FUNCTIONAL SPECIALIZATION AND SELF-PERSISTENT ROUTING OF NATIVE TOKENS IN VITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision Transformers (ViTs) increasingly employ auxiliary tokens such as [CLS] and register ([REG]) tokens, yet their computational roles remain poorly understood. We analyze register-token dynamics using dual-path activation patching and introduce two complementary measures: Self-Autonomy, capturing causal persistence of token states, and Patch Receptivity, capturing their recoverability from spatial patch information. Across four DINOv2 scales and nine controlled visual conditions, register streams progressively specialize with depth and exhibit a clear Autonomy-Receptivity duality: shallow states are strongly patch-driven, whereas deeper states become increasingly self-persistent. These specialization patterns remain consistent across visual conditions despite changes in routing strength. Test-time register early-exit interventions further show that shallow register states can exert substantial downstream influence despite low Self-Autonomy. Our results reveal register tokens as heterogeneous, depth-dependent information-routing streams whose functional roles emerge and specialize throughout the network.

1 INTRODUCTION

Vision Transformers (ViTs) (Dosovitskiy et al., 2021; Touvron et al., 2021a) have become a foundational backbone in computer vision, demonstrating strong visual representation capabilities across both supervised and self-supervised learning regimes (Caron et al., 2021; Radford et al., 2021; Oquab et al., 2024). A central component of ViTs is cross-token attention, through which information is exchanged between spatial patch tokens and auxiliary tokens. In particular, the [CLS] token has become an important object of analysis for understanding emergent representations, constructing attribution maps, and interpreting model predictions (Caron et al., 2021; Chefer et al., 2021; Shi et al., 2023). However, recent studies have shown that pretrained ViTs can spontaneously develop severe attention artifacts associated with high-norm outlier tokens at spatial locations containing little semantic information (Darcet et al., 2024; Oquab et al., 2024). Such artifacts can distort attention-based interpretations and adversely affect dense visual predictions.

The emergence of high-norm tokens is related to a broader phenomenon observed across Transformer architectures. In natural language processing, Transformer models can develop unusually large activations associated with particular tokens, often discussed in connection with attention sinks (Kobayashi et al., 2020; Chefer et al., 2021; Xiao et al., 2024; Sun et al., 2024). These tokens can absorb disproportionate amounts of attention and consequently influence information flow throughout the network (Yona et al., 2025). In vision models, Darcet et al. (2024) introduced additional learnable tokens, termed *registers*, to provide dedicated capacity for absorbing large vector norms that would otherwise emerge in spatial patch tokens. Subsequent studies have explored lightweight fine-tuning (Wang et al., 2024), inference-time attention modification (Nakamura et al., 2024), and feature denoising (Yang et al., 2024) to further mitigate related artifacts. Consequently, register tokens are commonly viewed as auxiliary buffers that absorb non-informative spatial signals.

Despite their growing adoption, this view leaves the internal function of register tokens largely unexplained. Register tokens are introduced as a homogeneous set without predefined semantic roles, yet their hidden states repeatedly interact with spatial patch tokens as information propagates through the network. It therefore remains unclear whether individual register streams develop distinct causal roles, how strongly their intermediate states persist toward their final representations, and to what extent those final representations remain recoverable from spatial information at different depths. More

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.

fundamentally, little is known about how the functional role of register streams changes throughout ViT computation. These questions cannot be fully resolved through attention visualization alone. Although attention maps reveal where tokens allocate attention, they do not establish whether an intermediate representation is causally responsible for a downstream state (Wiegreffe & Pinter, 2019; Vig et al., 2020). A mechanistic characterization therefore requires controlled interventions that selectively restore intermediate representations and trace their downstream consequences.

In this work, we present a fine-grained mechanistic dissection of ViT bottleneck streams using *dual-path activation patching*. Starting from matched clean and corrupted trajectories, we construct two complementary causal interventions at each network depth. The first restores the clean state of an individual bottleneck token into an otherwise corrupted trajectory and measures how strongly that state recovers its own final representation. The second restores the clean spatial patch stream while keeping the bottleneck states corrupted and measures how strongly the final bottleneck representation can be recovered through spatial information.

These two intervention paths give rise to two complementary causal measures. *Self-Autonomy* quantifies the causal persistence of an intermediate bottleneck state toward its final representation. *Patch Receptivity*, operationalized through end-to-end spatial recovery, quantifies the extent to which the final bottleneck representation can be causally recovered by restoring clean spatial patch information at an intermediate layer. Together, the two measurements characterize two distinct routes supporting the final bottleneck representation: persistence through the bottleneck state itself and recovery through the spatial stream. We evaluate these causal properties across four DINOv2 model scales (ViT-S/B/L/G) under nine controlled scene-corruption conditions, combining three levels of scene complexity with three background corruption types. Across these settings, our analysis reveals three consistent properties of ViT bottleneck dynamics.

Emergent and Condition-Robust Functional Specialization: Although register tokens have no predefined functional ordering, individual streams progressively develop distinct causal routing profiles. Register-wise heterogeneity increases systematically with depth, while the relative ordering of register streams remains highly stable across visual conditions within each model. This reveals specialization as an emergent property of the register population rather than an intrinsic role attached to a particular register index.

Autonomy-Receptivity Duality: Self-Autonomy and Patch Receptivity exhibit a pronounced inverse relationship across depth. When an intermediate bottleneck state has weak causal persistence, its final representation remains strongly recoverable through restored spatial information. As Self-Autonomy increases, Patch Receptivity decreases, revealing complementary self-persistent and spatially recoverable causal pathways.

Depth-Dependent and Scale-Dependent Causal Routing: The balance between these two pathways changes systematically throughout the network. Early bottleneck representations are predominantly supported by information residing in the spatial stream, whereas deeper states become increasingly self-persistent. Importantly, the crossover between these regimes is neither universal nor tied to a fixed layer: it is strongly dependent on model scale and register stream, occurring sparsely before the terminal depth in smaller models but becoming nearly universal in ViT-L and ViT-G. Visual conditions modulate the strength of routing without substantially reshuffling the learned register organization.

Finally, we ask whether these causal dynamics have observable functional consequences. As a lightweight validation, we perform a *Test-Time Register Early Exit* experiment in which register tokens are initialized and processed normally, but removed after a prescribed fraction of Transformer blocks while both the backbone and downstream classifier remain fixed. The results clarify an important distinction: low early Self-Autonomy does not imply functional irrelevance. Shallow register states are weakly self-persistent yet already exert substantial downstream influence, while most task-relevant contribution is progressively accumulated with depth. Together with the causal analysis, this suggests a depth-dependent functional transformation from transient but influential interaction states to increasingly self-persistent and specialized streams.

In summary, our main contributions are as follows:
1) **Dual-Path Causal Intervention Framework:** We introduce dual-path activation patching for mechanistically analyzing ViT bottleneck streams, together with Self-Autonomy for causal persistence and Patch Receptivity for end-to-end recoverability through restored spatial information.

2) **Emergent and Robust Register Specialization:** We show that initially similar register streams progressively develop distinct causal routing profiles, while their relative organization remains highly stable across visual conditions.
3) **Autonomy-Receptivity Duality and Depth-Dependent Routing:** We uncover a pronounced inverse relationship between Self-Autonomy and Patch Receptivity and show that bottleneck representations evolve from spatially recoverable toward increasingly self-persistent states through a scale- and stream-dependent causal transition.
4) **Functional Validation via Register Early Exit:** As a lightweight validation, we show that shallow registers can exert substantial downstream influence despite low Self-Autonomy, supporting a depth-dependent transformation from transient interaction to self-persistent specialization.

2 Related Work

Feature Visualization and ViT Artifacts: Visualization has long been used to interpret neural representations (Bau et al., 2017). In ViTs (Dosovitskiy et al., 2021; Touvron et al., 2021b; 2022), [CLS] attention maps are widely used for attribution and representation analysis (Caron et al., 2021; Chefer et al., 2021; Shi et al., 2023). However, pretrained ViTs can develop high-norm spatial outliers and attention artifacts (Darcet et al., 2024). Subsequent work has studied these effects through singular-feature repair, attention modification, denoising, and alternative artifact mechanisms (Wang et al., 2024; Nakamura et al., 2024; Yang et al., 2024; Shi et al., 2026).

Register Tokens and Their Functions: Registers were introduced as learnable auxiliary tokens that absorb large activations otherwise assigned to spatial patches (Darcet et al., 2024). Recent studies show that their roles are more complex. Register and [CLS] tokens can decouple local and global representations (Lappe & Giese, 2025); register-like behavior can arise without trained register embeddings (Jiang et al., 2025); and register representations can encode structured global concepts (Fel et al., 2026). Other work examines attention-sink mechanisms, cross-architecture generality, and alternative register construction (Fesser et al., 2026; Bach et al., 2025; Baxevanakis et al., 2026; Yan et al., 2025). Importantly, Parodi et al. (2026) show that zero-ablation can overstate dependence on exact register content. These works primarily characterize register semantics, aggregate effects, or ablation sensitivity, whereas we causally track individual register streams across depth.

Mechanistic Interpretability via Causal Interventions: Recent work increasingly complements visualization with mechanistic analysis, including semantic feature decomposition and sparse-autoencoder-based interpretation (Gandelsman et al., 2025; Robinson et al., 2024; Stevens et al., 2025; Joseph et al., 2025; Kim et al., 2025). Because attention alone does not establish causal responsibility (Wiegreffe & Pinter, 2019), causal interventions such as activation patching and mediation analysis provide a more direct approach (Vig et al., 2020). Our dual-path activation patching framework extends this perspective to ViT bottleneck streams: Self-Autonomy measures state persistence, while Patch Receptivity measures recoverability through restored spatial information. This enables stream-specific analysis of causal routing and explicitly separates state persistence from broader functional influence; register early exit is used only as a lightweight functional validation.

3 Quantifying Bottleneck Dynamics via Dual-Path Activation Patching

We introduce a dual-path activation patching framework to causally characterize ViT bottleneck streams. Given matched clean and corrupted trajectories, we selectively restore either (i) an individual bottleneck state or (ii) the spatial patch stream at an intermediate layer, and measure recovery of the corresponding final bottleneck representation. These interventions define complementary quantities: Self-Autonomy, which measures persistence through the bottleneck state itself, and Patch Receptivity, which measures recoverability through spatial information. Importantly, these quantities characterize the causal support of a bottleneck representation; they do not directly measure the total functional influence of a token on other streams. In particular, low Self-Autonomy indicates that an intermediate bottleneck state does not itself strongly persist to its final state, but does not imply that the token is functionally irrelevant to the subsequent computation.

3.1 Notation and Activation Patching

Let $f = f_L \circ \cdots \circ f_1$ denote an L-block ViT. At layer l, the hidden sequence is

$$X_l = [\mathbf{h}_l^{(\mathrm{CLS})}, \mathbf{h}_l^{(r_1)}, \ldots, \mathbf{h}_l^{(r_K)}, \mathbf{h}_l^{(p_1)}, \ldots, \mathbf{h}_l^{(p_N)}]^\top, \tag{1}$$

where $\mathcal{T} = \{\mathrm{CLS}, r_1, \ldots, r_K\}$ denotes the bottleneck tokens and $\mathcal{P} = \{p_1, \ldots, p_N\}$ the spatial patches. A clean image I and its corrupted counterpart I' produce X_l and $\widetilde{X}_l$, respectively.

For a set of token indices $\mathcal{I}$, we define the activation-patching operator

$$\mathrm{Patch}(X_l, \mathcal{I}, V)_i = \begin{cases} \mathbf{v}^{(i)}, & i \in \mathcal{I}, \\ X_{l,i}, & \text{otherwise.} \end{cases} \tag{2}$$

which replaces only the selected token states while leaving the remainder of the trajectory unchanged.

3.2 Self-Autonomy

Self-Autonomy measures how strongly an intermediate bottleneck state causally persists to its own final representation. For token $t \in \mathcal{T}$, we restore its clean state into the corrupted trajectory at layer l:

$$X_l^{(\mathrm{auto}:t)} = \mathrm{Patch}(\widetilde{X}_l, \{t\}, \{\mathbf{h}_l^{(t)}\}), \tag{3}$$

and propagate it to the final layer, obtaining $\mathbf{h}_{L,\mathrm{auto}}^{(t,l)}$.

Using cosine similarity relative to the clean and unpatched corrupted final representations:

$$\mathcal{S}_{\mathrm{corrupt}}^{(t)} = \mathrm{CosSim}\left(\widetilde{\mathbf{h}}_L^{(t)}, \mathbf{h}_{L,\mathrm{clean}}^{(t)}\right), \tag{4}$$

$$\mathcal{S}_{\mathrm{auto}}^{(t)}(l) = \mathrm{CosSim}\left(\mathbf{h}_{L,\mathrm{auto}}^{(t,l)}, \mathbf{h}_{L,\mathrm{clean}}^{(t)}\right), \tag{5}$$

we define

$$\mathcal{A}_l^{(t)} = \frac{\mathcal{S}_{\mathrm{auto}}^{(t)}(l) - \mathcal{S}_{\mathrm{corrupt}}^{(t)}}{\max\left(1 - \mathcal{S}_{\mathrm{corrupt}}^{(t)}, \epsilon\right)} \times 100\%, \tag{6}$$

where $\epsilon = 10^{-6}$. A high $\mathcal{A}_l^{(t)}$ indicates that restoring the intermediate bottleneck state substantially recovers its final clean representation, revealing strong causal persistence along that bottleneck stream. Conversely, a low value indicates weak *state persistence*; it should not be interpreted as evidence that the token has little influence on other token streams. Because $\mathcal{A}_l^{(t)}$ is a normalized recovery score, negative values are possible.

3.3 Patch Receptivity via End-to-End Spatial Recovery

Patch Receptivity measures how strongly the final bottleneck representation remains recoverable from spatial information available at layer l. Starting from the corrupted trajectory, we restore all clean patch states while leaving the bottleneck states corrupted:

$$X_l^{(\mathrm{rec})} = \mathrm{Patch}(\widetilde{X}_l, \mathcal{P}, \mathcal{P}_l), \qquad \mathcal{P}_l = \{\mathbf{h}_l^{(p)} \mid p \in \mathcal{P}\}. \tag{7}$$

After propagation to the final layer, let $\mathbf{h}_{L,\mathrm{rec}}^{(t,l)}$ denote the resulting representation of bottleneck token t. We define

$$\mathcal{S}_{\mathrm{rec}}^{(t)}(l) = \mathrm{CosSim}(\mathbf{h}_{L,\mathrm{rec}}^{(t,l)}, \mathbf{h}_{L,\mathrm{clean}}^{(t)}), \;\; \mathcal{R}_l^{(t)} = \frac{\mathcal{S}_{\mathrm{rec}}^{(t)}(l) - \mathcal{S}_{\mathrm{corrupt}}^{(t)}}{\max(1 - \mathcal{S}_{\mathrm{corrupt}}^{(t)}, \epsilon)} \times 100\%. \tag{8}$$

A higher $\mathcal{R}_l^{(t)}$ indicates that restoring spatial information at layer l substantially recovers the final bottleneck representation. Thus, Patch Receptivity measures end-to-end dependence on information carried by the spatial patch stream rather than local attention or immediate token sensitivity.

3.4 Joint Autonomy-Receptivity Representation

For each bottleneck stream, we characterize its causal state as

$$\mathcal{D}^{(t)}(l) = \left(\mathcal{A}_l^{(t)}, \mathcal{R}_l^{(t)}\right). \tag{9}$$

The two coordinates describe complementary recovery pathways: Self-Autonomy measures recovery through the intermediate bottleneck state itself, whereas Patch Receptivity measures recovery through the spatial patch stream.

To quantify their relative balance, we further define

$$\Delta_l^{(t)} = \mathcal{A}_l^{(t)} - \mathcal{R}_l^{(t)}. \tag{10}$$

Negative $\Delta_l^{(t)}$ indicates that the final bottleneck representation is more strongly recoverable through the spatial stream, whereas positive $\Delta_l^{(t)}$ indicates stronger persistence through the bottleneck state itself. Tracking $\mathcal{D}^{(t)}(l)$ and $\Delta_l^{(t)}$ across depth therefore reveals how bottleneck routing shifts from predominantly spatially recoverable toward increasingly self-persistent support. This distinction is central to our analysis: the framework characterizes how bottleneck states are represented and recovered, while their broader functional influence on downstream computation is examined separately through the register early-exit validation in Sec. 5.

4 Mechanistic Dissection of Bottleneck Streams

We apply the proposed dual-path activation patching framework to examine how the causal support of ViT bottleneck representations evolves across network depth. Our analysis focuses on three complementary properties: (i) how the balance between spatial recoverability and self-persistence changes with depth, (ii) whether individual register streams progressively develop distinct causal profiles, and (iii) whether the resulting organization remains stable across visual conditions. Importantly, the quantities analyzed in this section characterize the persistence and recoverability of bottleneck *states*; they do not by themselves measure the total functional influence of a token on other streams.

4.1 Experimental Setting

Models and data: We evaluate four pretrained DINOv2 models equipped with four register tokens: ViT-S/14, ViT-B/14, ViT-L/14, and ViT-G/14, containing 12, 12, 24, and 40 Transformer blocks, respectively. Clean-corrupted image pairs are constructed from Pascal VOC 2012 using segmentation masks. We select 90 images in total and divide them into three scene groups of 30 images each: (i) simple single-object scenes, where the background occupies more than 50% of the image area (Group A); (ii) single-object-dominated scenes with complex backgrounds (Group B); and (iii) cluttered multi-object scenes (Group C). For each group, we apply three background corruptions: Gaussian noise, black replacement, and blur, yielding nine controlled scene-corruption conditions.

Unless otherwise stated, results are averaged over the nine conditions. For cross-scale comparison, network depth is normalized as $\rho = l/L$. The terminal intervention point is excluded because restoring the target bottleneck state at the final block yields $\mathcal{A}_L^{(t)} \approx 100\%$, whereas restoring spatial patches yields $\mathcal{R}_L^{(t)} \approx 0\%$ by construction.

4.2 Autonomy-Receptivity Duality

Figure 1 reveals a pronounced inverse relationship between Self-Autonomy and Patch Receptivity. In early layers, bottleneck states exhibit weak causal persistence, while their final representations remain highly recoverable by restoring spatial information. Toward deeper layers, Self-Autonomy increases and Patch Receptivity decreases, indicating a shift in causal support from the spatial patch stream toward the bottleneck states themselves.

This relationship is not explained solely by their shared dependence on depth. After controlling for normalized depth ρ, the partial correlations between $\mathcal{A}$ and $\mathcal{R}$ remain negative across all four scales (-0.962, -0.879, -0.926, and -0.690 for ViT-S/B/L/G, respectively). These results support an empirical *Autonomy-Receptivity duality*: final bottleneck representations can be supported through complementary spatial-recovery and self-persistence pathways. We emphasize that low Self-Autonomy indicates weak persistence of a particular intermediate state to its own final representation; it does not imply that the corresponding token has little influence on the surrounding token streams. We revisit this distinction through a functional intervention in Sec. 5.

4.3 Crossover from Spatial Recoverability to Self-Persistent Routing

We define the *crossover depth* $\rho_*^{(t)}$ as the first normalized depth at which $\Delta^{(t)}$ crosses zero, using linear interpolation between adjacent blocks. Trajectories that do not cross before the terminal block are treated as right-censored. Figure 2 shows that crossover behavior is strongly scale- and stream-dependent. Across the nine conditions, only $9/36$ (25.0%) and $10/36$ (27.8%) register trajectories

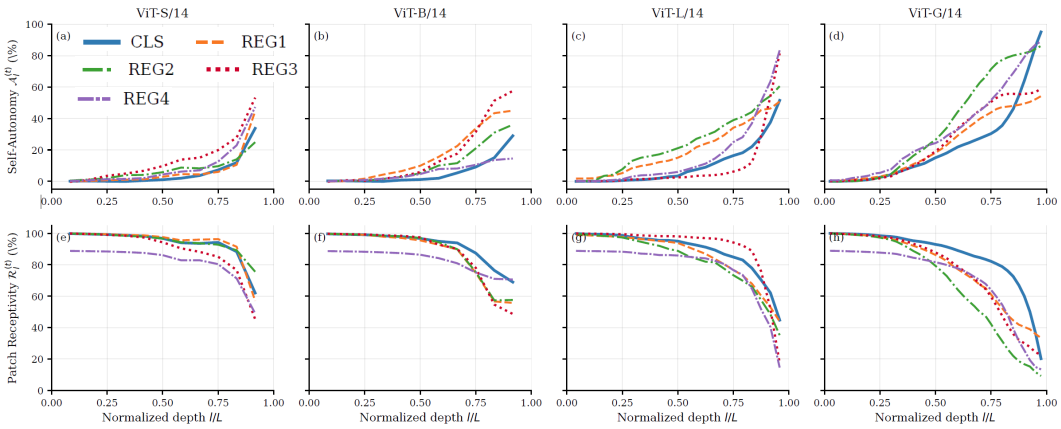


Figure 1: **Depth-dependent causal profiles of ViT bottleneck streams.** Self-Autonomy (top) and Patch Receptivity (bottom) are shown across normalized depth for DINOv2-S/B/L/G. Each curve denotes the nine-condition average for [CLS] or an individual register stream. Self-Autonomy generally increases while Patch Receptivity decreases, revealing a progressive shift from spatially supported recovery toward increasingly self-persistent routing. Register trajectories also diverge with depth, suggesting progressive functional specialization. The terminal intervention point is omitted.

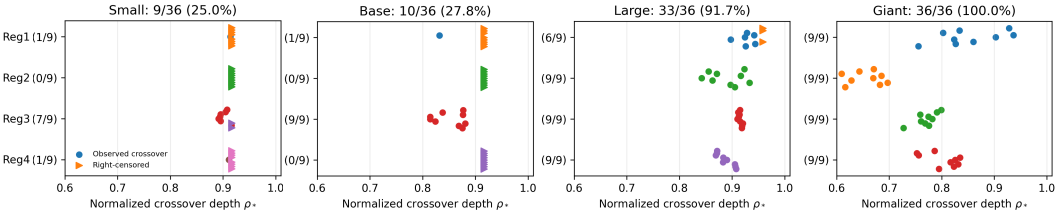


Figure 2: **Scale- and stream-dependent causal crossover.** Each point denotes the first non-terminal depth at which $\Delta_l^{(t)}$ crosses zero for one scene-corruption condition. Right-pointing markers denote trajectories that remain spatially recoverable up to the last non-trivial intervention depth. Crossover prevalence increases from $25.0\%/27.8\%$ in ViT-S/B to 91.7% in ViT-L and 100% in ViT-G, while transition depth remains strongly stream-dependent.

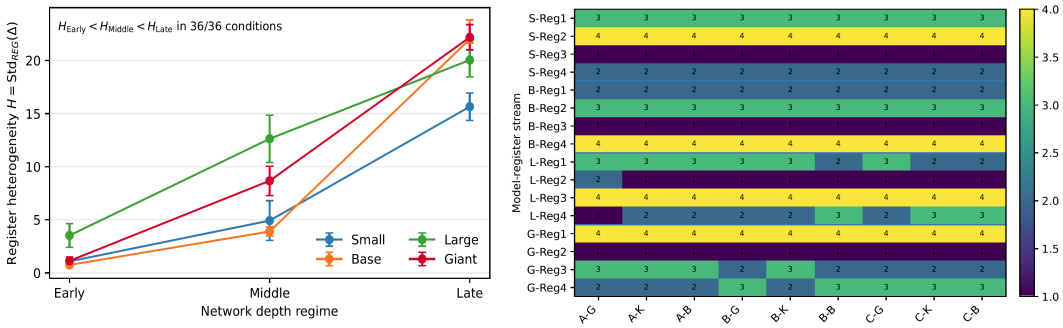
exhibit a non-trivial crossover in ViT-S and ViT-B, respectively. This fraction increases sharply to $33/36$ (91.7%) for ViT-L and $36/36$ (100%) for ViT-G. In ViT-S/B, crossover is concentrated primarily in REG_3; among trajectories exhibiting an observed crossover, the mean crossover depths are approximately 0.90 and 0.85, respectively. In contrast, every register stream in ViT-G crosses before the terminal layer, with REG_2 transitioning particularly early at $\rho_* \approx 0.66$. These results rule out a universal critical depth. Instead, the emergence of self-persistent routing depends jointly on model scale and individual register stream. Crossover prevalence provides the clearest cross-scale difference, increasing from sparse occurrence in ViT-S/B to near-universal or universal occurrence in ViT-L/G, while the crossover depth among observed crossers remains strongly stream-dependent.

4.4 Progressive Register Specialization

Functional specialization refers to stable differentiation in causal information-routing profiles across register streams, rather than to predefined semantic roles associated with particular register indices. Figure 1 also suggests that register streams become increasingly distinct with depth. We quantify this differentiation through the dispersion of routing balance across the four register tokens:

$$H_l = \mathrm{Std}_{t \in \mathrm{REG}} \left[\Delta_l^{(t)} \right]. \tag{11}$$

To summarize its depth dependence, we partition the non-terminal normalized depth into three equal regimes: Early : $\rho \in [0, \frac{1}{3})$, Middle : $\rho \in [\frac{1}{3}, \frac{2}{3})$, Late : $\rho \in [\frac{2}{3}, 1)$, where the terminal intervention point is excluded as described in Sec. 4.1. For each model-condition pair, H_l is averaged



(a) Progressive register heterogeneity (b) Condition-robust late-depth register ranking

Figure 3: **Emergent and robust functional specialization.** (a) The dispersion $H = \mathrm{Std}_{\mathrm{REG}}(\Delta)$ increases monotonically from early to late depth in all 36 model-condition combinations. (b) Late-depth register rankings remain highly consistent across the nine scene-corruption conditions (G: Gaussian, K: Black, B: Blur, Rank=1: most self-persistent), with Kendall's $W = 1.000$, 1.000, 0.857, and 0.901 for ViT-S/B/L/G.

over the layers belonging to each regime. As shown in Fig. 3(a), register heterogeneity increases strongly with depth. Averaged across models and conditions, H rises from 1.62 in early layers to 7.53 in middle layers and 19.97 in late layers. Remarkably,

$$H_{\mathrm{early}} < H_{\mathrm{middle}} < H_{\mathrm{late}} \tag{12}$$

holds for all 36 model-condition combinations (4 model scales $\times 9$ scene-corruption conditions), providing a highly consistent descriptive pattern of progressive register differentiation. As a secondary statistical summary, we apply a Friedman test to the 36 paired model-condition summaries across the three depth regimes, yielding $\chi^2(2) = 72.0$, $p = 2.32 \times 10^{-16}$, and Kendall's $W = 1.00$. Here, the statistical unit is a model–condition pair rather than an individual image; accordingly, this test quantifies the consistency of the depth-wise trend across experimental conditions rather than image-level sampling variability.

The same trend appears across model scales, with late-layer heterogeneity of 15.65, 22.00, 20.05, and 22.19 for ViT-S/B/L/G, respectively. Thus, register streams begin with relatively similar causal routing profiles but progressively separate into distinct regimes. This provides quantitative evidence that register specialization emerges through network depth rather than reflecting predefined differences among token indices.

4.5 Condition-Robust Functional Organization

We next ask whether the observed specialization persists across changes in scene structure and background corruption. For each of the nine conditions, we rank the four register streams according to their late-depth routing balance. Figure 3(b) shows remarkably stable rankings within each model. Kendall concordance coefficients are $W = 1.000$, 1.000, 0.857, and 0.901 for ViT-S/B/L/G, respectively ($p < 4 \times 10^{-5}$ in all cases).

The dominant late-Δ stream is REG_3 for ViT-S/B and REG_2 for ViT-L/G. Register specialization is therefore highly stable across visual conditions within a given model, while the identity of the dominant stream changes across model scales. This supports the view that specialization is an emergent property of the register population rather than an intrinsic semantic role associated with a fixed register index.

4.6 Effects of Scene Structure and Background Corruption

Although register ordering is highly stable, visual conditions can modulate the strength of information routing. Figure 4 separates the effects of background corruption and scene structure. Background corruption has little effect on late-depth register heterogeneity: $H_{\mathrm{late}} = 19.72$, 20.30, and 19.90, , but significantly changes the depth-averaged routing balance from -71.47 (Gaussian) to -68.46 (black) and -66.16 (blur). The Friedman test uses 4 models $\times$ 4 register streams $= 16$ matched units ($p = 7.97 \times 10^{-5}$, Kendall's $W = 0.590$), indicating a relative shift toward stronger self-persistent support. Scene structure more strongly affects late specialization:

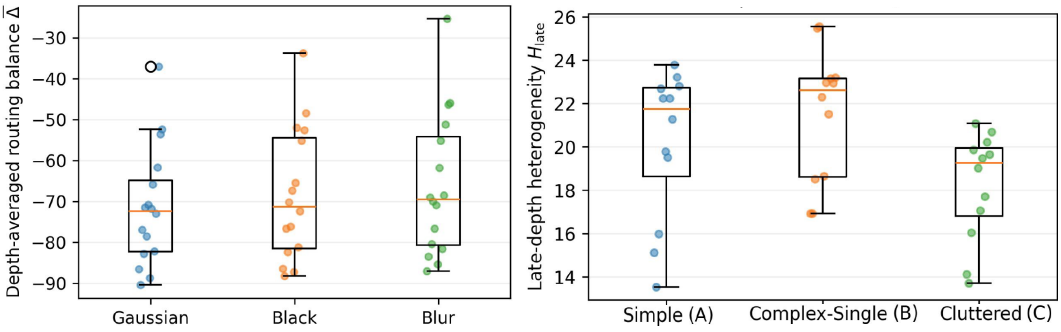


Figure 4: **Effects of scene structure and background corruption. Left:** Depth-averaged routing balance grouped by background corruption. Each point denotes one model-register summary after averaging over the three scene groups, yielding 4 model scales $\times$ 4 register streams $=$ 16 paired observations per corruption type. **Right:** Late-depth register heterogeneity grouped by scene structure. Each point denotes one model-corruption summary, yielding 4 model scales $\times$ 3 corruption types $=$ 12 paired observations per scene group.

Table 1: **Summary of crossover, register heterogeneity, and condition robustness.** Crossover denotes the fraction of register-condition trajectories crossing $\Delta = 0$ before the terminal intervention. H measures the dispersion of Δ across the four register streams, and Kendall's W measures agreement of late-depth register rankings across the nine conditions.

Model	Cross.	H_{E}	H_{M}	H_{L}	W	Dominant
ViT-S	25.0%	1.08	4.92	15.65	1.000	REG3 (9/9)
ViT-B	27.8%	0.74	3.90	22.00	1.000	REG3 (9/9)
ViT-L	91.7%	3.52	12.63	20.05	0.857	REG2 (8/9)
ViT-G	100%	1.13	8.67	22.19	0.901	REG2 (9/9)

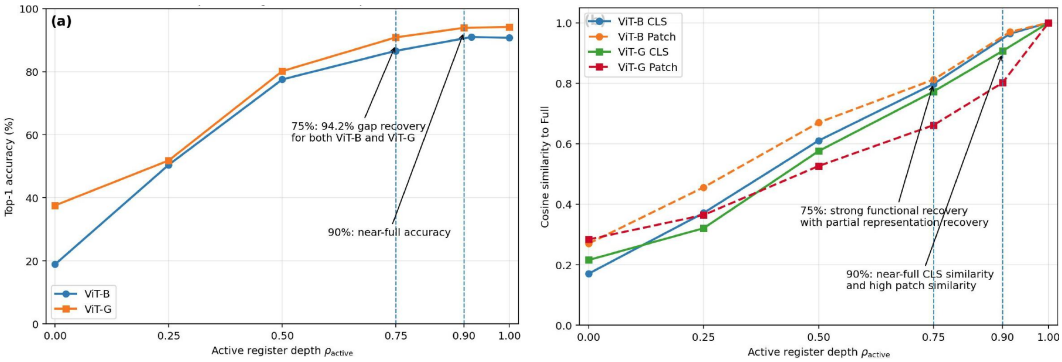
$H_{\text{late}} = 20.18, 21.51, 18.22$ for simple, complex-single, and cluttered scenes, respectively. Using 4 models $\times$ 3 corruption types $=$ 12 matched units, the effect is significant ($p = 5.09 \times 10^{-4}$, Kendall's $W = 0.632$). The lower heterogeneity in cluttered scenes indicates weaker separation among register routing profiles. Neither factor significantly shifts crossover depth when restricted to streams crossing under all nine conditions. Overall, visual conditions modulate routing strength and differentiation while largely preserving the relative organization of register streams.

4.7 From State Persistence to Functional Influence

The preceding analyses characterize how register states are causally supported, but they do not determine whether a weakly persistent state is functionally unimportant. In particular, the low Self-Autonomy observed in shallow layers means that those register states do not strongly persist to their own final representations; it does not exclude the possibility that they influence [CLS] or patch streams through cross-token interaction. This distinction motivates a complementary functional validation. Rather than inserting new register states at intermediate depth, which would alter the pretrained trajectory, we preserve the normal register computation from the input and remove registers only after a controlled fraction of the network. The following section uses this intervention to test how long register participation remains functionally important.

5 Test-Time Register Early Exit

To complement the causal-state analysis above, we perform a lightweight test-time register early-exit evaluation. Register tokens are initialized and processed normally from the input, but are removed after a prescribed fraction of Transformer blocks. The pretrained backbone remains frozen, and a linear classifier is trained only once using the full-register representation. During evaluation, both the backbone and classifier remain fixed; only the depth at which register tokens are removed is varied. We evaluate ViT-B/14 and ViT-G/14 on CIFAR-100 using register-active depths of 0%, 25%, 50%, 75%, 90%, and 100%. We report classification Top-1 accuracy together with cosine similarity of the final [CLS] and patch-token representations to those of the full-register model.



(a) CIFAR-100 Top-1 accuracy (b) Cosine similarity

Figure 5: **Functional validation through test-time register early exit.** (a) CIFAR-100 Top-1 accuracy as a function of the fraction of Transformer blocks for which register tokens remain active. Register participation is functionally important from early depth, but most of the full-register performance is recovered once registers remain active through 75% of the network. At approximately 90% active depth, both ViT-B and ViT-G recover near-full classification accuracy. (b) Cosine similarity of the final $[\mathrm{CLS}]$ and patch-token representations to the full-register model. Representation similarity increases monotonically with active depth, while task performance can approach the full model before exact representational equivalence is reached.

Figure 5(a) shows a consistent recovery pattern across ViT-B and ViT-G. Removing registers entirely reduces Top-1 accuracy from 90.72% to 18.91% for ViT-B and from 94.16% to 37.59% for ViT-G, confirming that register participation is functionally important. Allowing registers to remain active through the first 50% of the network recovers 81.6% and 75.2% of the full-versus-no-register performance gap for ViT-B and ViT-G, respectively. Extending register participation to 75% recovers approximately 94.2% of this gap for both models. At approximately 90% active depth, classification accuracy is essentially preserved: 90.97% versus 90.72% for ViT-B and 93.88% versus 94.16% for ViT-G. Figure 5(b) shows that final representation similarity also increases monotonically with register-active depth. Importantly, functional recovery can precede exact representational recovery. This effect is particularly clear in ViT-G, where retaining registers through 90% of the network preserves near-full classification accuracy despite a patch-token cosine similarity of only 0.802 to the full-register model.

These results reveal an important distinction between persistence and functional influence. The low Self-Autonomy of shallow register states does not imply that early register computation is dispensable. Rather, shallow registers are weakly self-persistent while already exerting substantial downstream influence through their interactions with other token streams. As depth increases, this functional contribution is progressively accumulated, while the register states themselves become increasingly self-persistent and heterogeneous. Together with the causal analyses in Sec. 4, the early-exit results therefore support a depth-dependent functional transformation: register tokens evolve from transient but influential interaction states at shallow depth toward increasingly self-persistent and specialized information-routing streams at deeper layers.

6 Conclusions

We presented a systematic causal analysis of register-token dynamics in DINOv2 Vision Transformers using dual-path activation patching. Our results show that register streams undergo a clear depth-dependent transformation: shallow registers are weakly self-persistent yet remain strongly coupled to the spatial stream, whereas deeper registers exhibit increasingly self-persistent routing and progressively differentiated causal profiles. This transition is strongly scale- and stream-dependent, while the resulting register organization remains highly consistent across visual conditions. A test-time register early-exit analysis further reveals that low Self-Autonomy does not imply functional irrelevance. Early register participation already exerts substantial downstream influence, while most task-relevant contribution is progressively consolidated by later layers. Together, these findings characterize register tokens not as homogeneous auxiliary storage, but as dynamic information-routing streams whose roles evolve from transient interaction to self-persistent specialization across network depth.

REFERENCES

Linus Ruben Bach, Emma Bakker, Rénan van Dijk, Jip de Vries, and Konrad Szewczyk. Registers in small vision transformers: A reproducibility study of vision transformers need registers. *Transactions on Machine Learning Research*, 2025.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3319–3327, 2017.

Spiros Baxevanakis, Platon Karageorgis, Ioannis Dravilas, and Konrad Szewczyk. Do all vision transformers need registers? a cross-architectural reassessment. *arXiv preprint arXiv:2603.25803*, 2026.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 782–791, 2021.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Thomas Fel, Binxu Wang, Michael A. Lepori, Matthew Kowal, Andrew Lee, Randall Balestriero, Sonia Joseph, Ekdeep Singh Lubana, Talia Konkle, Demba E. Ba, and Martin Wattenberg. Into the rabbit hull: From task-relevant concepts in DINO to minkowski geometry. In *International Conference on Learning Representations*, 2026.

Lukas Fesser, Mozes Jacobs, Thomas Fel, Andy Keller, and Sham Kakade. A unifying view of attention sinks: Two algorithms, two solutions. *arXiv preprint arXiv:2606.08105*, 2026.

Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting the second-order effects of neurons in clip. In *International Conference on Learning Representations (ICLR)*, 2025.

Nick Jiang, Amil Dravid, Alexei A. Efros, and Yossi Gandelsman. Vision transformers don't need trained registers. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Sonia Joseph, Praneet Suresh, Ethan Goldfarb, Lorenz Hufe, Yossi Gandelsman, Robert Graham, Danilo Bzdok, Wojciech Samek, and Blake Aaron Richards. Steering CLIP's vision transformer with sparse autoencoders. *arXiv preprint arXiv:2504.08729*, 2025.

Jinyeong Kim, Junhyeok Kim, Yumin Shim, Joohyeok Kim, Sunyoung Jung, and Seong Jae Hwang. Interpreting vision transformers via residual replacement model. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7057–7075, 2020.

Alexander Lappe and Martin A. Giese. Register and [CLS] tokens induce a decoupling of local and global features in large ViTs. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Kazumoto Nakamura, Yuji Nozawa, Yu-Chieh Lin, Kengo Nakata, and Youyang Ng. Improving image clustering with artifacts attenuation via inference-time attention engineering. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 277–295, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Felipe Parodi, Jordan K. Matelsky, and Melanie Segado. Zero-ablation overstates register content dependence in DINO vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4164–4170, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Brian S Robinson, Nathan Drenkow, Colin Conwell, and Michael Bonner. A sparse null code emerges in deep neural networks. In *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, pp. 302–314, 2024.

Baifeng Shi, Siyu Gai, Trevor Darrell, and Xin Wang. Toast: Transfer learning via attention steering. *arXiv preprint arXiv: 2305.15542*, 2023.

Cheng Shi, Yizhou Yu, and Sibei Yang. Vision transformers need more than registers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26328–26337, 2026.

Samuel Stevens, Wei-Lun Chao, Tanya Berger-Wolf, and Yu Su. Sparse autoencoders for scientifically rigorous interpretation of vision models. *arXiv preprint arXiv:2502.06755*, 2025.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. 2024.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, volume 139, pp. 10347–10357, July 2021a.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 32–42, October 2021b.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *European Conference on Computer Vision (ECCV)*, pp. 516–533, 2022.

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 *Advances in Neural Information Processing Systems*, volume 33, 2020.

Haoqi Wang, Tong Zhang, and Mathieu Salzmann. Sinder: Repairing the singular defects of dinov2. In *European Conference on Computer Vision (ECCV)*, pp. 20–35, 2024.

Sarah Wiegreffe and Yuval Pinter. Attention is not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, 2019.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024.

Zipeng Yan, Yinjie Chen, Chong Zhou, Bo Dai, and Andrew Luo. Vision transformers with self-distilled registers. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Jiawei Yang, Katie Z Luo, Jiefeng Li, Congyue Deng, Leonidas Guibas, Dilip Krishnan, Kilian Q Weinberger, Yonglong Tian, and Yue Wang. Denoising vision transformers. In *European Conference on Computer Vision (ECCV)*, pp. 453–469, 2024.

Srikar Yellapragada, Kowshik Thopalli, Vivek Narayanaswamy, Wesam Sakla, Yang Liu, Yamen Mubarka, Dimitris Samaras, and Jayaraman J. Thiagarajan. Leveraging registers in vision transformers for robust adaptation. *arXiv preprint arXiv:2501.04784*, 2025.

Itay Yona, Ilia Shumailov, Jamie Hayes, and Yossi Gandelsman. Interpreting the repeated token phenomenon in large language models. In *International Conference on Machine Learning (ICML)*, 2025.

A Appendix

A.1 Overview of the Dual-Path Activation Patching Framework

Figure 6 provides an overview of the proposed dual-path activation patching framework. The analysis consists of four main stages: controlled image construction, extraction of clean and corrupted ViT trajectories, dual-path causal intervention, and joint analysis of the resulting causal measurements.

First, clean-corrupted image pairs are constructed from Pascal VOC using semantic segmentation masks. The selected images are divided into three scene groups, with 30 images in each group, and three background-only corruptions are considered: Gaussian noise, black replacement, and Gaussian blur. All annotated foreground regions are preserved during corruption; for cluttered multi-object scenes, the foreground mask is defined as the union of all annotated object regions. Both the clean image I and its corrupted counterpart I' are then passed through the same frozen DINOv2 ViT, yielding the layer-wise clean trajectory $\{X_l\}_{l=1}^{L}$ and corrupted trajectory $\{\widetilde{X}_l\}_{l=1}^{L}$. At each depth l, the hidden sequence is decomposed into bottleneck tokens, $\mathcal{T} = \{[\mathrm{CLS}], \mathrm{REG}_1, \ldots, \mathrm{REG}_4\}$, and spatial patch tokens $\mathcal{P}$.

Starting from the corrupted trajectory, we perform two complementary causal interventions. In the *Self-Autonomy* path, only the clean state of a single bottleneck token t is restored at layer l, while all remaining states remain corrupted. The resulting recovery of the final representation of the same token defines Self-Autonomy $\mathcal{A}_l^{(t)}$. In the *Patch Receptivity* path, all clean spatial patch states are restored at layer l, while the bottleneck states remain corrupted. The resulting recovery of the final bottleneck representation defines Patch Receptivity $\mathcal{R}_l^{(t)}$.

The two quantities are jointly represented as

$$\mathcal{D}^{(t)}(l) = \left(\mathcal{A}_l^{(t)}, \mathcal{R}_l^{(t)}\right), \tag{13}$$

and their relative balance is summarized by

$$\Delta_l^{(t)} = \mathcal{A}_l^{(t)} - \mathcal{R}_l^{(t)}. \tag{14}$$

Negative $\Delta_l^{(t)}$ indicates stronger recoverability through the spatial stream, whereas positive $\Delta_l^{(t)}$ indicates stronger self-persistent support through the bottleneck state itself. These quantities form the basis for the analyses of depth-dependent causal dynamics, crossover depth, progressive register specialization, and condition-robust register organization presented in the main paper. Register early exit is evaluated separately as a lightweight functional validation of the distinction between state persistence and broader functional influence.

A.2 Dataset Construction and Corruption Details

A.2.1 Scene Groups and Foreground-Background Masks

We construct controlled clean-corrupted image pairs from Pascal VOC using the provided semantic segmentation annotations. We select 90 images in total and divide them into three scene groups of 30 images each: (i) *Group A*, simple single-object scenes in which the background occupies more than 50% of the image area; (ii) *Group B*, single-object-dominated scenes with more complex background structure; and (iii) *Group C*, cluttered multi-object scenes.

For each image, the segmentation annotation is used to separate foreground object regions from the background. Let $M_{\mathrm{fg}} \in \{0,1\}^{H \times W}$ denote the foreground mask and $M_{\mathrm{bg}} = 1 - M_{\mathrm{fg}}$ its complementary background mask. All corruptions are applied *only* to pixels within M_{bg}; foreground pixels

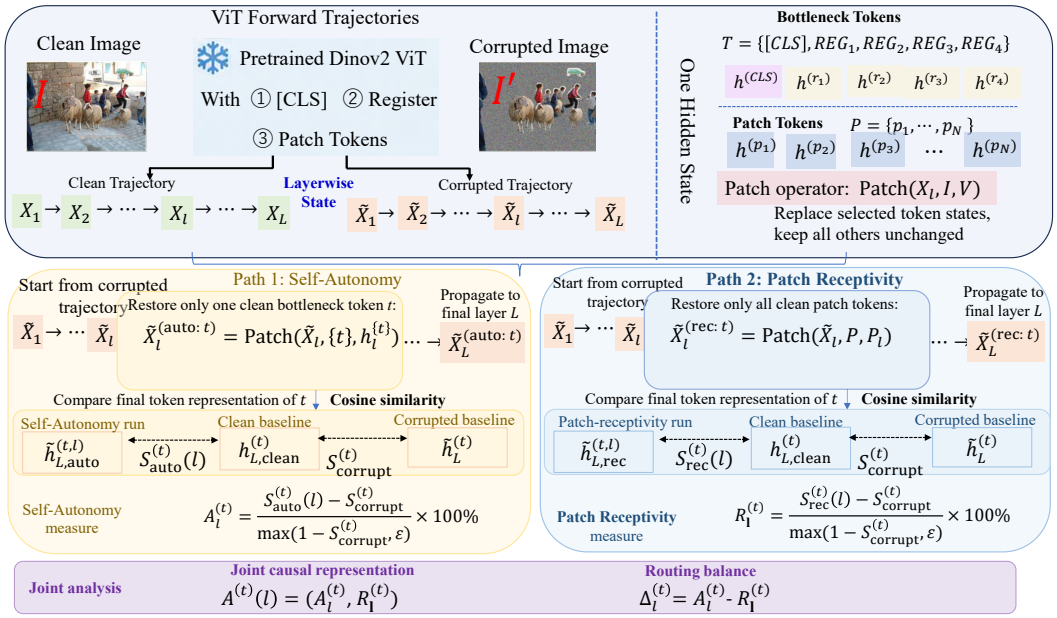


Figure 6: **Overview of the dual-path activation patching framework.** Given the clean-corrupted image pairs, the clean and corrupted images are propagated through the same frozen DINOv2 ViT to obtain layer-wise trajectories. At each intervention depth, the Self-Autonomy path restores only one clean bottleneck-token state, whereas the Patch Receptivity path restores the clean spatial patch stream while leaving bottleneck states corrupted. The resulting recovery scores, $\mathcal{A}_l^{(t)}$ and $\mathcal{R}_l^{(t)}$, jointly characterize the causal support of each bottleneck stream and are used to analyze routing balance, crossover depth, and progressive register specialization.

are preserved exactly as in the clean image. For multi-object scenes, all annotated object regions are combined to form the foreground mask, and the remaining pixels are treated as background. This design allows us to perturb contextual information while keeping the principal foreground content unchanged.

A.2.2 Background Corruptions

For each clean image, we generate three corrupted counterparts by modifying only the background region: Gaussian noise, black replacement, and Gaussian blur. Combining the three scene groups with the three corruption types yields nine scene-corruption conditions.

Gaussian Noise: For the Gaussian-noise condition, an RGB noise image is independently sampled from

$$N_{ijk} \sim \mathcal{N}(128, 50^2), \tag{15}$$

where i, j index spatial position and k indexes the RGB channel. The sampled values are clipped to the valid $[0, 255]$ intensity range and converted to 8-bit integers. Background pixels are then replaced by the corresponding noise values:

$$I'_{ijk} = \begin{cases} N_{ijk}, & M_{\text{bg}}(i,j) = 1, \\ I_{ijk}, & \text{otherwise.} \end{cases} \tag{16}$$

Thus, the corruption removes the original background structure while preserving a spatially unstructured high-variance signal.

Black Replacement: For the black-background condition, all background pixels are replaced by zero:

$$I'_{ijk} = \begin{cases} 0, & M_{\text{bg}}(i,j) = 1, \\ I_{ijk}, & \text{otherwise.} \end{cases} \tag{17}$$

This removes both background appearance and local texture while preserving the foreground object unchanged.

Gaussian Blur. For the blur condition, we first apply a Gaussian blur with radius 15 to the full image using the PIL `GaussianBlur` operator. Only the background pixels from this blurred image are then copied into the corrupted image:

$$I'_{ijk} = \begin{cases} \mathcal{G}_{15}(I)_{ijk}, & M_{\mathrm{bg}}(i,j) = 1, \\ I_{ijk}, & \text{otherwise}, \end{cases} \tag{18}$$

where $\mathcal{G}_{15}$ denotes Gaussian filtering with PIL blur radius 15. Consequently, large-scale background appearance is retained while fine spatial details are strongly suppressed.

A.2.3 Image Preprocessing and Metric Aggregation

Before being passed to DINOv2, both the clean and corrupted images undergo the same preprocessing pipeline, including resizing to 224×224 using bilinear interpolation and normalization using the mean and standard deviation of the corresponding pretrained DINOv2 checkpoint. No additional augmentation is applied during the causal intervention analysis.

For each image, Self-Autonomy $\mathcal{A}_l^{(t)}$ and Patch Receptivity $\mathcal{R}_l^{(t)}$ are computed independently for every bottleneck stream and intervention depth. The resulting measurements are first averaged over the 30 images belonging to each scene-corruption condition. Unless otherwise stated, the curves reported in the main paper are subsequently averaged over the nine scene-corruption conditions. Condition-level analyses retain the nine condition-specific averages rather than collapsing them into a single overall mean.

A.2.4 Qualitative Examples of Controlled Corruptions

Figure 7 provides representative examples of the clean and corrupted inputs used in our analysis. We show five-to-six images from each of the three scene groups together with their Gaussian-noise, black, and blurred-background counterparts. Across all corruption types, only the background region is modified according to the Pascal VOC segmentation mask, while foreground object pixels remain unchanged. The examples illustrate the increasing scene complexity from simple single-object scenes (Group A), to single-object scenes with structured backgrounds (Group B), and finally to cluttered multi-object scenes (Group C).

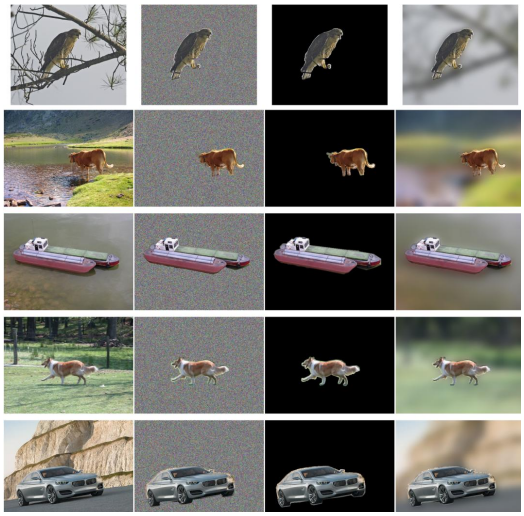
A.3 Experimental and Statistical Protocol

Models and preprocessing: We evaluate pretrained DINOv2 ViT-S/14, ViT-B/14, ViT-L/14, and ViT-G/14 models equipped with four register tokens, containing 12, 12, 24, and 40 Transformer blocks, respectively. All model parameters are frozen and the models are evaluated in inference mode. Clean and corrupted images undergo identical preprocessing, including resizing to 224, BILINEAR interpolation, and the normalization associated with the pretrained DINOv2 checkpoints. No data augmentation is applied during the causal intervention experiments.

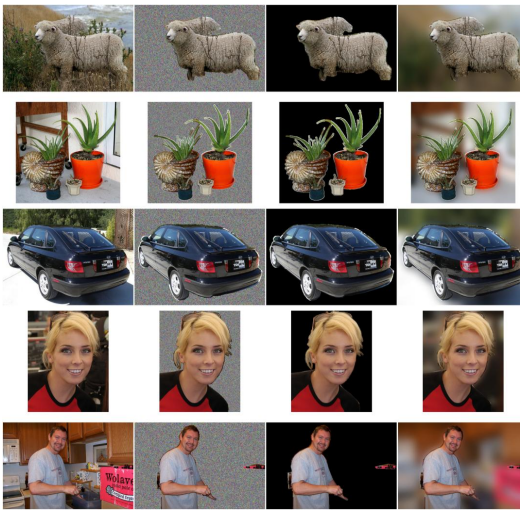
Intervention location: We define X_l as the hidden token sequence at the output of Transformer block l and before it is passed to block $l+1$. Activation patching is performed at this location. After replacing the selected token states, the modified sequence is propagated through all remaining blocks without any further intervention. The same intervention location is used for Self-Autonomy and Patch Receptivity. For Self-Autonomy, exactly one bottleneck token is restored at a time, whereas all remaining bottleneck and patch tokens retain their corrupted states. For Patch Receptivity, all spatial patch tokens are restored simultaneously, while all bottleneck tokens remain corrupted.

Metric aggregation: Self-Autonomy $\mathcal{A}_l^{(t)}$ and Patch Receptivity $\mathcal{R}_l^{(t)}$ are first computed independently for each image, bottleneck stream, and intervention depth. For each of the nine scene–corruption conditions, the resulting scores are then averaged over the 30 images belonging to that condition. Unless otherwise stated, the trajectories shown in the main paper are subsequently averaged over the nine condition-level means. Thus, the default aggregation order is

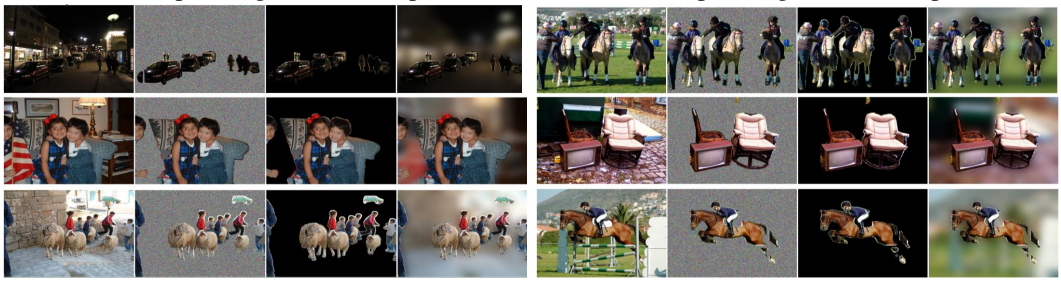
$$\text{image} \rightarrow \text{condition mean} \rightarrow \text{nine-condition mean}.$$



(a) Sample images from Group A



(b) Sample images from Group B



(c) Sample images from Group C

Figure 7: **Representative clean-corrupted image pairs used in the causal analysis.** Five-to-six examples are shown for each scene group: (a) simple single-object scenes (Group A), (b) single-object-dominated scenes with complex backgrounds (Group B), and (c) cluttered multi-object scenes (Group C). For each clean image, we generate three background-only corruptions: Gaussian noise ($\mu = 128$, $\sigma = 50$), black replacement, and Gaussian blur (PIL radius 15). Foreground regions defined by the Pascal VOC segmentation masks remain unchanged.

Depth regimes and register heterogeneity: For summary analyses, normalized non-terminal depth is divided into three equal regimes:

$$\text{Early} : \rho \in [0, 1/3), \qquad \text{Middle} : \rho \in [1/3, 2/3), \qquad \text{Late} : \rho \in [2/3, 1).$$

The terminal intervention point is excluded. For each model-condition pair, $H_l = \mathrm{Std}_{t \in \mathrm{REG}}[\Delta_l^{(t)}]$ is averaged within each depth regime, yielding H_{early}, H_{middle}, and H_{late}.

The Friedman test for progressive heterogeneity treats each of the $4 \times 9 = 36$ model-condition pairs as one paired statistical unit across the three depth regimes. It therefore assesses the consistency of the depth-wise ordering across model-condition summaries rather than image-level sampling variability.

Crossover analysis: For each register stream and scene-corruption condition, crossover is defined as the first non-terminal change of routing balance $\Delta_l^{(t)}$ from negative to non-negative. The crossover location is estimated by linear interpolation between the two adjacent blocks. Trajectories that remain negative through the final non-trivial intervention depth are treated as right-censored and are not assigned an artificial crossover depth. Crossover prevalence is therefore reported separately from the mean crossover depth among observed crossers.

Statistical analysis: All statistical tests are performed on aggregated experimental summaries rather than treating individual image-level measurements as independent replicates. For progres-

Table 2: Register-wise crossover statistics. "Cross." denotes the number of observed non-terminal crossovers among the nine conditions. Mean ρ_* is computed only over observed crossers.

	REG_1		REG_2		REG_3		REG_4	
Model	Cross.	Mean ρ_*	Cross.	Mean ρ_*	Cross.	Mean ρ_*	Cross.	Mean ρ_*
ViT-S	1/9	–	0/9	–	7/9	0.90	1/9	–
ViT-B	1/9	–	0/9	–	9/9	0.85	0/9	–
ViT-L	6/9	0.93	9/9	0.89	9/9	0.92	9/9	0.89
ViT-G	9/9	0.85	9/9	0.66	9/9	0.77	9/9	0.80

sive register heterogeneity, the paired unit is a model-condition combination ($4 \times 9 = 36$ units). For the effect of background corruption on depth-averaged routing balance, the paired unit is a model-register combination after averaging over scene groups ($4 \times 4 = 16$ units). For the effect of scene structure on late-depth heterogeneity, the paired unit is a model-corruption combination ($4 \times 3 = 12$ units). Friedman tests are performed across the corresponding three matched conditions, and Kendall's W is reported as an effect-size measure. Accordingly, the reported p-values quantify consistency across these aggregated experimental units rather than image-level sampling variability.

Register early-exit protocol: For the CIFAR-100 validation experiment, the pretrained DINOv2 backbone is kept frozen and a single linear classifier is trained on the final full-register $[\mathrm{CLS}]$ representation. The classifier is optimized with AdamW for 15 epochs using a batch size of 1024, a learning rate of 10^{-3}, and weight decay of 10^{-4}. Cross-entropy loss is used, and the classifier obtained after the final training epoch is used for all subsequent evaluations; no validation-based checkpoint selection is performed. The random seed is fixed to 42.

All CIFAR-100 images are deterministically resized to 224×224 using bicubic interpolation, converted to tensors, and normalized with mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$. No data augmentation is applied. The same trained linear classifier is reused for every register early-exit depth without further optimization.

Top-1 accuracy is evaluated on the complete CIFAR-100 test set ($N_{\mathrm{test}} = 10{,}000$). Representation similarity is computed on a fixed subset of $1{,}000$ test images. Because the test loader is deterministic and not shuffled, this subset corresponds to the first $1{,}000$ test samples and is identical across all early-exit configurations. For each sample, $[\mathrm{CLS}]$ similarity is computed between the early-exit and full-register representations, while patch similarity is obtained by averaging cosine similarity over all spatial patch tokens.

A.4 Additional results

A.4.1 Condition-Wise Causal Trajectories

The main paper reports causal trajectories averaged over the nine scene-corruption conditions. Figures 8 and 9 provide the corresponding condition-wise trajectories for ViT-S and ViT-G, respectively. For clarity, each curve summarizes the mean trajectory over the four register streams, with solid lines denoting Self-Autonomy and dashed lines denoting Patch Receptivity. The bold black curves show the corresponding nine-condition averages. Despite condition-dependent variation in routing strength, the overall pattern is consistent across conditions in both models: Self-Autonomy tends to increase with depth, whereas Patch Receptivity decreases. Moreover, the qualitative transition from spatially recoverable to increasingly self-persistent routing remains visible across all nine scene–corruption conditions, supporting the robustness claims in the main text.

A.4.2 Register-Wise Crossover Statistics

Table 2 reports the number of observed crossovers for each register stream across the nine conditions, together with the mean normalized crossover depth among trajectories exhibiting an observed crossover. Trajectories without a non-terminal crossover are treated as right-censored and are excluded from the mean crossover-depth calculation.

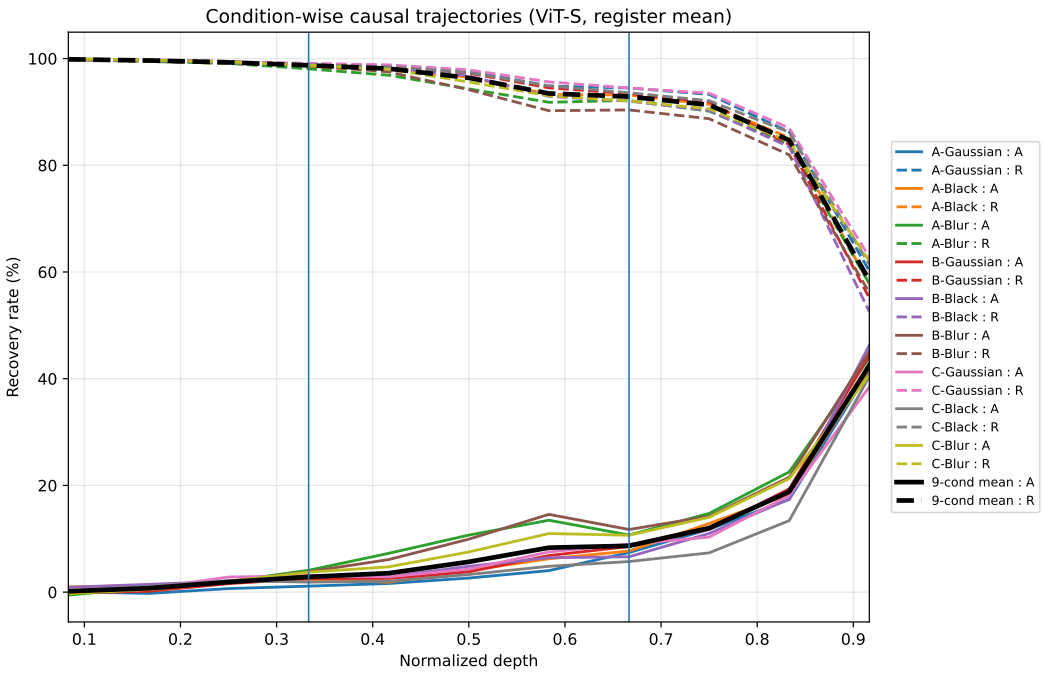


Figure 8: **Condition-wise causal trajectories for ViT-S.** Each colored pair of curves corresponds to one of the nine scene-corruption conditions, with solid lines denoting Self-Autonomy and dashed lines denoting Patch Receptivity. Each trajectory is averaged over the four register streams. The bold black curves show the nine-condition mean. Vertical lines mark the boundaries between the early, middle, and late depth regimes.

A.4.3 Detailed Register Early-Exit Results

Table 3 reports the complete numerical results corresponding to Fig. 5. We evaluate register early exit by allowing register tokens to participate in the first fraction of Transformer blocks and removing them from all subsequent blocks. The backbone remains frozen throughout evaluation. The linear classifier is trained only once using representations produced by the full-register model and is kept fixed for all early-exit configurations. We report Top-1 classification accuracy, its difference relative to the corresponding full-register model (ΔTop-1), and cosine similarity of the final $[\mathrm{CLS}]$ and mean patch representations relative to the full-register representations. Classification accuracy is evaluated on all $10{,}000$ CIFAR-100 test images, while representation similarity is computed on a fixed subset of $1{,}000$ test images. Because the number of Transformer blocks is discrete, the requested exit ratio does not always coincide exactly with the realized ratio. In particular, the requested 90% setting for ViT-B corresponds to retaining registers through 11 of 12 blocks, yielding an actual ratio of 0.917. At the deepest early-exit setting, performance is essentially preserved: ViT-B reaches 90.97% compared with 90.72% for the full-register baseline, while ViT-G reaches 93.88% compared with 94.16%. Interestingly, task-level recovery can precede exact representational recovery. At the 90% exit setting, ViT-G retains 93.88% Top-1 accuracy compared with 94.16% for the full model, despite a substantially lower patch-representation cosine similarity of 0.802. This suggests that exact reconstruction of the full-register representation is not necessary for preserving downstream classification performance.

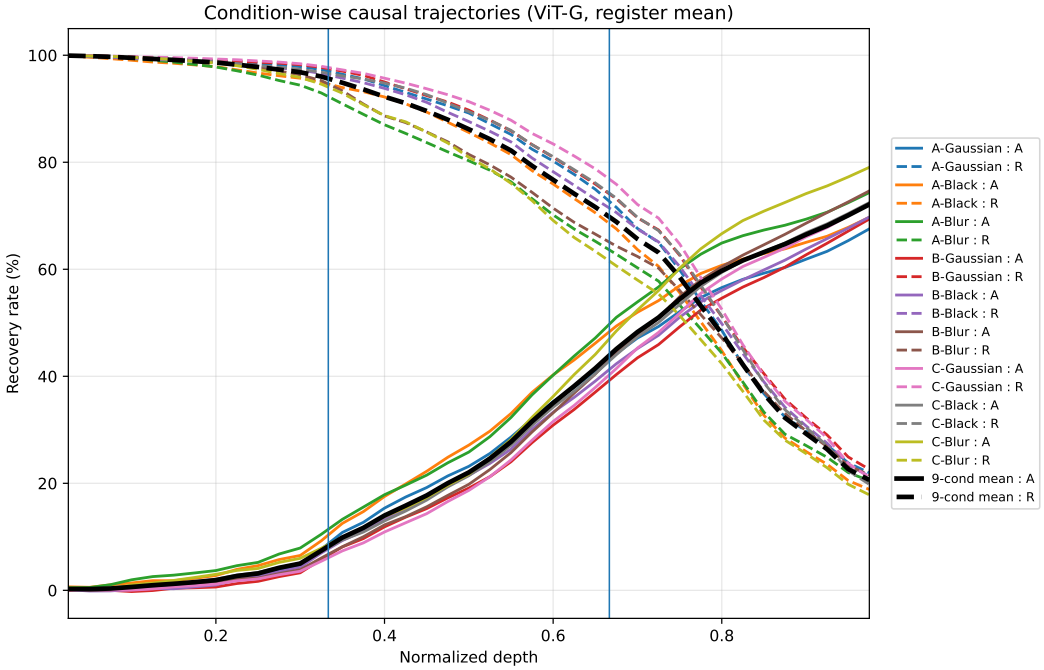


Figure 9: **Condition-wise causal trajectories for ViT-G.** Each colored pair of curves corresponds to one of the nine scene-corruption conditions, with solid lines denoting Self-Autonomy and dashed lines denoting Patch Receptivity. Each trajectory is averaged over the four register streams. The bold black curves show the nine-condition mean. Vertical lines mark the boundaries between the early, middle, and late depth regimes.

Table 3: **Detailed register early-exit results on CIFAR-100.** "Req." denotes the requested fraction of Transformer blocks for which register tokens remain active, while "Actual" denotes the realized fraction after mapping to an integer number of blocks. ΔTop-1 is measured relative to the corresponding full-register model. $\mathrm{Cos}_{\mathrm{CLS}}$ and $\mathrm{Cos}_{\mathrm{Patch}}$ denote cosine similarity of the final [CLS] representation and mean patch representation, respectively, relative to the full-register model.

Model	Setting	Req.	Active / Total	Actual	Top-1 (%)	ΔTop-1	$\mathrm{Cos}_{\mathrm{CLS}}$	$\mathrm{Cos}_{\mathrm{Patch}}$
ViT-B/14	None	0.00	0/12	0.000	18.91	-71.81	0.171	0.270
	Early-25%	0.25	3/12	0.250	50.45	-40.27	0.371	0.456
	Early-50%	0.50	6/12	0.500	77.52	-13.20	0.610	0.670
	Early-75%	0.75	9/12	0.750	86.59	-4.13	0.798	0.812
	Early-90%	0.90	11/12	0.917	90.97	+0.25	0.964	0.970
	Full	1.00	12/12	1.000	90.72	0.00	1.000	1.000
ViT-G/14	None	0.00	0/40	0.000	37.59	-56.57	0.216	0.284
	Early-25%	0.25	10/40	0.250	51.83	-42.33	0.321	0.365
	Early-50%	0.50	20/40	0.500	80.15	-14.01	0.576	0.526
	Early-75%	0.75	30/40	0.750	90.86	-3.30	0.773	0.662
	Early-90%	0.90	36/40	0.900	93.88	-0.28	0.907	0.802
	Full	1.00	40/40	1.000	94.16	0.00	1.000	1.000

Top-1 accuracy is evaluated on $N_{\mathrm{test}} = 10{,}000$ images. Representation similarities are computed using $N_{\mathrm{sim}} = 1{,}000$ images for every configuration.