mean reciprocal rank (MRR), precision at one (P@1), and precision at ten (P@10). The primary Dream/LLaDA study evaluates every eligible source layer: 26 for Dream and 30 for LLaDA. The preregistered four-layer grids remain a sensitivity analysis and are used for the matched-scale LLaMA control.

5.2 A staged evaluation of route conditioning

The experiments follow the three questions posed in the introduction. We first ask whether conditioning improves vocabulary recovery without using the answer to choose either a source or a layer. For this all-layer ranking comparison, all four transport methods use the same benchmark records, checkpoints, final transport endpoints (27 for Dream and 31 for LLaDA), and VJP batch size 32. Global J/R are fitted independently on a deterministic $1{,}000 \times 128$ sample from the Dolma 3 Dolmino Mix introduced with OLMo 3 (Olmo et al., 2026), drawn across all 100 local shards. The exact sampling and corpus hash are reported in the appendix. We evaluate every eligible source layer and report one fixed concept-disjoint 40/60 split, 100 repeated splits, fixed-last controls, and a target-aware all-layer oracle diagnostic.

Having established whether conditioning helps, we test what the extra signal depends on. Recipient-controlled prompts place two unrelated facts and two masked answers in one forward state. We hold every source activation fixed and change only the recipient used by the cross-Jacobian, making the own-versus-decoy comparison an isolated test of dependence on q. We then follow that recipient through denoising at 0%, 25%, 50%, 75%, and 100% context restoration with three mask seeds. The recipient remains masked throughout. Oracle trajectories restore gold context tokens; model-filled trajectories use the model's own reconstructions. We compare the current output distribution, logit lens, global J/R fields, and CCJ using MRR area under the denoising curve, early P@10, and hidden-reveal events in which a route ranks a concept that remains weak in the current output.

Finally, we ask whether the readout predicts what happens when the source is changed. Finite-dose interventions apply both signs at 1%, 2.5%, 5%, and 10% of source-state norm and compare predicted with observed target-logit changes. Because this test uses the same derivative that defines CCJ-J, a second experiment removes the derivative from the intervention itself: CCJ-J selects one layer without answer labels, after which we swap actual donor residual states between two argument positions. Random-site, decoy-site, unrelated-donor, and norm-matched random controls test whether the resulting two-way answer swap is specific to the selected binding.

Uncertainty follows the natural independent unit of each experiment: normalized concept for denoising, matched pair for swaps, and relation/entity clusters for transplants. Unless stated otherwise, intervals use 5,000 bootstrap replicates. Intervals over the 100 reused development/test partitions describe *split sensitivity*, not population sampling uncertainty. Full prompts, filtering decisions, sample counts, and implementation details appear in the appendix.

6 RESULTS

6.1 CONDITIONING RECOVERS INFORMATION THAT GLOBAL FIELDS MISS

A contextual lens is useful only if retaining the route improves readout without target leakage. The most direct alternative explanation is inadequate layer search, so Table 1 gives every contextual and global method all 26 Dream or 30 LLaDA source layers. The oracle columns choose the best layer after observing the target and are diagnostic only. The 40/60 columns freeze one development-selected layer before scoring held-out concepts. Appendix Figure 5 shows the corresponding target-blind development profiles.

Table 1: **All-layer multihop comparison (MRR).** Every method searches all eligible layers. Oracle selection sees the target; 40/60 selection freezes one development-selected layer for held-out concepts. The source is selected by transported-vector norm. D/L denotes Dream/LLaDA.

		Dream		LLaDA	
Method	Layers D/L	Oracle	40/60	Oracle	40/60
GLOBAL-J	26/30	.0158	.0004	.0169	.0066
CCJ-J	26/30	**.3889**	**.0779**	**.2112**	**.0701**
GLOBAL-R	26/30	.0111	.0017	.0143	.0067
CCJ-R	26/30	**.4453**	**.0893**	**.2901**	**.0832**

The four norm-selected fixed-split paired intervals are clearly positive. Relative to GLOBAL-J, CCJ-J improves MRR by $+0.0775$, 95% CI $[0.0203, 0.1467]$, on Dream and $+0.0635$ $[0.0190, 0.1177]$ on LLaDA. Relative to GLOBAL-R, CCJ-R improves MRR by $+0.0876$ $[0.0319, 0.1540]$ and $+0.0765$ $[0.0260, 0.1398]$. Direct CCJ-R–CCJ-J differences are small and their intervals cross zero on both models, so the supported conclusion is that conditioning matters, not that one backward rule decisively wins.

The result is also stable to the concept split. Across 100 deterministic splits, norm-selected multihop MRR is $0.0525/0.0536$ for CCJ-J and $0.1258/0.0600$ for CCJ-R on Dream/LLaDA, versus $0.0026/0.0067$ for GLOBAL-J and $0.0111/0.0124$ for GLOBAL-R. Association is harder but shows the same global-versus-contextual separation. Notably, Dream CCJ-R obtains MRR 0.0883 on the fixed split and 0.0628 averaged over repeated splits, versus 0.0003 and 0.0002 for GLOBAL-R. This Dream gain depends on norm-based source routing and weakens under the fixed-last control. Complete P@10, association, source-rule, and paired results appear in Appendix E. We retain CCJ-J as the theory-aligned primary estimator rather than choosing J versus R after observing each model's result.

These comparisons rule out sparse baseline depth coverage, but they do not yet show that the additional information is tied to the intended destination. We test that stronger claim next.

6.2 THE SIGNAL FOLLOWS THE RECIPIENT AND EMERGES DURING DENOISING

In two-fact, two-mask prompts, own-recipient readouts outperform decoy readouts despite identical forward activations. For the mean-state CCJ-R diagnostic, P@10 is 95.0% versus 55.0% on Dream and 97.5% versus 60.0% on LLaDA. The forward pass and source state are identical in the two conditions. The rank change must therefore come from conditioning the backward route on q, rather than from attaching a stronger semantic decoder to the source.

We next trace this recipient-conditioned signal through denoising.

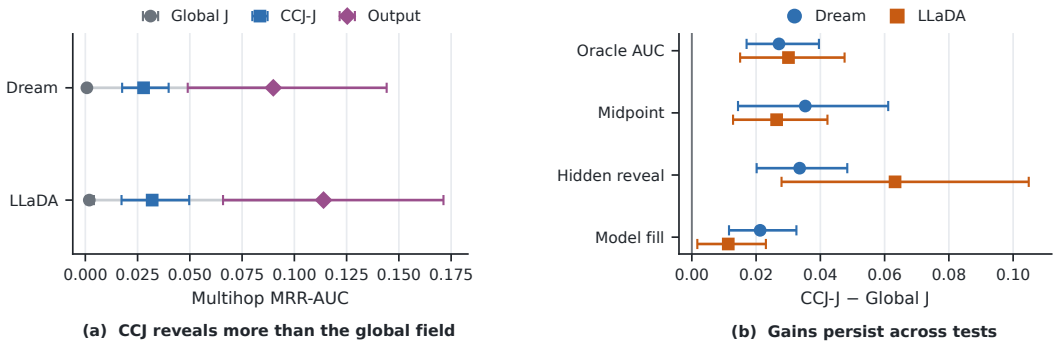

Figure 2: **Contextual transport reveals latent concepts during denoising.** **(a)** Multihop oracle MRR-AUC with normalized-concept bootstrap intervals. CCJ-J substantially exceeds GLOBAL-J; output logits are shown as a complementary reference. **(b)** Paired CCJ-J gains over GLOBAL-J are positive for oracle AUC, the midpoint, hidden-reveal events, and collision-free model-filled trajectories in both models.

On multihop oracle trajectories, Dream obtains CCJ-J MRR-AUC 0.02787 versus 0.00075 for GLOBAL-J (gain 0.02712, 95% CI $[0.01704, 0.03956]$). LLaDA obtains 0.03198 versus 0.00190 (gain 0.03008, $[0.01503, 0.04754]$). Output logits remain stronger overall; CCJ-J instead localizes route-level evidence, with positive midpoint, hidden-reveal, and collision-free model-filled gains over GLOBAL-J in both models (model-filled: 0.02126 Dream, 0.01134 LLaDA).

CCJ-J further identifies hidden-reveal states in which route-conditioned evidence becomes visible before the concept is highly ranked by the current output distribution. The hidden-reveal advantage is 0.03356 on Dream and 0.06325 on LLaDA. These events expose information specific to an incoming source-to-recipient route rather than duplicating the model's current answer distribution.

Figure 3 makes both forms of conditioning concrete. The same source state yields different vocabulary rankings when only its recipient changes, and CCJ-J can enter the top 10 during denoising while both the output and GLOBAL-J remain outside it. Subsequent rank changes show how the contribution of that route evolves as the surrounding diffusion state is restored.

Recipient controls give the gain a semantic interpretation, and denoising trajectories give it a temporal one. Both remain observational readouts. We next ask whether the identified routes predict what happens when the source activation is changed.

6.3 THE SELECTED ROUTES PREDICT INTERVENTION EFFECTS

We first test local calibration. Across 50 unseen multihop concepts per model, both signs, and doses from 1–10% of source-state norm, CCJ-J predictions track observed concept-logit changes (Figure 4(a)). Pearson r is 0.883 on Dream and 0.990 on LLaDA; origin slopes are 0.711 and 0.931, with sign agreement 98.3% and 100%. Across both models, CCJ-J provides the most consistent joint correlation and magnitude calibration. CCJ-R is stronger on Dream ($r = 0.958$, slope 0.799) but substantially weaker on LLaDA ($r = 0.560$, slope 0.193). At 10%, CCJ-J changes the model's top token in 32% of Dream and 22% of LLaDA interventions.

Because this agreement is measured with the derivative that defines CCJ-J, we next remove the derivative from the intervention itself. We construct two-fact prompts from a country/function grid.

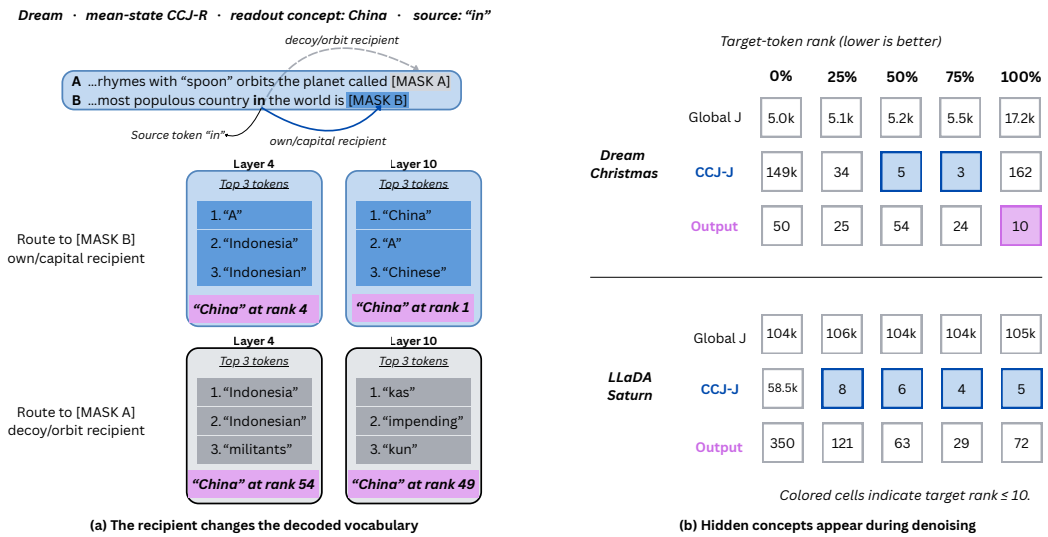

Figure 3: **Route-conditioned vocabulary readouts. (a)** A qualitative vocabulary trace from one fixed Dream forward pass. With the source token and layer held fixed, mean-state CCJ-R ranks *China* fourth/first when routed to its own capital recipient, but 54th/49th when routed to the unrelated orbit recipient. Cells list the three highest-scoring tokens. **(b)** Target ranks for two pre-indexed held-out denoising trajectories; colored cells denote the top 10. CCJ-J reveals *Christmas* at 50% restoration while the output and GLOBAL-J rank it 54th and 5,156th, and reveals *Saturn* at 25% while they rank it 121st and 106,492nd. Figure 2 quantifies the aggregate effect.

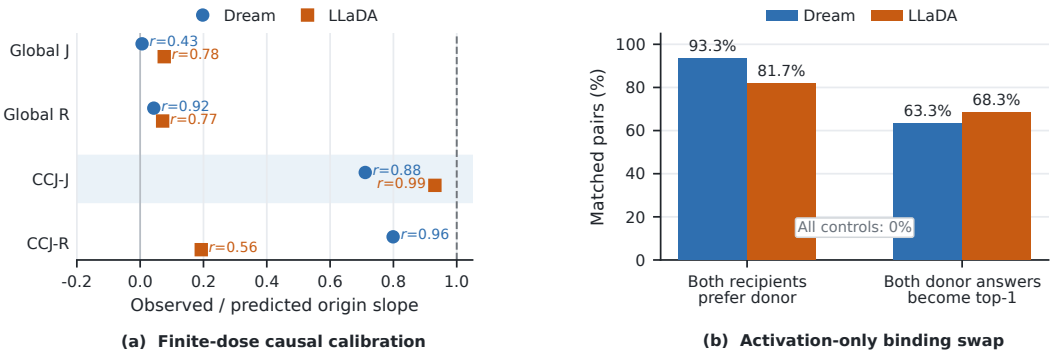

Figure 4: **Two complementary causal tests. (a)** For finite writes, CCJ-J remains strongly correlated and reasonably scaled across both models; CCJ-R is strongest on Dream but degrades on LLaDA; correlations and sign agreement are reported in text. **(b)** After CCJ-J selects a layer without answer labels, swapping actual argument residuals causes the two masked recipients to prefer the corresponding donor answers.

CCJ-J uses no answer labels to select one shared layer; the intervention then swaps the actual residual states at the two country arguments. Both recipients prefer their corresponding donor answers in 93.3% of Dream pairs and 81.7% of LLaDA pairs; both donors become top-1 in 63.3% and 68.3% (Figure 4(b)). The mean swap-margin changes are $+8.684$ and $+9.109$. Random-site, decoy-site, unrelated-donor, and norm-matched random controls never produce a both-recipient donor-preference success.

Together, the two tests close different gaps. Finite writes show that the local map predicts the sign and scale of small changes; activation-only swaps show that a layer selected by that map contains a manipulable binding even when the causal write is an ordinary model activation.

6.4 What is general, and what is specific to diffusion?

The same target-blind protocol also improves over global fields on LLaMA-3.1-8B-Instruct. Mean multihop MRR is 0.0256 for CCJ-J versus 0.0006 for Global J, and 0.0492 for CCJ-R versus 0.0006 for Global R; association gains are also positive across split sensitivity. What diffusion adds is the ability to keep a recipient unresolved and watch its incoming evidence develop.

That resolution has a computational cost. Exact CCJ builds one route-specific map per (X, p, q) tuple and takes 8.25–11.75 seconds per example with 4.55–6.21 GB additional peak memory in the controlled profiling setup. Global J/R shift this cost offline: fitting one complete reusable all-layer field on the $1{,}000$-document Dolmino corpus required 9.8–13.7 GPU-hours, after which applying the fitted field took approximately 0.04–0.05 seconds per example. CCJ requires no fitting corpus but recomputes transport for each route. This trade-off provides a concrete target for structured and low-rank approximations. Appendix F reports the corresponding sketches and scope analyses.

7 Limitations and Conclusion

CCJ gains resolution by giving up amortization: each route-specific $d \times d$ operator requires hundreds of reverse-mode batches and more memory and latency than a global field. Exact CCJ is therefore a high-fidelity analysis tool; structured, low-rank, and matrix-free approximations are natural next steps. Evidence spans two masked diffusion models and one autoregressive comparison; the vocabulary metrics focus on single-token concepts, and the causal tests use country/function grids with predefined spans. Broader relations and multi-token concepts remain open.

Because a bidirectional association reporter can alter earlier states, that task measures query-conditioned construction and routing rather than passive extraction. Residual swaps may transfer attributes beyond entity identity, although donor, site, decoy, and norm controls isolate the two-way answer effect. Dream's association result for CCJ-R also depends on norm-based routing and weakens under the fixed-last control. The conditional-risk result concerns operator approximation, not token rank. All-layer coverage rules out sparse depth search, but some selected layers remain split-sensitive. Broader model families, tokenizers, and depths are important tests of generality.

Within this scope, CCJ exposes one source-to-recipient route: it recovers concepts missed by global fields, follows recipients through denoising, predicts intervention and activation-swap effects, and extends to LLaMA. Diffusion additionally shows how route evidence develops over time.

AI Use Statement

Generative AI tools, were used to provide coding assistance, including implementation support and debugging, and to improve the clarity and presentation of the manuscript. This included assistance with grammar, phrasing, sentence structure, and the organization and framing of written material. The authors reviewed and verified all AI-assisted outputs and take full responsibility for the accuracy, content, claims, and final presentation of the work.

References

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens, 2025. URL https://arxiv.org/abs/2303.08112.

Camila Blank, Agam Bhatia, and Neel Nanda. R-lens: Making j-lens more faithful on early layers. LessWrong, 2026. URL https://www.lesswrong.com/posts/nv8oedrnLXKRzNEL9/r-lens-making-j-lens-more-faithful-on-early-layers. AI Alignment Forum.

Aaron Grattafiori et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models, 2026. URL https://arxiv.org/abs/2607.15495.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32819–32848. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/lou24a.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models, 2025. URL https://arxiv.org/abs/2502.09992.

nostalgebraist. interpreting gpt: the logit lens. LessWrong, August 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2026. URL https://arxiv.org/abs/2512.13961.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 130136–130184. Curran Associates, Inc., 2024. doi: 10.52202/079017-4135. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/eb0b13cc515724ab8015bc978fdde0ad-Paper-Conference.pdf.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small, 2022. URL https://arxiv.org/abs/2211.00593.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models, 2025. URL https://arxiv.org/abs/2508.15487.

A Additional Experimental Details

Target-blind ranking. For each split, normalized concept strings are hashed into 40% development and 60% test sets. Repeated concepts cannot cross the boundary. Each method selects its layer by development MRR; test targets never affect layer or source selection. Source positions maximize transported-vector norm, with the fixed-last eligible token as a pre-specified control. The primary Dream/LLaDA comparison evaluates all 26/30 eligible source layers. The preregistered grids $4, 10, 16, 24$ for Dream and $4, 12, 20, 28$ for LLaDA remain a four-layer sensitivity analysis; the LLaMA control uses the latter grid.

Denoising trajectories. We evaluate 0, 25, 50, 75, and 100% context restoration with three mask seeds while keeping the recipient masked. The primary multihop analysis has 189 observations and 39 normalized concepts for each model and fill mode. Association has 174 observations/58 concepts for Dream and 162/54 for LLaDA. Oracle reveal restores the gold context token; model fill commits the model's prediction. Hidden reveal records cases where the lens exposes a concept that remains weak in the current output distribution.

Finite writes and swaps. Finite-dose calibration uses 50 unseen concepts per model, both intervention signs, and doses of 1, 2.5, 5, and 10% of source-state norm. Binding swaps use ten countries, four functions (capital, language, continent, currency), and 60 matched pairs per model. The same selected layer is used for both argument sites.

Readout and CCJ-R implementation. For every lens, N is the model's native final normalization, applied independently to the transported d-vector with its frozen learned parameters immediately before the model's own unembedding. CCJ-R leaves the forward computation unchanged. In the backward pass, it substitutes the activation-aware R-lens rules only at RMSNorm and gated SiLU paths; all other modules retain their ordinary derivatives. These rules are fixed across models and are never selected using evaluation targets.

B Conditional-Risk Proof

Let $\mu_{\mathcal{G}} = \mathbb{E}[C \mid \mathcal{G}]$ and $\mu_{\mathcal{H}} = \mathbb{E}[C \mid \mathcal{H}]$ for $\mathcal{G} \subseteq \mathcal{H}$. Write

$$C - \mu_{\mathcal{G}} = (C - \mu_{\mathcal{H}}) + (\mu_{\mathcal{H}} - \mu_{\mathcal{G}}).$$

The cross term vanishes because the second factor is $\mathcal{H}$-measurable and $\mathbb{E}[C - \mu_{\mathcal{H}} \mid \mathcal{H}] = 0$. Expanding the squared Frobenius norm and taking expectations gives

$$\mathcal{R}(\mathcal{G}) - \mathcal{R}(\mathcal{H}) = \mathbb{E}\|\mu_{\mathcal{H}} - \mu_{\mathcal{G}}\|_F^2.$$

For the nested information sets revealing recipient Q, then source P, then prompt state X, repeated application yields

$$\mathcal{R}(\mathcal{G}_0) = \mathcal{H}_Q + \mathcal{H}_{P|Q} + \mathcal{H}_{X|P,Q} + \mathcal{R}(\mathcal{G}_{QPX}).$$

For a deterministic network, C is fixed by (X, P, Q), so the final term is zero. For any state h and linear vocabulary map U,

$$\|U(C - \widehat{C})h\|_2 \le \|U\|_2 \|C - \widehat{C}\|_F \|h\|_2,$$

but this score bound does not imply a monotone token-rank guarantee.

C Preregistered Four-Layer Sensitivity

Before expanding every method across depth, we preregistered the same four layers for each method. We retain this analysis as a sensitivity check and repeat concept-disjoint development/test selection over 100 deterministic splits. It includes both multihop and association.

Table 2: Preregistered four-layer target-blind MRR (mean over 100 concept-disjoint splits).

Model	Task	Logit	Global J	Global R	CCJ-J	CCJ-R
Dream	Multihop	.0046	.0002	.0038	.0228	.0904
Dream	Association	.0012	.0005	.0002	.0117	.0639
LLaDA	Multihop	.0048	.0095	.0032	.0600	.0368
LLaDA	Association	.0001	.0004	.0007	.0060	.0110

Table 3: All-layer association MRR with norm-based source selection.

	Dream			LLaDA		
Method	Oracle	40/60	Repeated	Oracle	40/60	Repeated
GLOBAL-J	.00318	.00067	.00061	.02516	.00015	.00066
CCJ-J	**.14990**	.00201	.00964	**.07306**	.00484	.00650
GLOBAL-R	.00136	.00026	.00019	.02046	.00047	.00290
CCJ-R	**.19612**	**.08833**	**.06276**	**.09116**	**.00763**	**.00898**

D ALL-LAYER ASSOCIATION RESULTS

Association provides a harder, non-arithmetic test of the same target-independent routing rule. Every method searches all eligible layers. The oracle columns remain target-aware diagnostics; the repeated columns are means over 100 concept-disjoint development/test splits.

On the fixed split, Dream CCJ-R exceeds GLOBAL-R by $+0.08808$, 95% CI $[0.03747, 0.15130]$, and CCJ-J by $+0.08633$ $[0.03631, 0.14681]$. LLaDA CCJ-R exceeds GLOBAL-R by $+0.00716$ $[0.00154, 0.01468]$, while its difference from CCJ-J is inconclusive. The Dream effect is specific to norm-based routing: the fixed-last CCJ-R MRR is 0.00317.

E ALL-LAYER RANKING DETAILS

Global J and Global R are fitted independently on 1,000 deterministic length-128 documents sampled from 1,000,000 local Dolma 3 Dolmino records across all 100 shards. We mask 50% of positions after the first 16 with seed 20260915. The corpus SHA-256 is `e6d8a4c106292e0865f05cd6f3dec250c33e78f75a8da9e7d81ad6eecfabbf58`. Every method evaluates the same records and all 26 Dream or 30 LLaDA source layers, using final transport endpoints 27/31 and VJP batch size 32. Oracle scores choose the best layer after observing each target. Test scores freeze one layer selected on the 40% development concepts. The frozen norm-selected layers (Dream/LLaDA) are 25/28 for GLOBAL-J, 19/28 for CCJ-J, 3/28 for GLOBAL-R, and 23/0 for CCJ-R. Under the fixed-last source rule they are 25/29, 14/28, 25/29, and 17/21, respectively.

Table 1 reports the absolute oracle and target-blind multihop MRR values. Table 4 gives the corresponding P@10 values. Table 5 reports the fixed-last source control.

Table 4: All-layer multihop comparison: absolute P@10 under norm-based source selection.

		Dream		LLaDA	
Method	Layers D/L	Oracle	40/60	Oracle	40/60
GLOBAL-J	26/30	4.8%	0.0%	4.8%	1.6%
CCJ-J	26/30	**66.9%**	**9.7%**	**43.5%**	**12.9%**
GLOBAL-R	26/30	1.6%	0.0%	3.2%	1.6%
CCJ-R	26/30	**66.1%**	**14.5%**	**47.6%**	**13.7%**

The full analysis contains 94 multihop concept observations in 84 valid records per model. The fixed test partition contains 63 observations grouped into 62 item clusters. Metrics first average observations within each item and then average across items, so a two-concept item may contribute

Table 5: Fixed-last source control: held-out MRR.

Method	Dream multihop	LLaDA multihop	Dream assoc.	LLaDA assoc.
GLOBAL-J	.02351	.00870	.00144	.00012
CCJ-J	.04394	.06040	.02503	.01781
GLOBAL-R	.02991	.01067	.00270	.00060
CCJ-R	.03523	.07433	.00317	.00788

Table 6: Paired norm-selected CCJ-minus-global multihop MRR differences.

Model	Pair	Difference [95% CI]
Dream	CCJ-J minus GLOBAL-J	+.0775 [+.0203,+.1467]
Dream	CCJ-R minus GLOBAL-R	+.0876 [+.0319,+.1540]
LLaDA	CCJ-J minus GLOBAL-J	+.0635 [+.0190,+.1177]
LLaDA	CCJ-R minus GLOBAL-R	+.0765 [+.0260,+.1398]

0.5 to P@10. Association contains 58 held-out Dream and 54 held-out LLaDA observations. Paired intervals use 5,000 item-clustered bootstrap replicates.

Prompt–concept grouped sensitivity. As a stricter independence check, we join all concepts from the same prompt and every repeated normalized concept into connected components. Components, rather than individual concepts, are assigned to development or test, and layer selection is repeated using development components only.

Table 7: Prompt–concept grouped sensitivity. Entries are primary-split ΔMRR with component-clustered paired 95% bootstrap intervals. Components denotes the total number of connected components.

Model	Suite	Components	CCJ-J − GLOBAL-J	CCJ-R − GLOBAL-R
Dream	Multihop	59	+.0502 [.0059,.1159]	+.0731 [-.0228,.1798]
LLaDA	Multihop	59	+.0679 [.0151,.1389]	+.0744 [.0264,.1390]
LLaMA	Multihop	67	+.0223 [.0094,.0395]	+.0651 [.0298,.1104]
Dream	Association	99	+.0105 [.0011,.0232]	+.0649 [.0274,.1117]
LLaDA	Association	94	-.0032 [-.0092,.0000]	+.0056 [.0004,.0138]
LLaMA	Association	96	+.0102 [.0003,.0282]	+.0423 [.0111,.0844]

Concept-clustered bootstraps give the same inferential conclusions. Across 100 component-disjoint splits, the positive-split fraction is reported only as a stability diagnostic, not as a confidence interval or p-value. The primary CCJ-J multihop advantage remains positive for all three models; Dream CCJ-R multihop and LLaDA association CCJ-J remain inconclusive on the primary split.

F ADDITIONAL CONTROLS AND SCOPE ANALYSES

Association source clamp. Clamping source states to their pre-reporter values changes CCJ-J MRR from .0095 to .0002 on Dream and .0096 to .0014 on LLaDA. CCJ-R changes from .0779 to .0328 on Dream and .0026 to .0020 on LLaDA. The appended query is therefore part of the computation being interpreted in a bidirectional model.

Naturalistic transplants. Cross-prompt residual transplants produce a pooled swap-margin change of $+8.965$, with 82.1% donor preference and 38.1% donor top-1. A wrong-relation control is similarly strong, locating the dominant effect in entity-identity transfer and motivating relation-specific transplant designs.

Route heterogeneity. Structured recipient/source variation exceeds shuffled and norm-matched controls. The preregistered mean-state scalar has Pearson $r = -0.3105$ with ranking gain, 95%

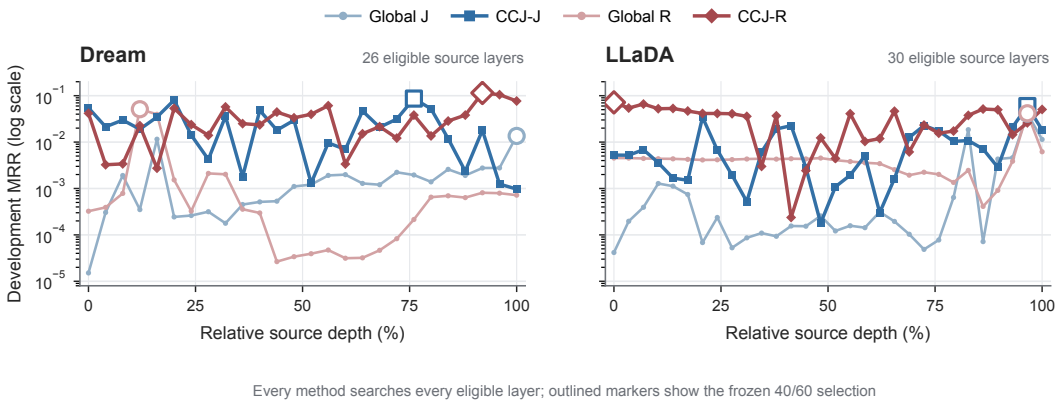

Figure 5: **All-layer target-blind development profiles.** Every contextual and global method is evaluated at each eligible source layer. Outlined markers identify the layer frozen by the primary 40/60 development split. Test and oracle claims rely on Tables 1, 3, and 4.

clustered interval $[-0.3670, -0.2365]$, showing that the empirical advantage depends on richer route structure than this single summary captures.

Efficiency. Exact CCJ costs 8.25–11.75 seconds per example with 4.55–6.21 GB additional peak memory in the controlled profiling setup. Global J/R instead incur a substantial one-time fitting cost: the complete Dream Global-J and Global-R fits required 10.44 and 9.77 GPU-hours, respectively, while the corresponding LLaDA fits required 13.74 and approximately 12.64 GPU-hours, respectively, on the $1{,}000$-document Dolmino corpus. Once fitted, applying a global field takes roughly 0.04–0.05 seconds per example. Rank-512 contextual sketches preserve useful vector geometry and partial ranking or calibration signal. These measurements define concrete fidelity and runtime targets for future structured approximations.