CAUSAL DARWINIAN HEAD SELECTION: A MECHANISTIC FRAMEWORK FOR INTERPRETING ATTENTION HEADS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding how internal components of Transformer models contribute to prediction remains a central challenge for mechanistic interpretability. We propose *Causal Darwinian Head Selection* (CDHS), a Darwinian, causal intervention framework for analyzing attention heads as a population of computational units. CDHS uses temporary ordered ablations to estimate head-level causal contribution and to measure the resulting changes in model behavior; it is intended as a post hoc diagnostic method. Applied to Transformer-based image classifiers, CDHS jointly characterizes three aspects of head organization: sparse causal fitness, class-conditional behavioral effects, and checkpoint-wise emergence under task-aligned supervision. We find that a small subset of dominant heads contributes disproportionately to predictive performance, whereas temporarily intervening on many low-fitness heads has limited immediate effect. Interventions on dominant heads induce structured class-level shifts in prediction behavior, suggesting that some heads support class-discriminative computations rather than only global accuracy. Using checkpoint-wise analysis and a matched label-permutation control, we further observe that fitness concentration strengthens during training under real labels more than under permuted labels. These results position CDHS as a framework for linking causal intervention, class-level model behavior, and training dynamics in vision Transformers.

1 INTRODUCTION

Transformer architectures have become a dominant paradigm in vision, language, and multimodal learning, with multi-head attention as a central computational mechanism (Banerjee & Banik, 2025; Jin et al., 2025). Despite their architectural discreteness, attention heads remain poorly understood as mechanistic units: it is still unclear whether they implement distinct causal computations or primarily provide interchangeable representational capacity (Wang et al., 2025; Chen et al., 2024; Basile et al., 2025). This distinction is critical for mechanistic interpretability, since explanations based on architectural components are meaningful only when those components make identifiable causal contributions to model behavior. Much prior interpretability work analyzes attention weights, yet attention patterns alone do not establish causal relevance, as they may change without corresponding effects on predictions. This limitation motivates causal interventions on internal components. Attention heads are a natural unit for such analysis: they are discrete, repeatedly applied across inputs, and directly involved in Transformer computation. However, existing studies often emphasize efficiency, pruning, or representational capacity, leaving the causal organization, functional differentiation, and training-time emergence of head specialization less well characterized.

We introduce *Causal Darwinian Head Selection* (CDHS), an ordered-intervention framework for measuring head-level causal fitness. CDHS treats attention heads as a population of computational units and evaluates their contribution through counterfactual silencing. Heads whose temporary silencing substantially degrades performance are assigned high causal fitness, whereas heads with limited immediate effect are treated as low-fitness or functionally neutral under the evaluated task. Importantly, CDHS is used only as an evaluation-time diagnostic: the architecture and learned parameters remain fixed, and head silencing does not alter the training objective. Using CDHS, we analyze Transformer-based image classifiers along three axes. First, we measure whether causal importance

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.

is uniformly distributed across heads or concentrated in a small subset. Second, we examine whether high-fitness heads affect only global accuracy or induce structured class-conditional behavioral changes. Third, we track head-level causal fitness across training checkpoints to study when sparse causal organization emerges. To distinguish task-aligned specialization from structure induced by architecture or optimization alone, we compare standard training with a matched label-permutation control that preserves the architecture, optimizer, schedule, and marginal label distribution while removing input–label alignment. Our results indicate that attention heads form a sparse and functionally differentiated population of causal components. A small subset of high-fitness heads accounts for a disproportionate share of predictive performance: silencing these heads rapidly degrades accuracy, whereas silencing low-fitness heads has limited immediate effect. Dominant-head intervention also produces structured changes in class-level confusion patterns, indicating that high-fitness heads support specific class distinctions rather than causing uniform degradation. Checkpoint-wise analysis further shows that this sparse causal organization emerges during learning. Under real labels, head-fitness distributions become more unequal and concentrated than under the label-permutation control, separating task-aligned specialization from specialization-like effects induced by architecture, optimization, or final-checkpoint ranking.

Together, these findings suggest that attention heads are not merely interchangeable architectural replicas, but differentiated causal components shaped by task-aligned supervision. CDHS provides a unified framework for connecting component-level causal intervention, class-conditional function, and training dynamics, offering a step toward more mechanistic and modular explanations of learned Transformer computation.

2 Related Work

2.1 Interpretability and Causal Perspectives on Attention

Attention weights need not faithfully explain predictions: alternative distributions can produce similar outputs (Jain & Wallace, 2019), and removing highly weighted connections may have limited performance impact (Serrano & Smith, 2019). Their explanatory value depends on the definition of explanation, and attention may remain informative under appropriate evaluation settings (Wiegreffe & Pinter, 2019). Comparisons with gradient-based and rollout methods indicate that attention combines computational roles and does not reliably reflect causal importance (Abnar & Zuidema, 2020; Kobayashi et al., 2020; Brunner et al., 2020; Tutek & Šnajder, 2020), with related limitations in vision and multimodal transformers (Chefer et al., 2021b;a). Head ablation, pruning, and masking reveal widespread redundancy alongside a subset of heads with consistent, task-specific causal effects (Voita et al., 2019; Michel et al., 2019; Kovaleva et al., 2019; Dalvi et al., 2020). Analyses of parameter structure and activation statistics further characterize transformer components without equating attention scores with explanations (Clark et al., 2019; Vig & Belinkov, 2019; Geva et al., 2021), while mechanistic studies identify recurring head types and circuits through controlled interventions (Gould et al., 2024; Elhage et al., 2021; Conmy et al., 2023; Wang et al., 2023). We likewise treat heads as intervenable causal components, examining the structure and stability of their behavioral effects.

2.2 Mechanistic Interpretability and Circuit Analysis

Mechanistic interpretability (MI) reverse-engineers how internal components and subnetworks implement behavior, viewing networks as compositions of causal computational circuits (Olah et al., 2020; Saphra & Wiegreffe, 2024). Circuit discovery seeks compact computational subgraphs that explain specific behaviors (Conmy et al., 2023; Dunefsky et al., 2024), using interventions such as activation patching and causal tracing to assess the necessity and sufficiency of neurons and heads (Zhang & Nanda, 2024; Lepori et al., 2024; Zhong et al., 2023). Activation patching localizes function through activation ablation and substitution (Zhang et al., 2025b), while causal tracing and mediation connect internal variables to behavior within broader causal frameworks (Geiger et al., 2025). In language models, circuit analyses identify interpretable features and head–MLP computations for tasks including indirect object identification and entity tracking (Wang et al., 2023; Prakash et al., 2024; Merullo et al., 2024), as well as numerical comparisons (Hanna et al., 2023). Other studies examine circuit generalization across prompt formats and tasks (Sun, 2025). Advances in automated discovery and attribution improve subcircuit identification (Zhang et al., 2025a; Rezaei Jafari et al.,

2025; Nam et al., 2025), with gradient-based attribution patching (Syed et al., 2024) and edge pruning (Bhaskar et al., 2024) enabling efficient analysis. Further work extends circuit analysis beyond architectural components to feature representations (Ahmad et al., 2025). We extend these predominantly language-centered investigations to vision transformers, linking head-level causal structure to visual classification failures.

2.3 Darwinian Perspectives on Functional Specialization

Network overparameterization and redundancy motivate a Darwinian view of specialization under shared objectives, echoing selection and stabilization among competing neural circuits in Neural Darwinism (Edelman, 1987). Lottery-ticket studies identify sparse subnetworks capable of matching dense-model performance, suggesting implicit selection among redundant components (Frankle et al., 2020; Malach et al., 2020; Natale et al., 2024), with related pruning analyses in multi-head attention (Behnke & Heafield, 2020). Pruning and representation refinement remove redundant units or heads while retaining important components with limited performance loss (Voita et al., 2019; Ashkboos et al., 2024); feature collapse links similar task roles to convergent representations, highlighting competition during learning (Laurent et al., 2024). Specialized substructures can also emerge under task or resource constraints without explicit architectural modularity (Béna & Goodman, 2025; Qiu et al., 2024; Musat, 2025). Surveys of modular deep learning describe specialized routing, typically imposed as an inductive bias, to reduce interference and improve generalization (Pfeiffer et al., 2023), whereas representational convergence across independent training runs suggests reproducible functional "winners" (Li et al., 2016). Attention studies characterize head specialization through pruning and functional analysis (Gould et al., 2024; Basile et al., 2025), while causal frameworks offer taxonomies and intervention-based estimates of task impact (Nam et al., 2025; Rohekar et al., 2023; Ranaldi, 2025). However, specialization or survival alone does not establish a component's causal function. We systematically ablate individual heads as explicit computational units at evaluation time, providing mechanistic evidence for Darwinian-style functional differentiation beyond metaphor or statistical association.

3 Methodology

We introduce **Causal Darwinian Head Selection (CDHS)**, a diagnostic framework for measuring the causal contribution of attention heads in a trained Transformer. CDHS treats attention heads as intervenable computational units and evaluates their importance through temporary counterfactual masking at evaluation time. The model parameters are always kept fixed, and no additional training or parameter update is introduced.

3.1 Diagnostic Head Masking

Let $\mathcal{X}$ denote the input space and $\mathcal{Y}$ the label space. We consider a multi-class Transformer classifier parameterized by θ,

$$f(x;\theta) : \mathcal{X} \to \Delta(\mathcal{Y}), \tag{1}$$

where $\Delta(\mathcal{Y})$ is the probability simplex over classes. The model is evaluated on a fixed data distribution $\mathcal{D}$ over input–label pairs (x, y).

Let $\mathcal{H} = \{h_1, h_2, \ldots, h_N\}$ denote the set of all attention heads in the model. Each attention head h_i produces an intermediate representation $h_i(x)$ that contributes to the final prediction through residual connections and downstream layers. We write the model as

$$f(x) = g\big(x, h_1(x), h_2(x), \ldots, h_N(x)\big), \tag{2}$$

where $g(\cdot)$ denotes the remaining computation of the network, including residual connections, normalization layers, and MLP blocks.

For any subset of heads $S \subseteq \mathcal{H}$, we define a diagnostic masking operator

$$\mathcal{A}(f, S)(x) = g\big(x, \tilde{h}_1^S(x), \tilde{h}_2^S(x), \ldots, \tilde{h}_N^S(x)\big), \tag{3}$$

where

$$\tilde{h}_i^S(x) = \begin{cases} 0, & h_i \in S, \\ h_i(x), & h_i \notin S. \end{cases} \tag{4}$$

The masking operation is applied only to head outputs prior to residual addition. All model parameters and all other computations remain unchanged. Thus, CDHS is used as an evaluation-time diagnostic probe rather than a pruning or retraining method.

3.2 Head-Level Causal Fitness

We measure the contribution of an individual attention head by comparing the original model with a counterfactual model in which that head is masked. For head h_i, the interventional model is

$$f^{(i)}(x) = \mathcal{A}(f, \{h_i\})(x). \tag{5}$$

Let the task loss be

$$\mathcal{L}(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\ell(f(x), y)\big], \tag{6}$$

where ℓ is an evaluation loss. We define the average causal effect of head h_i as

$$\mathrm{ACE}(h_i) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\ell(f^{(i)}(x), y) - \ell(f(x), y)\big]. \tag{7}$$

A larger $\mathrm{ACE}(h_i)$ means that masking h_i causes a larger increase in task loss, indicating a stronger causal contribution to model performance.

In practice, we estimate $\mathrm{ACE}(h_i)$ on K evaluation samples:

$$\widehat{\mathrm{ACE}}(h_i) = \frac{1}{K}\sum_{k=1}^{K}\Big(\ell(f^{(i)}(x_k), y_k) - \ell(f(x_k), y_k)\Big), \quad (x_k, y_k) \sim \mathcal{D}. \tag{8}$$

We define the causal fitness of head h_i as its positive causal contribution:

$$\mathcal{F}(h_i) = \max\{\widehat{\mathrm{ACE}}(h_i), 0\}. \tag{9}$$

Heads with high causal fitness are interpreted as beneficial contributors under the fixed environment $(\mathcal{D}, \theta)$. Heads with non-positive causal effect may reflect redundancy, compensatory computation, or interaction effects not captured by single-head masking.

3.3 Ordered Counterfactual Evaluation

CDHS evaluates how causal influence is distributed across attention heads by progressively masking heads according to their estimated causal fitness. Let

$$\pi = \mathrm{argsort}_i\ \widehat{\mathrm{ACE}}(h_i) \tag{10}$$

denote the ranking of heads in decreasing order of estimated causal contribution.

We define the top-down and bottom-up diagnostic sets of size k as

$$S_k^{\mathrm{TD}} = \{h_{\pi(1)}, \ldots, h_{\pi(k)}\}, \quad S_k^{\mathrm{BU}} = \{h_{\pi(N)}, \ldots, h_{\pi(N-k+1)}\}. \tag{11}$$

Top-down masking removes the most causally influential heads first, while bottom-up masking removes the least influential heads first.

We evaluate the corresponding counterfactual loss changes:

$$\Delta_k^{\mathrm{TD}} = \mathcal{L}\big(\mathcal{A}(f, S_k^{\mathrm{TD}})\big) - \mathcal{L}(f), \quad \Delta_k^{\mathrm{BU}} = \mathcal{L}\big(\mathcal{A}(f, S_k^{\mathrm{BU}})\big) - \mathcal{L}(f). \tag{12}$$

A sharp increase in the top-down curve together with a flat bottom-up curve indicates that causal contribution is concentrated in a small subset of dominant heads. Conversely, similar top-down and bottom-up curves suggest a more diffuse distribution of causal influence. We also include random-order masking as a neutral baseline. This allows us to distinguish targeted removal of high-fitness heads from degradation caused by removing heads arbitrarily.

3.4 Class-Conditional Causal Effects

To study whether high-fitness heads support global accuracy or class-specific computations, we estimate class-conditional causal effects. For class $c \in \mathcal{Y}$, we define

$$\mathrm{ACE}_c(h_i) = \mathbb{E}_{(x,y)\sim\mathcal{D}|y=c}\big[\ell(f^{(i)}(x), y) - \ell(f(x), y)\big]. \tag{13}$$

We summarize the class selectivity of head h_i using the specialization score

$$\mathrm{Spec}(h_i) = \frac{\max_c \max(\mathrm{ACE}_c(h_i), 0)}{\sum_{c'} \max(\mathrm{ACE}_{c'}(h_i), 0) + \varepsilon}, \tag{14}$$

where $\varepsilon > 0$ is a small constant for numerical stability. Negative class-conditional effects are excluded from the ratio because they may correspond to compensatory, suppressive, or regularizing behavior.

A high specialization score indicates that the positive causal contribution of a head is concentrated on a small number of classes. We use this measure to characterize whether attention heads behave as class-specific specialists, broader semantic specialists, or general-purpose contributors.

3.5 Causal Darwinian Head Selection

CDHS provides a Darwinian interpretation of the above causal measurements. The fixed data distribution and trained parameters define the environment $(\mathcal{D}, \theta)$, while the attention heads $\mathcal{H}$ form a population of computational units. Under this view, causal fitness $\mathcal{F}(h_i)$ measures how much each head contributes to task performance under counterfactual evaluation. The term "selection" refers only to analytic ranking and diagnostic masking. It does not imply architectural pruning, deployment-time compression, or additional optimization. CDHS therefore unifies three diagnostic views of attention heads: individual causal fitness, ordered population-level ablation, and class-conditional functional specialization. Additional formal details, including identifiability, finite-sample concentration, non-additivity, survival indicators, and the ecological niche interpretation, are provided in Appendix A.1.

4 Experiments

We conduct experiments on Tiny-ImageNet using standard training and evaluation splits, and report additional results on CIFAR-10 and CIFAR-100 in the appendix. To account for training stochasticity, all experiments are repeated with multiple random seeds, and we report mean performance with standard deviation. Beyond final-checkpoint ablations, we analyze intermediate training checkpoints with a label-permutation control to test whether sparse causal head specialization emerges during learning. To study the causal role of attention heads, we introduce *Causal Darwinian Head Selection (CDHS)*, a mechanistic evaluation framework inspired by Darwinian selection principles. In CDHS, attention heads are treated as competing computational units subject to explicit selection pressure imposed via causal interventions. Concretely, CDHS operates by progressively removing attention heads from the model according to a predefined selection criterion and evaluating the resulting impact on model fitness, measured by classification accuracy. Heads whose removal induces a large fitness drop are interpreted as causally advantageous components, while heads whose removal yields negligible impact are considered functionally redundant under the evaluated task.

4.1 Head Fitness and Selection Strategies

Head Fitness Estimation. Prior to ablation, we estimate the causal fitness of each attention head using a class-conditional fitness score derived from its contribution to model outputs. This score is computed independently of the ablation process and held fixed throughout all experiments to avoid circularity.

Selection Regimes. CDHS is instantiated under three distinct selection regimes, corresponding to different evolutionary pressures: **Top-down (Positive Selection)** removes heads in descending order of estimated fitness, simulating strong selection against the most causally advantageous components; **Bottom-up (Negative Selection)** removes heads in ascending order of fitness, corresponding to the elimination of weakly contributing components; and **Random Drift** removes heads in a uniformly random order, serving as a neutral evolutionary baseline.

4.2 EXPERIMENTS ON TINY-IMAGENET

4.2.1 CAUSAL IMPACT OF ATTENTION HEAD ABLATION

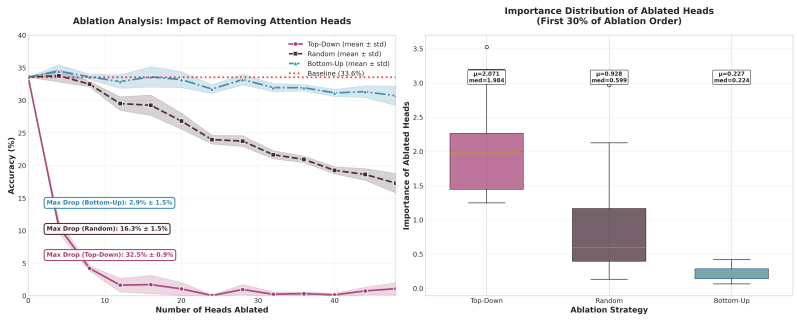


Figure 1: CDHS under progressive attention head ablation on Tiny-ImageNet.

Figure 1 left reports test accuracy on Tiny-ImageNet as a function of the number of ablated attention heads under the three CDHS selection regimes. Top-down ablation produces a sharp and catastrophic drop in performance: removing only a small subset of the most causally important heads drives accuracy from 33.6% to near zero. By contrast, random ablation yields a smoother, approximately linear degradation, whereas bottom-up ablation preserves accuracy even after a large fraction of heads has been removed. These results indicate that, despite the greater class diversity and semantic complexity of Tiny-ImageNet, performance is bottlenecked by a sparse set of causally dominant heads rather than by aggregate model capacity. Figure 1 right shows the distribution of causal fitness scores for the first 30% of ablated heads under each strategy. Heads eliminated early under top-down selection exhibit substantially higher mean and median fitness than those removed under random or bottom-up regimes. This separation confirms that the observed performance collapse is driven by targeted removal of highly influential heads, revealing a strongly skewed and heterogeneous distribution of causal fitness that persists at larger scale and finer-grained classification settings.

4.2.2 OUT-OF-SAMPLE VALIDATION OF CAUSAL HEAD RANKINGS

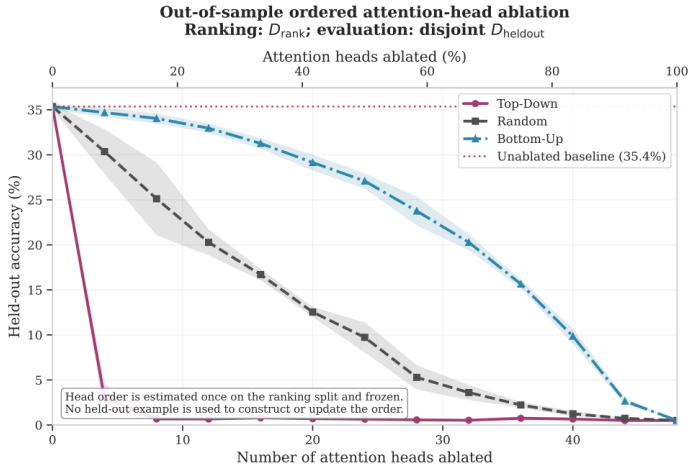


Figure 2: Out-of-sample ordered attention-head ablation on Tiny-ImageNet.

To test the out-of-sample validity of CDHS rankings, we estimate head-level causal contributions on D_{rank}, freeze the ranking, and evaluate cumulative ablations on disjoint D_{heldout}. Top-down and bottom-up ablations mask heads in descending and ascending order of estimated contribution, respectively, with random ablation as a reference independent of these scores. Held-out examples never inform the ranking, and evaluation uses fixed model parameters without retraining. Figure 2 shows that ablating fewer than 20% of the highest-ranked heads reduces held-out accuracy from 35.4%

to below 1%. With 20 heads ablated, accuracy remains approximately 29% under bottom-up ablation, compared with approximately 13% under random ablation and below 1% under top-down ablation. This separation supports the out-of-sample utility of CDHS rankings and cannot be explained solely by reusing ranking examples for evaluation.

4.2.3 Class-Conditional Effects of Causally Important Head Ablation

We next examine the class-conditional consequences of eliminating the single most causally important ("winner") attention head identified by CDHS on Tiny-ImageNet.

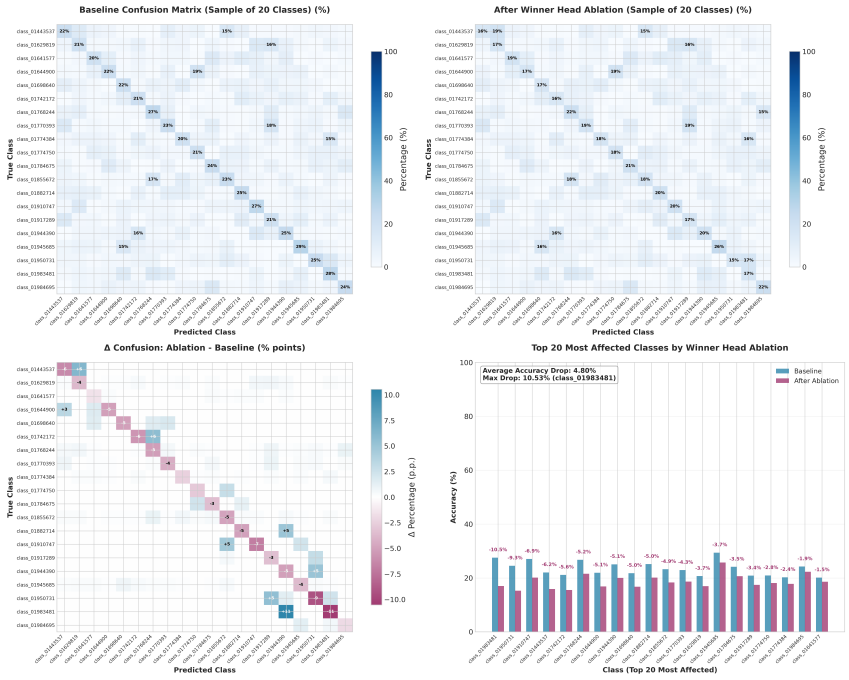


Figure 3: Class-wise causal effects of winner head ablation on Tiny-ImageNet.

Confusion Matrix Shift. Figure 3 top compares the baseline confusion matrix with that obtained after winner head ablation for a representative subset of Tiny-ImageNet classes. The resulting changes are highly structured: certain classes exhibit pronounced drops in correct classification, while others are comparatively resilient. Importantly, errors are redirected toward specific alternative classes rather than being uniformly diffused, indicating structured perturbations of decision boundaries.

Differential Confusion Analysis. The difference between post-ablation and baseline confusion matrices (Figure 3, bottom left) reveals systematic redistribution of predictions across particular class pairs. These non-random shifts suggest that the ablated head supported discriminative computations for specific semantic distinctions that are not redundantly encoded by the remaining heads.

Class-wise Accuracy Impact. Figure 3 bottom right reports per-class accuracy before and after winner head ablation, sorted by magnitude of degradation. The average accuracy drop across affected classes is 4.80%, with a maximum drop of 10.53% observed for a single class. The uneven impact across classes further supports the presence of strong class-specific functional dependencies on a small number of causally dominant attention heads.

4.2.4 Causal Functional Specialization of Attention Heads

Head–Class Specialization Matrix. Figure 4 top visualizes the normalized class-conditional specialization strength of attention heads across a subset of Tiny-ImageNet classes. Rows correspond to attention heads sorted by causal fitness, and columns to semantic classes. The resulting matrix exhibits pronounced heterogeneity: some heads display sharp, high-intensity responses to specific classes, while others contribute more diffusely across categories.

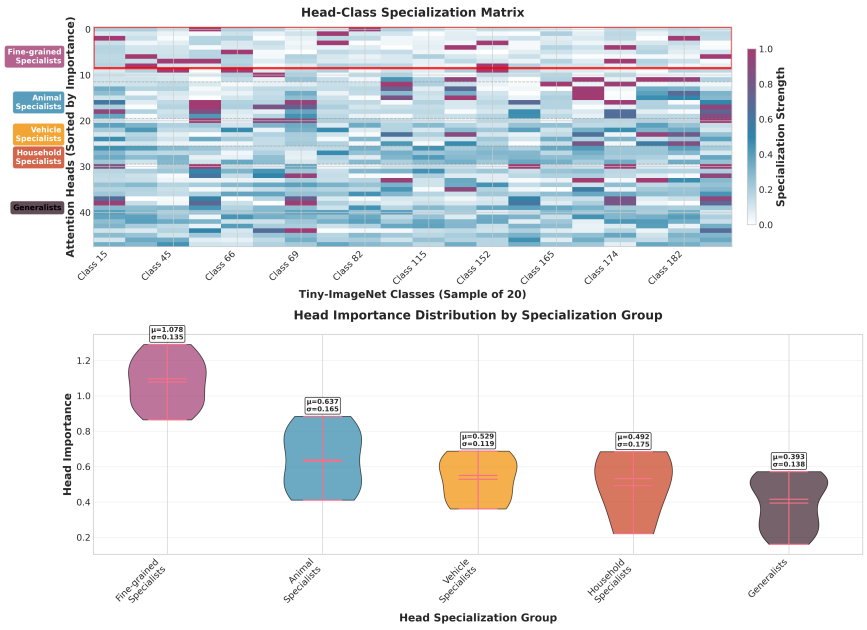


Figure 4: Emergent functional niches of attention heads identified via causal ablation on Tiny-ImageNet.

Emergent Functional Niches. Based on their specialization profiles, attention heads can be grouped into interpretable functional niches, including fine-grained specialists, animal specialists, vehicle specialists, household specialists, and generalists. These niches arise without explicit architectural constraints or supervision, indicating that CDHS uncovers latent modular organization that emerges naturally during standard training at scale.

Fitness Distribution by Niche. Figure 4 bottom reports the distribution of causal fitness scores across specialization groups. Fine-grained specialists exhibit the highest average fitness with relatively low variance, indicating consistent and non-redundant causal roles. In contrast, generalist heads show lower average fitness and higher redundancy, consistent with their greater robustness under bottom-up ablation and their more distributed functional contributions.

4.2.5 Causal Head Specialization Emerges During Training

The preceding analyses identify sparse causal fitness and functional specialization at the final checkpoint. However, final-checkpoint measurements alone do not rule out the possibility that the observed hierarchy is merely a post hoc ordering of trained attention heads. To examine whether this structure forms during learning, we track the distribution of estimated head fitness at intermediate checkpoints and compare standard training with a label-permutation control. The control uses the same architecture, optimizer, training schedule, and marginal label distribution, but removes task-aligned input–label structure. Figure 5 shows that causal head organization develops progressively under real labels. The Gini coefficient of the head-fitness distribution separates from the label-permutation control after the early training phase and remains higher across later checkpoints. This indicates that head fitness becomes increasingly unequal as task-relevant structure is learned. A complementary top-k analysis shows that the top 10% of heads capture a larger fraction of total positive fitness under real-label training than under the control. The corresponding excess curves remain positive across most checkpoints, suggesting that causal fitness is progressively concentrated into a small subset of heads during training. These results strengthen the interpretation of attention heads as a differentiated population of causal components. Dominant heads are not only identified by sorting the final model; rather, their disproportionate fitness emerges over the course of learning. Together with the class-conditional specialization results above, the training-time analysis suggests that sparse importance and functional differentiation reflect task-aligned causal organization rather than a final-checkpoint ranking artifact.

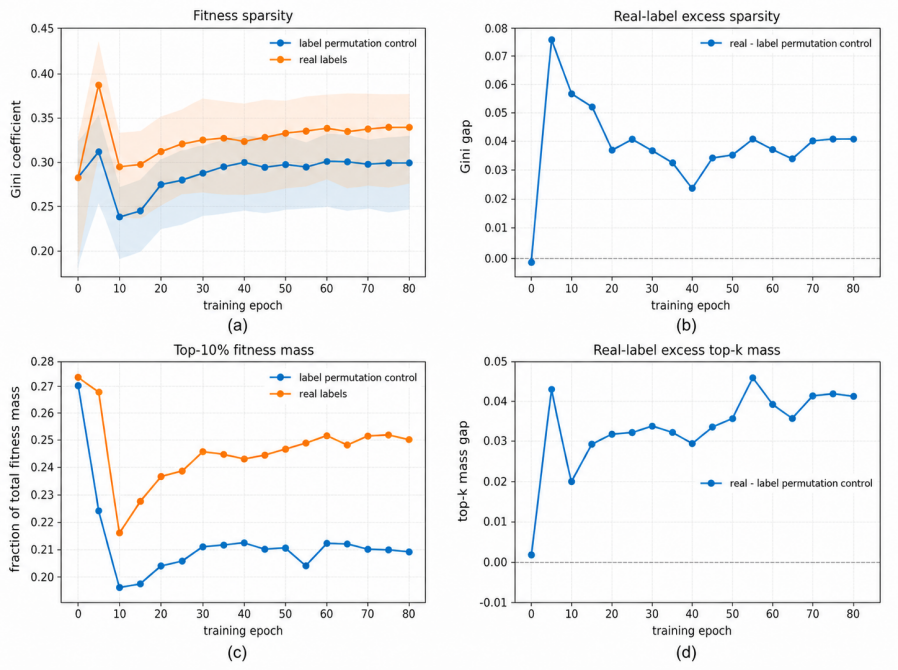


Figure 5: Training-time emergence of causal head specialization.

5 Conclusion

We introduced *Causal Darwinian Head Selection (CDHS)*, a mechanistic framework for quantifying attention-head organization through controlled causal interventions. In CDHS, "selection" refers to ranking attention heads by causal fitness. By applying ordered, temporary head ablations, CDHS links head-level causal importance to measurable changes in model behavior. Experiments on Tiny-ImageNet, with supporting results on CIFAR, show that causal fitness is highly heterogeneous across attention heads. A small subset of high-fitness heads contributes disproportionately to model performance, while many low-fitness heads have limited immediate impact under ablation. Targeted perturbation of high-fitness heads leads to rapid performance degradation, suggesting that attention heads in supervised vision Transformers are not uniformly redundant or interchangeable. Beyond aggregate accuracy, high-fitness head ablations produce structured class-level effects, including selective shifts in confusion patterns and category-specific degradation. Specialization analyses further indicate that heads occupy distinct functional roles, ranging from fine-grained specialists to semantically grouped specialists and broader generalists. Checkpoint-wise analysis shows that this organization emerges during training. Under real-label supervision, the head-fitness distribution becomes increasingly unequal and concentrated, as reflected by the separation of the Gini coefficient from a label-permutation control and the increasing share of positive causal fitness captured by top-ranked heads. These results suggest that task-aligned supervision shapes sparse and differentiated causal structure within attention modules.

As with most causal intervention studies, our analysis is conducted within a controlled experimental scope. We focus on supervised vision Transformers for image classification, which provides a tractable setting for isolating and interpreting head-level causal effects. Extending CDHS to broader architectures, modalities, and tasks is a natural direction for future work. In addition, the present study emphasizes single-head interventions to obtain clean causal rankings; future extensions could incorporate interaction-aware interventions to study non-additive effects across heads and other Transformer components.

Overall, CDHS provides a controlled framework for identifying sparse causal structure, characterizing functional specialization, and tracking the emergence of component importance during training. Our findings support a mechanistic view of attention heads as differentiated causal components shaped by task-aligned supervision.

AI USE STATEMENT

Disclosure of LLM usage. We used large language models (LLMs) only for polishing the writing; all scientific content is the authors' own.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Annual Meeting of the Association for Computational Linguistics*, 2020.

Areeb Ahmad, Abhinav Joshi, and Ashutosh Modi. Beyond components: Singular vector-based interpretability of transformer circuits. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Saleh Ashkboos, Maximilian L Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. SliceGPT: Compress large language models by deleting rows and columns. In *The Twelfth International Conference on Learning Representations*, 2024.

Anasua Banerjee and Debajyoty Banik. A comprehensive survey on transformer-based machine translation: Identifying research gaps and solutions for large language models. *ACM Computing Surveys*, 58(5):1–33, 2025.

Lorenzo Basile, Valentino Maiorca, Diego Doimo, Francesco Locatello, and Alberto Cazzaniga. Head pursuit: Probing attention specialization in multimodal transformers. In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025.

Maximiliana Behnke and Kenneth Heafield. Losing heads in the lottery: Pruning transformer attention in neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2664–2674, 2020.

Gabriel Béna and Dan FM Goodman. Dynamics of specialization in neural modules under resource constraints. *Nature Communications*, 16(1):187, 2025.

Adithya Bhaskar, Alexander Wettig, Dan Friedman, and Danqi Chen. Finding transformer circuits with edge pruning. *Advances in Neural Information Processing Systems*, 37:18506–18534, 2024.

Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers. In *International Conference on Learning Representations*, 2020.

Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 397–406, 2021a.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 782–791, 2021b.

Xingwu Chen, Lei Zhao, and Difan Zou. How transformers utilize multi-head attention in in-context learning? a case study on sparse linear regression. *Advances in Neural Information Processing Systems*, 37:119573–119613, 2024.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. In *BlackboxNLP@ACL*, 2019.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Fahim Dalvi, Hassan Sajjad, Nadir Durrani, and Yonatan Belinkov. Analyzing redundancy in pretrained transformer models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4908–4926, 2020.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable llm feature circuits. *Advances in Neural Information Processing Systems*, 37:24375–24410, 2024.

Gerald M Edelman. *Neural Darwinism: The theory of neural group selection*. Basic Books, 1987.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pp. 3259–3269. PMLR, 2020.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, et al. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26 (83):1–64, 2025.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021.

Rhys Gould, Euan Ong, George Ogden, and Arthur Conmy. Successor heads: Recurring, interpretable attention heads in the wild. In *The Twelfth International Conference on Learning Representations*, 2024.

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *Advances in Neural Information Processing Systems*, 36:76033–76060, 2023.

Sarthak Jain and Byron C Wallace. Attention is not explanation. In *North American Chapter of the Association for Computational Linguistics*, 2019.

Yizhang Jin, Jian Li, Tianjun Gu, Yexin Liu, Bo Zhao, Jinxiang Lai, Zhenye Gan, Yabiao Wang, Chengjie Wang, Xin Tan, et al. Efficient multimodal large language models: A survey. *Visual Intelligence*, 3(1):27, 2025.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention module is not only a weight: Analyzing transformers with vector norms. *Conference on Empirical Methods in Natural Language Processing*, 2020.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of bert. *Conference on Empirical Methods in Natural Language Processing*, 2019.

Thomas Laurent, James H von Brecht, and Xavier Bresson. Feature collapse. In *The Twelfth International Conference on Learning Representations*, 2024.

Michael A Lepori, Alexa R Tartaglini, Wai K Vong, Thomas Serre, Brenden M Lake, and Ellie Pavlick. Beyond the doors of perception: Vision transformers represent relations between objects. *Advances in Neural Information Processing Systems*, 37:131503–131544, 2024.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? In *ICLR*, 2016.

Eran Malach, Gilad Yehudai, Shai Shalev-Schwartz, and Ohad Shamir. Proving the lottery ticket hypothesis: Pruning is all you need. In *International Conference on Machine Learning*, pp. 6682–6691. PMLR, 2020.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Circuit component reuse across tasks in transformer language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32, 2019.

Tiberiu Musat. Mechanism and emergence of stacked attention heads in multi-layer transformers. In *The Thirteenth International Conference on Learning Representations*, 2025.

Andrew Nam, Henry Conklin, Yukang Yang, Thomas Griffiths, Jonathan Cohen, and Sarah-Jane Leslie. Causal head gating: A framework for interpreting roles of attention heads in transformers. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Emanuele Natale, Davide Ferré, Giordano Giambartolomei, Frédéric Giroire, and Frederik Mallmann-Trenn. On the sparsity of the strong lottery ticket hypothesis. *Advances in Neural Information Processing Systems*, 37:40565–40592, 2024.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

Jonas Pfeiffer, Sebastian Ruder, Ivan Vulić, and Edoardo Maria Ponti. Modular deep learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. Survey Certification.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *The Twelfth International Conference on Learning Representations*, 2024.

Zihan Qiu, Zeyu Huang, and Jie Fu. Unlocking emergent modularity in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2638–2660, 2024.

Leonardo Ranaldi. Survey on the role of mechanistic interpretability in generative ai. *Big Data and Cognitive Computing*, 9(8):193, 2025.

Farnoush Rezaei Jafari, Oliver Eberle, Ashkan Khakzar, and Neel Nanda. Relp: Faithful and efficient circuit discovery via relevance patching. In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025.

Raanan Y Rohekar, Yaniv Gurwicz, and Shami Nisimov. Causal interpretation of self-attention in pre-trained transformers. *Advances in Neural Information Processing Systems*, 36:31450–31465, 2023.

Naomi Saphra and Sarah Wiegreffe. Mechanistic? In *The 7th BlackboxNLP Workshop*, 2024.

Sofia Serrano and Noah A Smith. Is attention interpretable? In *Annual Meeting of the Association for Computational Linguistics*, 2019.

Alan Sun. Circuit stability characterizes language model generalization. In *ACL (1)*, pp. 9025–9040, 2025.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 407–416, 2024.

Martin Tutek and Jan Šnajder. Staying true to your word: (how) can attention become explanation? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pp. 131–142, 2020.

Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. In *BlackboxNLP@ACL*, 2019.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 5797–5808, 2019.

George Wang, Jesse Hoogland, Stan van Wingerden, Zach Furman, and Daniel Murfet. Differentiation and specialization of attention heads via the refined local learning coefficient. In *The Thirteenth International Conference on Learning Representations*, 2025.

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. In *The Eleventh International Conference on Learning Representations*, 2023.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Conference on Empirical Methods in Natural Language Processing*, 2019.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024.

Lin Zhang, Wenshuo Dong, Zhuoran Zhang, Shu Yang, Lijie Hu, Ninghao Liu, Pan Zhou, and Di Wang. EAP-GP: Mitigating saturation effect in gradient-based automated circuit identification. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Lin Zhang, Lijie Hu, and Di Wang. Mechanistic unveiling of transformer circuits: Self-influence as a key to model reasoning. In *North American Chapter of the Association for Computational Linguistics*, 2025b.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. *Advances in neural information processing systems*, 36:27223–27250, 2023.

A Appendix

A.1 Additional Methodological and Theoretical Details

This appendix provides additional formal details for the Causal Darwinian Head Selection (CDHS) framework introduced in Section 3. These results support the diagnostic interpretation of head-level causal fitness, ordered counterfactual evaluation, and class-conditional specialization. They are not required for implementing the main evaluation procedure.

A.1.1 Deterministic Interventional Functional Graph

CDHS adopts a mechanistic interpretability perspective in which attention heads are treated as intervenable sub-computations of a fixed trained Transformer. Recall from Section 3.1 that the model can be written as

$$f(x) = g\big(x, h_1(x), h_2(x), \ldots, h_N(x)\big), \tag{15}$$

where h_i denotes the output of attention head i, and $g(\cdot)$ represents the remaining computation of the network, including residual connections, normalization layers, MLP blocks, and downstream layers.

We model this computation as a deterministic interventional functional graph. Let $(X, Y) \sim \mathcal{D}$ denote an input–label pair drawn from the fixed evaluation distribution. For each attention head $h_i \in \mathcal{H}$, define the structural assignment

$$H_i := h_i(X), \tag{16}$$

where $h_i(\cdot)$ is a deterministic sub-computation parameterized by the fixed model parameters θ. The model output is then

$$f(X) = g\big(X, H_1, \ldots, H_N\big). \tag{17}$$

Randomness enters the framework only through sampling (X, Y) from $\mathcal{D}$. Conditional on X and fixed parameters θ, all computations are deterministic. This enables direct evaluation of counterfactual interventions through repeated forward passes without retraining or modifying model parameters.

Assumption A.1 (Approximate Modular Decomposition)**.** Each attention head h_i constitutes an intervenable sub-computation whose output can be counterfactually modified while the remaining model parameters and computations are held fixed.

This assumption is approximate because residual mixing, normalization, and downstream nonlinearities couple the effects of different heads. Nevertheless, it provides a tractable unit of intervention for head-level causal diagnostics.

A.1.2 Well-Defined Head-Level Interventions

A head-level diagnostic intervention corresponds to replacing the output of a selected attention head by the zero vector before residual addition. For a single head h_i, this intervention can be written as

$$\mathrm{do}(H_i = 0). \tag{18}$$

This coincides with the diagnostic masking operator $\mathcal{A}(f, \{h_i\})$ defined in Section 3.1. The resulting interventional model is

$$f^{(i)}(x) = \mathcal{A}(f, \{h_i\})(x). \tag{19}$$

More generally, for a subset of heads $S \subseteq \mathcal{H}$, the intervened model is

$$f^{(S)}(x) = \mathcal{A}(f, S)(x). \tag{20}$$

Lemma A.2 (Interventional Closure)**.** *The class of models obtained through diagnostic head masking preserves determinism and acyclicity of the original computation graph.*

Proof. The original Transformer computation induces a directed acyclic computation graph over intermediate variables within a forward pass. The intervention $\mathrm{do}(H_i = 0)$ replaces the structural assignment $H_i := h_i(X)$ with the constant assignment $H_i := 0$. This operation removes incoming dependencies into H_i but does not introduce new dependencies among existing variables. Since the modified assignment is still deterministic and the graph structure remains acyclic, the intervened model preserves determinism and acyclicity. □

Corollary A.3 (Invariant Evaluation Distribution)**.** *For any diagnostic head-masking intervention, the distribution over input–label pairs remains $\mathcal{D}$.*

This corollary emphasizes that CDHS changes only the internal computation of the trained model during evaluation. It does not alter the data distribution, the learned parameters, or the training procedure.

A.1.3 Identifiability of Head-Level Causal Effects

Recall that the average causal effect of head h_i is defined as

$$\mathrm{ACE}(h_i) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\ell(f^{(i)}(x), y) - \ell(f(x), y)\big]. \tag{21}$$

Theorem A.4 (Identifiability)**.** *Under the deterministic interventional graph defined above and fixed parameters θ, $\mathrm{ACE}(h_i)$ is identifiable for every $h_i \in \mathcal{H}$ from empirical diagnostic masking experiments.*

Proof. For fixed parameters θ, both the original model f and the intervened model $f^{(i)}$ are deterministic functions of the input x. The intervention $\mathrm{do}(H_i = 0)$ replaces only the structural assignment of H_i while leaving the input–label distribution unchanged. Therefore, the interventional expectation

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\ell(f^{(i)}(x), y)\big] \tag{22}$$

can be estimated directly by evaluating the masked model on samples from $\mathcal{D}$. Since both the original and intervened models are deterministic conditional on X, no additional counterfactual ambiguity arises. Hence $\mathrm{ACE}(h_i)$ is identifiable from empirical masking. □

The empirical estimator used in CDHS is

$$\widehat{\mathrm{ACE}}(h_i) = \frac{1}{K}\sum_{k=1}^{K}\Big(\ell(f^{(i)}(x_k), y_k) - \ell(f(x_k), y_k)\Big), \quad (x_k, y_k) \sim \mathcal{D}. \tag{23}$$

The corresponding empirical causal fitness is

$$\widehat{\mathcal{F}}(h_i) = \max\{\widehat{\mathrm{ACE}}(h_i), 0\}. \tag{24}$$

A.1.4 Finite-Sample Concentration and Ranking Stability

Define

$$Z_k^{(i)} := \ell(f^{(i)}(x_k), y_k) - \ell(f(x_k), y_k). \tag{25}$$

Then $\widehat{\mathrm{ACE}}(h_i) = K^{-1}\sum_{k=1}^{K} Z_k^{(i)}$.

Theorem A.5 (Uniform Concentration over Heads)**.** *Assume $\ell \in [0, 1]$. Then $Z_k^{(i)} \in [-1, 1]$, and for any $\epsilon > 0$,*

$$\Pr\left(\max_{h_i\in\mathcal{H}}\left|\widehat{\mathrm{ACE}}(h_i) - \mathrm{ACE}(h_i)\right| > \epsilon\right) \le 2|\mathcal{H}|\exp\left(-\frac{K\epsilon^2}{2}\right). \tag{26}$$

Proof. For a fixed head h_i, the random variables $\{Z_k^{(i)}\}_{k=1}^{K}$ are i.i.d. because each (x_k, y_k) is sampled independently from $\mathcal{D}$. Since $\ell \in [0, 1]$, the difference $Z_k^{(i)}$ lies in $[-1, 1]$. Applying Hoeffding's inequality gives

$$\Pr\left(\left|\widehat{\mathrm{ACE}}(h_i) - \mathrm{ACE}(h_i)\right| > \epsilon\right) \le 2\exp\left(-\frac{K\epsilon^2}{2}\right). \tag{27}$$

Taking a union bound over all $|\mathcal{H}|$ heads yields the result. □

If the evaluation loss is unbounded, such as the standard cross-entropy loss without clipping, an analogous concentration statement can be obtained under sub-Gaussian or bounded-difference assumptions on $Z_k^{(i)}$. The bounded-loss version above is sufficient to formalize the finite-sample behavior of CDHS under normalized or clipped evaluation losses.

Let $\pi^\star$ denote the population ranking induced by decreasing $\mathrm{ACE}(h_i)$, and let $\widehat{\pi}$ denote the empirical ranking induced by decreasing $\widehat{\mathrm{ACE}}(h_i)$.

Theorem A.6 (Ranking Consistency). *Suppose there exists a margin $\gamma > 0$ such that*

$$|\mathrm{ACE}(h_i) - \mathrm{ACE}(h_j)| \geq \gamma, \quad \forall i \neq j. \tag{28}$$

Then if

$$K \geq \frac{8}{\gamma^2} \log\left(\frac{2|\mathcal{H}|}{\delta}\right), \tag{29}$$

the empirical ranking $\widehat{\pi}$ matches the population causal ranking $\pi^\star$ with probability at least $1 - \delta$.

Proof. By Theorem A.5, with probability at least $1 - \delta$, all empirical estimates deviate from their true values by at most $\gamma/2$. Under this event, every pairwise ordering is preserved, and therefore the full ranking is preserved. □

Thus, the ordered counterfactual evaluation in Section 3.3 is statistically stable when causal effects are sufficiently separated and the evaluation set is large enough.

A.1.5 Ordered Counterfactual Evaluation and Non-Additivity

CDHS uses single-head causal effects to define an empirical ordering over attention heads. However, cumulative masking is not assumed to be additive.

For a set of heads $S \subseteq \mathcal{H}$, define the set-level causal effect

$$\mathrm{ACE}(S) = \mathcal{L}\big(\mathcal{A}(f, S)\big) - \mathcal{L}(f). \tag{30}$$

For the empirical top-down diagnostic set

$$S_k^{\mathrm{TD}} = \{h_{\widehat{\pi}(1)}, \ldots, h_{\widehat{\pi}(k)}\}, \tag{31}$$

the cumulative top-down loss change is

$$\Delta_k^{\mathrm{TD}} = \mathcal{L}\big(\mathcal{A}(f, S_k^{\mathrm{TD}})\big) - \mathcal{L}(f). \tag{32}$$

Lemma A.7 (Additive Baseline Interpretation). *For any k, the quantity*

$$\sum_{r=1}^{k} \mathrm{ACE}(h_{\widehat{\pi}(r)}) \tag{33}$$

corresponds to the cumulative loss change under a hypothetical additive, non-interacting causal model. Deviations from this value quantify interaction effects among attention heads.

Proof. Consider a hypothetical additive causal model in which the effect of masking a set of heads S is the sum of the individual marginal effects:

$$L_{\mathrm{add}}(S) = L(\emptyset) + \sum_{h_i \in S} \mathrm{ACE}(h_i). \tag{34}$$

In this model, masking the top-k heads induces exactly the cumulative loss change

$$\sum_{r=1}^{k} \mathrm{ACE}(h_{\widehat{\pi}(r)}). \tag{35}$$

In the true Transformer, residual streams, normalization, MLP blocks, and downstream nonlinearities may introduce redundancy, synergy, or suppressive interactions. Therefore, the actual cumulative loss change may deviate from the additive baseline. □

Theorem A.8 (Non-Additivity). *There exist models and head pairs (h_i, h_j) such that*

$$\mathrm{ACE}(\{h_i, h_j\}) \neq \mathrm{ACE}(h_i) + \mathrm{ACE}(h_j). \tag{36}$$

In particular, both sub-additive and super-additive interactions are possible.

Proof. For sub-additivity, consider two heads whose outputs are redundant in the downstream computation, for example through a dependence of the form

$$g(x, h_i, h_j) = g(x, h_i + h_j). \tag{37}$$

Masking either head alone may cause a moderate loss increase, while masking both together need not produce the sum of the two individual effects.

For super-additivity, consider two heads that compensate for each other, such that either one alone is sufficient for a downstream feature but both are required for robust prediction under masking. Then masking one head may have little effect, while masking both may cause a large degradation. Hence cumulative causal effects need not equal the sum of individual causal effects. □

Corollary A.9 (Diagnostic Nature of Ordered Masking)**.** *Ordered counterfactual evaluation identifies dominant causal contributors under the evaluated intervention, but it does not recover a globally optimal pruning set.*

This distinction is central to CDHS. The method ranks and masks heads only for evaluation-time diagnosis. It does not claim that high-fitness heads should be kept for compression, nor that low-fitness heads can be safely removed from a deployed model.

A.1.6 Concentration of Causal Fitness

A characteristic empirical signature of CDHS is a separation between top-down and bottom-up masking curves. One sufficient explanation for such separation is a concentrated distribution of positive causal fitness.

Define the population-level causal fitness as

$$\mathcal{F}(h_i) = \max\{\mathrm{ACE}(h_i), 0\}. \tag{38}$$

Let $\pi^\star$ rank heads by decreasing $\mathcal{F}(h_i)$.

Assumption A.10 (Concentrated Fitness Distribution)**.** There exist constants $\beta \in (0, 1)$ and $C > 0$ such that

$$\mathcal{F}(h_{\pi^\star(i)}) = Ci^{-\beta} + o(i^{-\beta}). \tag{39}$$

Proposition A.11 (Front-Loaded Cumulative Fitness)**.** *Under the concentrated fitness assumption,*

$$\sum_{i=1}^{k} \mathcal{F}(h_{\pi^\star(i)}) = \Theta(k^{1-\beta}). \tag{40}$$

Moreover, for $1 \le k \le N$*,*

$$\frac{\sum_{i=1}^{k} \mathcal{F}(h_{\pi^\star(i)})}{\sum_{i=1}^{N} \mathcal{F}(h_{\pi^\star(i)})} \approx \left(\frac{k}{N}\right)^{1-\beta}. \tag{41}$$

Proof. By assumption,

$$\mathcal{F}(h_{\pi^\star(i)}) = Ci^{-\beta} + o(i^{-\beta}). \tag{42}$$

Ignoring lower-order terms,

$$\sum_{i=1}^{k} i^{-\beta} \sim \int_1^k x^{-\beta}\, dx = \frac{k^{1-\beta} - 1}{1 - \beta}. \tag{43}$$

Multiplying by C yields the stated $\Theta(k^{1-\beta})$ scaling. Applying the same approximation to the denominator with upper limit N gives the cumulative fraction. □

This result formalizes the intuition that when causal fitness is concentrated among a relatively small number of heads, top-down masking produces a sharper degradation curve than bottom-up masking. Conversely, when causal fitness is diffuse, the top-down and bottom-up curves are expected to differ less strongly.

A.1.7 Counterfactual Selection States and Survival Indicators

To connect causal attribution with the Darwinian terminology used in CDHS, we define an analytic threshold-based selection trajectory. This construction is interpretive only and does not correspond to training-time evolution, parameter updates, or architectural pruning.

Definition A.12 (Counterfactual Selection State)**.** At diagnostic threshold $\tau^{(t)}$, the counterfactual selection state is

$$\mathcal{H}^{(t)} = \left\{ h_i \in \mathcal{H} : \mathcal{F}(h_i) > \tau^{(t)} \right\}, \tag{44}$$

where $\tau^{(t)}$ is a non-decreasing sequence of thresholds.

The set $\mathcal{H}^{(t)}$ describes the heads whose causal fitness exceeds a given diagnostic threshold. It is an analytic summary of the fixed trained model under the fixed evaluation environment $(\mathcal{D}, \theta)$.

Definition A.13 (Causal Survival Indicator)**.** At threshold step t, the causal survival status of head h_i is defined as

$$\mathbb{I}^{(t)}_{\text{surv}}(h_i) = \begin{cases} 1, & \text{if } h_i \in \mathcal{H}^{(t)}, \\ 0, & \text{otherwise.} \end{cases} \tag{45}$$

Proposition A.14 (Fitness–Survival Monotonicity)**.** *Assume that $\tau^{(t)}$ is non-decreasing in t. If*

$$\mathcal{F}(h_i) > \mathcal{F}(h_j), \tag{46}$$

then for every threshold step t,

$$\mathbb{I}^{(t)}_{\text{surv}}(h_i) \geq \mathbb{I}^{(t)}_{\text{surv}}(h_j). \tag{47}$$

Proof. If $\mathcal{F}(h_i) > \mathcal{F}(h_j)$, then any threshold exceeded by h_j is also exceeded by h_i. Therefore, whenever h_j belongs to $\mathcal{H}^{(t)}$, h_i also belongs to $\mathcal{H}^{(t)}$. This proves the monotonicity of the survival indicator. □

This threshold-based survival notion is distinct from top-down masking. In top-down masking, high-fitness heads are deliberately removed first to test their causal importance. In threshold-based survival analysis, high-fitness heads remain above increasing causal-fitness thresholds for longer.

A.1.8 Class-Conditional Fitness and Specialization

For class $c \in \mathcal{Y}$, recall the class-conditional causal effect from Section 3.4:

$$\text{ACE}_c(h_i) = \mathbb{E}_{(x,y)\sim\mathcal{D}|y=c}\big[\ell(f^{(i)}(x), y) - \ell(f(x), y)\big]. \tag{48}$$

This quantity measures the causal contribution of head h_i within the class-conditioned environment $Y = c$.

Definition A.15 (Class-Conditional Fitness)**.** For class $c \in \mathcal{Y}$, define

$$\mathcal{F}_c(h_i) = \max\{\text{ACE}_c(h_i), 0\}. \tag{49}$$

Under the ecological interpretation of CDHS, classes define diagnostic niches, and attention heads may exhibit niche-specific causal utility. A head with high $\mathcal{F}_c(h_i)$ for a particular class contributes disproportionately to performance in that class-conditioned environment.

The specialization score is

$$\text{Spec}(h_i) = \frac{\max_c \max(\text{ACE}_c(h_i), 0)}{\sum_{c'} \max(\text{ACE}_{c'}(h_i), 0) + \varepsilon}, \tag{50}$$

where $\varepsilon > 0$ is a small numerical constant.

Theorem A.16 (Specialization Identifiability)**.** *Assume the class-conditional distributions*

$$\mathcal{D}_c = \mathcal{D}(\cdot \mid Y = c) \tag{51}$$

are well-defined. Then $\text{ACE}_c(h_i)$ *and* $\text{Spec}(h_i)$ *are identifiable from class-conditioned empirical masking experiments.*

Proof. The argument follows the identifiability proof for $\mathrm{ACE}(h_i)$, restricted to samples from $\mathcal{D}_c$. Since the intervention modifies only the internal head output and leaves the class-conditioned input distribution unchanged, $\mathrm{ACE}_c(h_i)$ can be estimated by evaluating the original and masked models on samples with $y = c$. The specialization score is a deterministic function of the collection of identifiable class-conditional causal effects, and is therefore identifiable. □

Proposition A.17 (Bounded Specialization)**.** *For every head h_i,*

$$\mathrm{Spec}(h_i) \in [0, 1]. \tag{52}$$

Proof. The numerator is the maximum of a collection of non-negative terms. The denominator is the sum of all terms in the same collection plus $\varepsilon > 0$. Therefore, the ratio is non-negative and at most one. □

A high specialization score indicates that the positive causal contribution of a head is concentrated on a small number of classes. A low specialization score indicates that its positive contribution is more broadly distributed across classes.

A.1.9 Well-Posedness and Stability of Causal Attribution

Theorem A.18 (Existence and Finiteness of ACE)**.** *Assume the loss function ℓ is bounded and measurable, and $f(\cdot;\theta)$ is measurable for fixed θ. Then for every attention head $h_i \in \mathcal{H}$, the average causal effect $\mathrm{ACE}(h_i)$ exists, is finite, and is uniquely defined.*

Proof. By assumption, $\ell \circ f$ and $\ell \circ f^{(i)}$ are bounded measurable functions of (x, y). Since $(X, Y) \sim \mathcal{D}$ is defined on a probability space, both expectations

$$\mathbb{E}[\ell(f(X), Y)] \quad \text{and} \quad \mathbb{E}[\ell(f^{(i)}(X), Y)] \tag{53}$$

exist and are finite. Linearity of expectation yields existence and uniqueness of $\mathrm{ACE}(h_i)$. □

Corollary A.19 (Well-Defined Fitness Landscape)**.** *The induced fitness values*

$$\mathcal{F}(h_i) = \max\{\mathrm{ACE}(h_i), 0\} \tag{54}$$

define a non-negative, finite-valued fitness landscape over the discrete population of attention heads $\mathcal{H}$.

Assumption A.20 (Lipschitz Stability of Loss)**.** There exists $L > 0$ such that for all x, y and all model outputs $p, q \in \Delta(\mathcal{Y})$,

$$|\ell(p, y) - \ell(q, y)| \le L\|p - q\|_1. \tag{55}$$

Theorem A.21 (Stability under Small Interventions)**.** *Let h_i be an attention head whose output magnitude is bounded:*

$$\|h_i(x)\| \le M \quad \textit{for all } x. \tag{56}$$

Assume that the downstream computation g is C_g-Lipschitz with respect to the intervention site of head h_i. Then

$$|\mathrm{ACE}(h_i)| \le L \cdot \mathbb{E}_{X\sim\mathcal{D}}\big[\|f^{(i)}(X) - f(X)\|_1\big] \le LC_gM. \tag{57}$$

Proof. The first inequality follows directly from the Lipschitz continuity of ℓ:

$$\Big|\ell(f^{(i)}(X), Y) - \ell(f(X), Y)\Big| \le L\|f^{(i)}(X) - f(X)\|_1. \tag{58}$$

Taking expectations gives the first bound. The second bound follows from the Lipschitz continuity of the downstream computation g with respect to the head output and from the boundedness of $h_i(X)$. □

This result ensures that causal effects remain controlled under bounded head outputs and stable downstream computation, preventing pathological domination by numerically unstable heads.

A.1.10 SUMMARY

The main text presents CDHS as a practical diagnostic pipeline: mask attention heads at evaluation time, estimate head-level causal fitness, perform ordered counterfactual evaluation, and analyze class-conditional specialization. This appendix provides the corresponding formal interpretation. Diagnostic head masking is a well-defined counterfactual intervention under fixed parameters; head-level causal effects are identifiable from empirical masking; finite-sample estimates concentrate under standard boundedness assumptions; ordered masking is diagnostic rather than additive or pruning-oriented; and the Darwinian interpretation can be formalized through causal fitness, threshold-based selection states, survival indicators, and class-conditioned niches.

A.2 MORE EXPERIMENTS

A.2.1 EXPERIMENT SETUP

The experiments were conducted on an institutional high-performance computing cluster designed to support high-performance and high-throughput computational workloads. The facility provides large-scale CPU and GPU-accelerated computing resources, including multi-core CPU nodes, high-memory compute nodes, and GPU-enabled nodes. It also provides shared parallel storage and large-capacity storage for data-intensive workloads. The cluster runs a Linux-based operating system, and job submission and scheduling are managed using the SLURM Workload Manager.

A.2.2 MORE EXPERIMENTS ON CIFAR-10

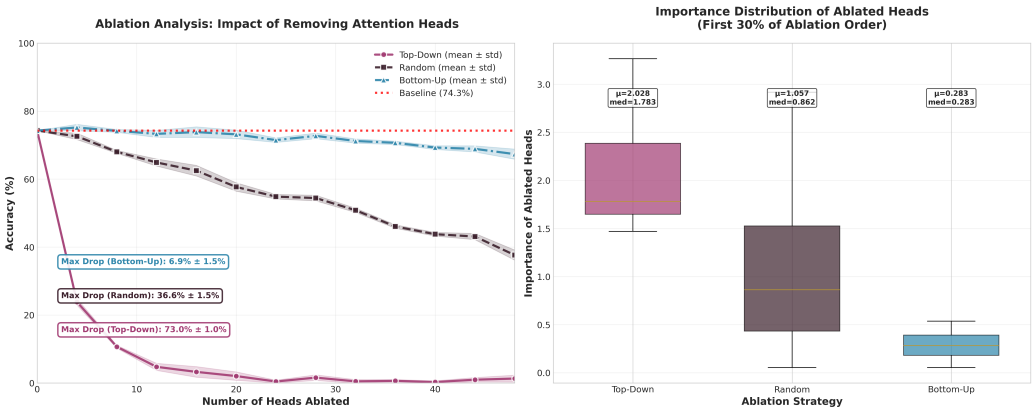


Figure 6: Causal Darwinian Head Selection (CDHS) under progressive attention head ablation. Left: Test accuracy as a function of the number of ablated attention heads for top-down (highest estimated causal fitness first), random, and bottom-up (lowest estimated causal fitness first) removal, averaged over multiple seeds with standard deviation. Top-down ablation induces a rapid collapse in performance, while bottom-up ablation leaves accuracy largely intact, indicating a Darwinian-like selection pressure in which a small subset of heads exerts disproportionately large causal impact on model performance. Right: Fitness distributions of the first 30% of ablated heads under each strategy, showing that heads eliminated early in the top-down order exhibit substantially higher estimated causal fitness than those removed randomly or last.

Causal Impact of Attention Head Ablation Figure 6 left shows model accuracy as a function of the number of eliminated attention heads under the three CDHS selection regimes. Under top-down selection, model performance collapses rapidly: removing only a small fraction of the most causally important heads reduces accuracy from 74.3% to near-zero levels. In contrast, random drift results in a gradual decline in accuracy, while bottom-up selection preserves performance even after a substantial number of heads have been removed. This divergence indicates that attention heads are not functionally interchangeable; rather, overall model performance is bottlenecked by a small subset of causally dominant heads.

Figure 6 right reports the distribution of estimated causal fitness scores for the first 30% of heads eliminated under each selection regime. Heads removed under top-down selection exhibit significantly

higher mean and median fitness than those removed under random drift or bottom-up selection. This result demonstrates that the observed performance collapse is driven by targeted removal of highly influential components rather than by cumulative capacity reduction, revealing a sparse and highly uneven distribution of causal fitness across attention heads.

Class-Conditional Effects of Causally Important Head Ablation To further probe the functional role of dominant heads, we analyze the effect of eliminating the single highest-fitness ("winner") attention head identified by CDHS.

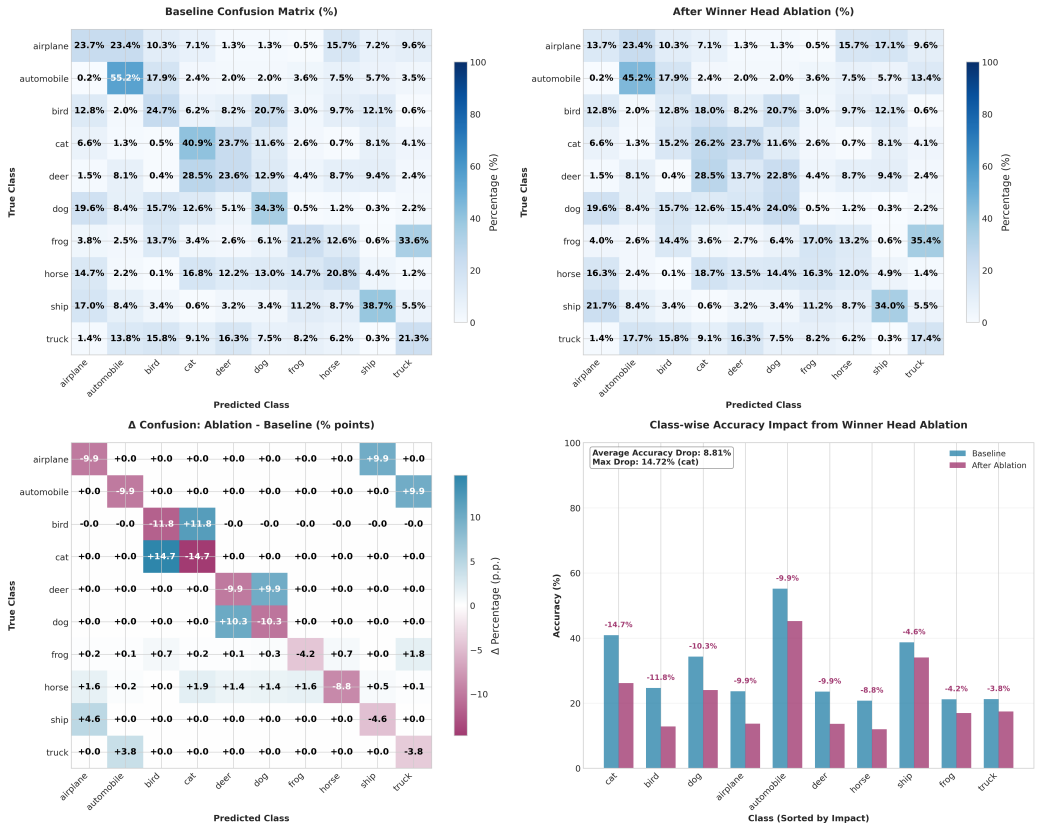


Figure 7: Class-wise causal effects of winner head ablation. Top: Confusion matrices before and after ablating the most causally important ("winner") attention head, revealing structured shifts in misclassification patterns rather than uniform degradation. Bottom left: Difference between post-ablation and baseline confusion matrices, exposing class-specific redistribution of prediction errors. Bottom right: Per-class accuracy before and after ablation, sorted by impact. The uneven degradation across classes suggests that attention heads differ substantially in their causal contributions to class discrimination, with certain class distinctions critically dependent on a small number of highly important heads.

Confusion Matrix Shift. Figure 7 top compares the baseline confusion matrix with that obtained after winner head ablation. The resulting changes are highly structured: certain classes exhibit pronounced drops in correct classification, while others show comparatively smaller changes. Moreover, misclassifications are redirected toward specific alternative classes rather than being uniformly distributed.

Differential Confusion Analysis. The difference between post-ablation and baseline confusion matrices (Figure 7, bottom left) reveals systematic rerouting of predictions across particular class pairs. These structured shifts suggest that the eliminated head contributed to class-discriminative computations that are not redundantly encoded by the remaining heads.

Class-wise Accuracy Impact. Figure 7 bottom right reports class-wise accuracy before and after winner head ablation. The average accuracy drop across classes is 8.81%, with a maximum drop

of 14.72% observed for the *cat* class, further supporting the presence of class-specific functional dependencies.

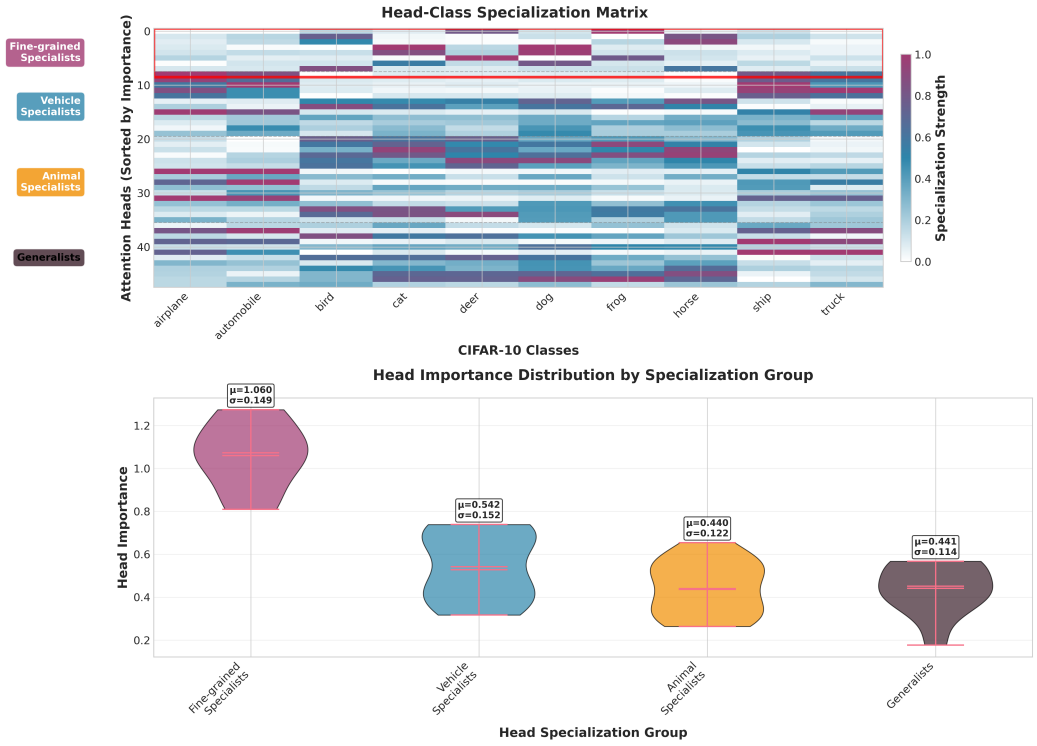


Figure 8: Emergent functional niches of attention heads identified via causal ablation. Top: Head–class specialization matrix, where rows correspond to attention heads (sorted by causal fitness as measured by accuracy degradation under ablation) and columns to semantic classes; color intensity reflects class-specific causal contribution. Distinct block structures indicate functional specialization, with individual heads exhibiting strong class-selective causal roles. Bottom: Distributions of causal fitness grouped by specialization type (fine-grained specialists, vehicle specialists, animal specialists, and generalists). Specialized heads exhibit higher causal fitness than generalists, suggesting heterogeneous and non-redundant functional roles across attention heads.

Causal Functional Specialization of Attention Heads **Head–Class Specialization Matrix.** Figure 8 top visualizes the normalized class-conditional specialization strength of each attention head across CIFAR-10 classes. The resulting matrix reveals clear specialization patterns: some heads exhibit narrow, high-intensity responses to specific classes, while others display broader, lower-magnitude contributions.

Emergent Functional Niches. Based on their specialization profiles, attention heads can be grouped into interpretable functional niches, including fine-grained specialists, vehicle specialists, animal specialists, and generalists. These niches emerge without explicit supervision, suggesting that causal selection induced by CDHS reveals latent modular structure that forms during standard training.

Fitness Distribution by Niche. Figure 8 bottom reports the distribution of fitness scores across specialization groups. Fine-grained specialists exhibit the highest average fitness with low variance, indicating consistent causal fitness. In contrast, generalist heads show lower average fitness, suggesting a more redundant or distributed functional role and, as corroborated by bottom-up ablation results, greater robustness to elimination.

Causal Head Specialization Emerges During Training To distinguish training-time specialization from a purely post hoc ranking of heads at the final checkpoint, we measure head-level causal fitness throughout CIFAR-10 training. Models are trained under two matched supervision regimes: standard real-label training and a label-permutation control. The control preserves the image distribution,

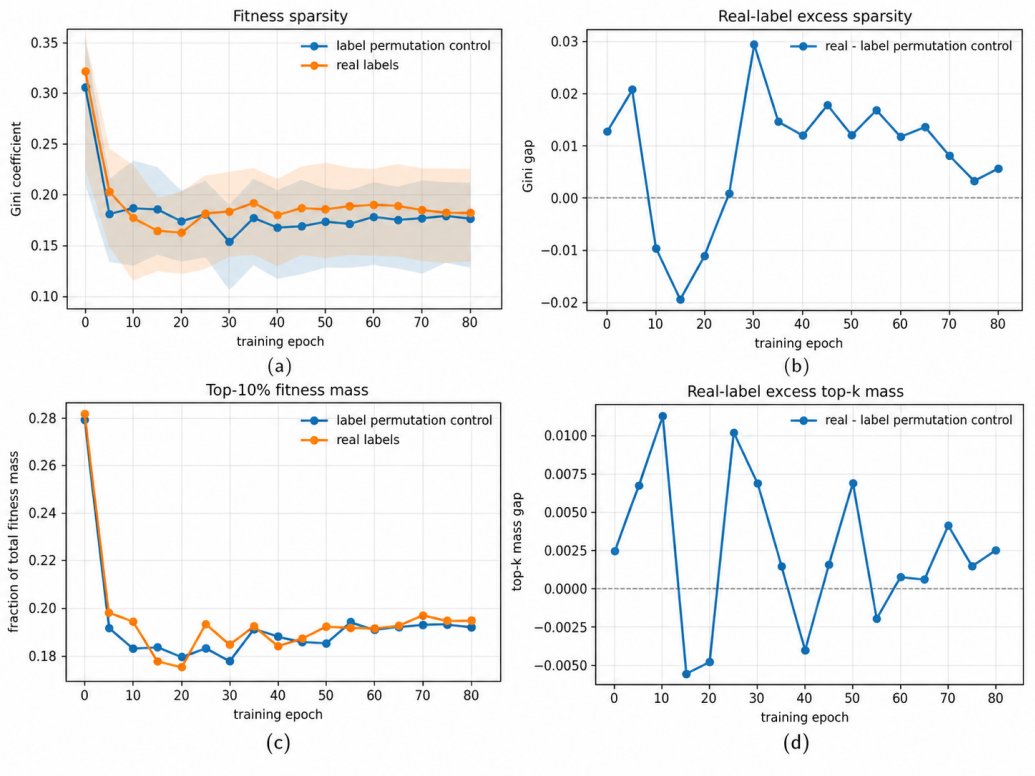


Figure 9: Training-time emergence of causal head specialization. We compare head-level causal fitness throughout CIFAR-10 training under real-label supervision and a label-permutation control. (a) Real-label training yields higher head-fitness sparsity than the permutation control across much of training. (b) The real-minus-control Gini gap is predominantly positive after the initial transient phase, indicating that semantic supervision induces additional concentration of causal importance. (c) The top-10% fitness mass is similar across conditions, suggesting that specialization is not driven solely by a small set of heads. (d) The excess top-k mass is weaker and more variable than the sparsity effect.

architecture, optimizer, and training schedule, while removing the semantic correspondence between images and labels. Differences between the two conditions therefore isolate the effect of task-aligned supervision beyond generic optimization dynamics and memorization pressure.

At each checkpoint, we estimate the causal fitness of every attention head by intervening on that head and measuring the resulting change in validation performance. We summarize the distribution of head fitness using two concentration statistics. The Gini coefficient measures the global sparsity of causal fitness across all heads, while the top-10% fitness mass measures the fraction of total causal fitness assigned to the highest-ranked heads. For both statistics, we also report the real-label excess over the label-permutation control, directly measuring the additional specialization induced by semantically meaningful supervision.

Figure 9 shows that real-label supervision produces a more structured allocation of causal importance across attention heads during training. In the raw Gini trajectories, both conditions undergo a sharp early reorganization, but the real-label model subsequently maintains higher fitness sparsity across much of training than the permutation control (Fig. 9a). This effect is clearer in the real-minus-control trajectory, where the excess Gini is predominantly positive after the initial transient phase (Fig. 9b). Thus, the emergence of high- and low-fitness heads is not fully explained by the architecture or optimizer alone; it is amplified when the labels preserve semantic structure.

The top-10% fitness-mass results provide a complementary but weaker signal. The raw trajectories are close between the two conditions, and the excess top-k mass fluctuates around zero with small positive deviations at several checkpoints (Fig. 9c,d). This suggests that the specialization is not dominated

by a single small set of heads that immediately captures most of the causal contribution. Instead, the stronger and more stable Gini effect indicates a broader redistribution of causal fitness across the head population, with real-label training producing a more uneven and organized importance profile.

Together, these results support the interpretation that attention-head specialization develops during learning rather than being merely discovered by sorting heads at the final checkpoint. On CIFAR-10, semantic supervision induces a training-time organization of causal importance across heads, most clearly reflected in the persistent excess sparsity of the real-label model relative to the label-permutation control.

A.2.3 More Experiments on CIFAR-100

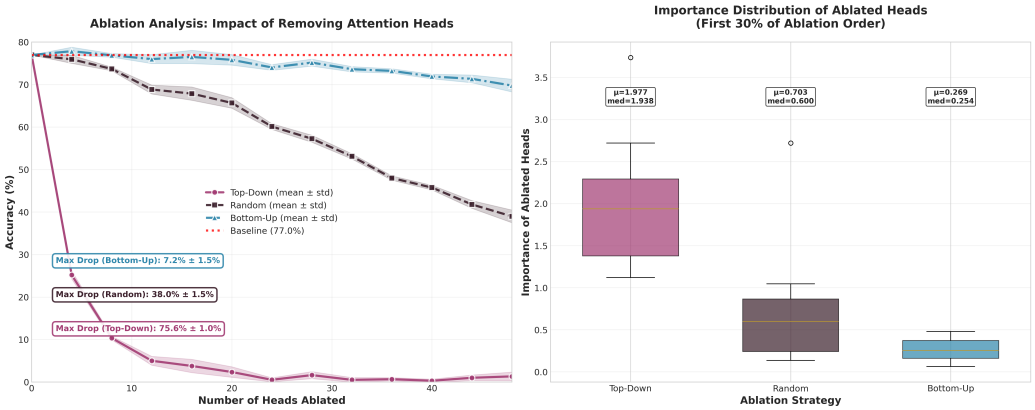


Figure 10: Causal Darwinian Head Selection (CDHS) under progressive attention head ablation on CIFAR-100. Left: Top-1 test accuracy as a function of the number of ablated attention heads for top-down (highest causal fitness first), random, and bottom-up (lowest causal fitness first) removal, averaged over multiple seeds with standard deviation. Top-down ablation induces a rapid collapse in performance despite the large number of heads, while bottom-up ablation preserves accuracy even after substantial removal, indicating strong Darwinian selection pressure in which a small subset of heads dominates causal fitness in fine-grained recognition. Right: Fitness distributions of the first 30% of ablated heads under each strategy, showing that heads eliminated early in the top-down order possess substantially higher causal fitness than those removed randomly or last.

Causal Impact of Attention Head Ablation Figure 10 left reports the accuracy as a function of the number of ablated attention heads on CIFAR-100 under three CDHS regimes. Similar to CIFAR-10, top-down elimination induces a rapid collapse in model performance: removing a small fraction of the highest-fitness heads drives accuracy sharply downward from 77.0% toward near-zero levels. In contrast, random ablation yields a gradual degradation, while bottom-up elimination preserves accuracy even after a large proportion of low-fitness heads are removed. The increased class granularity of CIFAR-100 amplifies this separation between regimes. In particular, the steepness of the top-down curve indicates that a very small subset of heads constitutes a critical causal bottleneck for fine-grained recognition, whereas the majority of heads contribute only marginally to global fitness.

Figure 10 right shows the fitness distribution of the first 30% of ablated heads under each regime. Heads eliminated by top-down selection have substantially higher mean and median fitness than those removed under random or bottom-up strategies. Bottom-up ablation predominantly removes heads with near-zero fitness, explaining the minimal accuracy loss observed in this regime. Together, these results demonstrate that performance degradation is not driven by uniform capacity reduction, but by targeted removal of a sparse set of causally dominant heads, revealing a highly skewed fitness landscape over attention heads in CIFAR-100.

Class-Conditional Effects of Causally Important Head Ablation To probe class-conditional dependencies, we ablate the single highest-fitness (”winner”) attention head identified by CDHS and evaluate its impact on CIFAR-100 predictions.

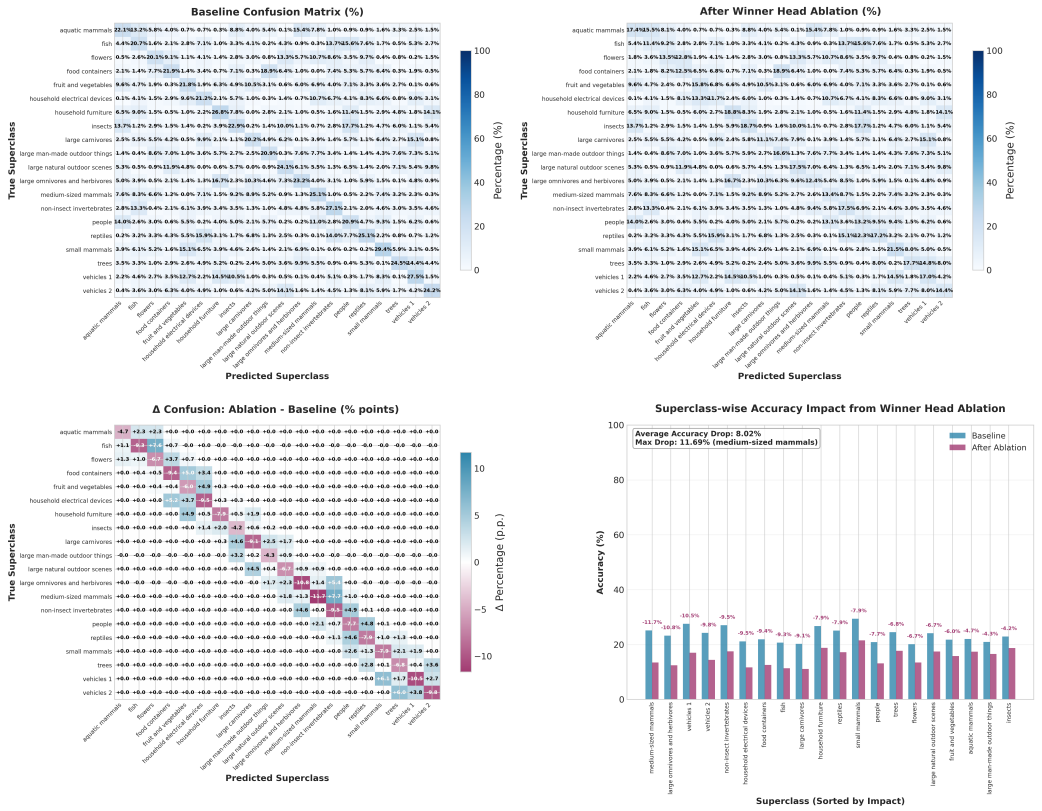


Figure 11: Superclass-wise Darwinian survival signatures induced by Causal Darwinian Head Selection on CIFAR-100. Top: Confusion matrices before and after ablating the most causally fit ("winner") attention head, revealing structured and superclass-specific shifts in misclassification patterns rather than uniform degradation. Bottom left: Difference between post-ablation and baseline confusion matrices, exposing systematic redistribution of prediction errors between semantically related superclasses. Bottom right: Superclass-wise accuracy before and after ablation, sorted by impact. The uneven degradation across superclasses indicates that Darwinian selection at the head level induces non-uniform survival of semantic discriminations, with certain fine-grained category distinctions critically dependent on a small number of high-fitness heads.

Confusion Matrix Shift. Figure 11 top compares the baseline confusion matrix with that obtained after winner head ablation. The resulting degradation is highly structured: specific superclasses exhibit pronounced drops in correct classification, while others are comparatively stable. Errors are not diffused uniformly across classes but are redirected toward semantically related superclasses.

Differential Confusion Analysis. The difference matrix (Figure 11, bottom left) highlights systematic rerouting of predictions between particular superclass pairs. These structured shifts indicate that the removed head implemented class-discriminative computations that are not redundantly represented elsewhere in the model.

Superclass-wise Fitness Impact. Figure 11 bottom right reports superclass-wise accuracy before and after ablation. The average accuracy drop across CIFAR-100 superclasses is 8.02%, with a maximum drop of 11.69% observed for *medium-sized mammals*. This heterogeneity confirms that dominant heads encode fine-grained, class-specific features whose removal selectively impairs related categories.

Causal Functional Specialization of Attention Heads Head–Superclass Specialization Matrix. Figure 12 top visualizes the normalized contribution of each attention head to each CIFAR-100 superclass. The matrix reveals clear specialization patterns despite the increased task complexity:

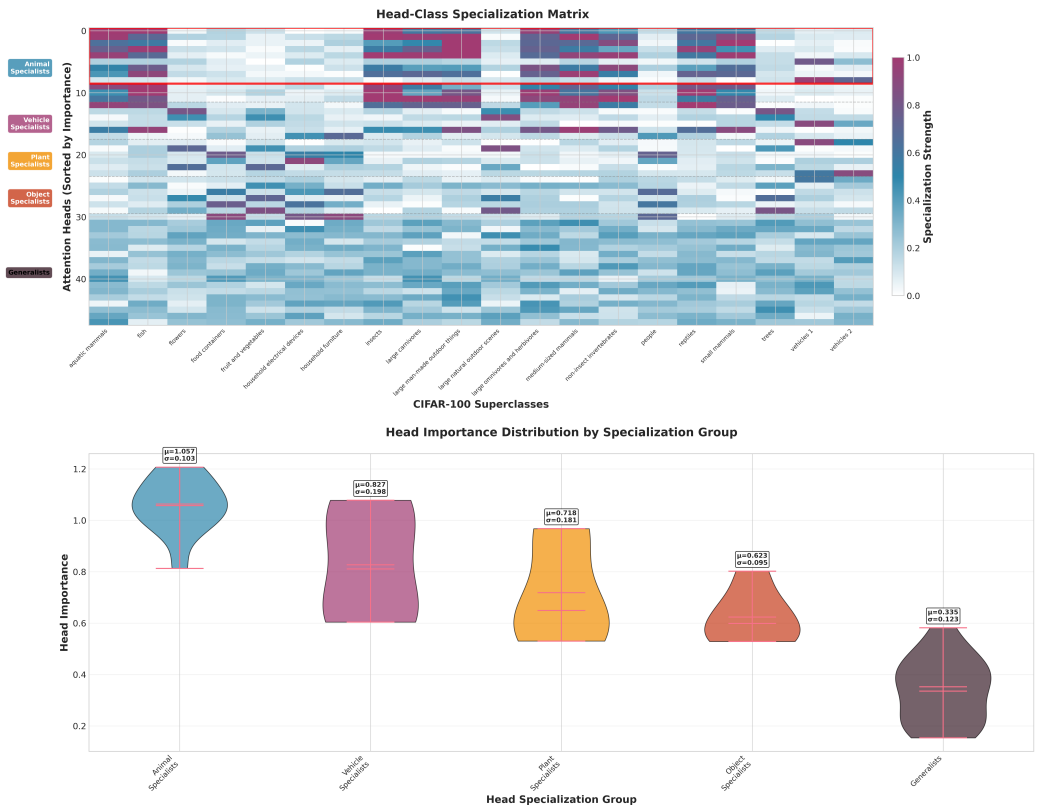


Figure 12: Emergent functional niches of attention heads under Causal Darwinian Head Selection (CDHS) on CIFAR-100. Top: Head–superclass specialization matrix, where rows correspond to attention heads (sorted by causal fitness) and columns to CIFAR-100 superclasses; color intensity reflects superclass-specific causal contribution. Distinct block structures reveal Darwinian niche formation, in which heads specialize for particular semantic domains such as animals, vehicles, plants, and objects. Bottom: Distributions of causal fitness grouped by specialization type (animal specialists, vehicle specialists, plant specialists, object specialists, and generalists). Heads occupying specialized functional niches exhibit higher causal fitness than generalists, consistent with a Darwinian interpretation of attention heads as a population of competing computational units shaped by fine-grained task-driven selection pressure.

a subset of heads exhibits strong, localized responses to particular semantic groups (e.g., animals, vehicles, plants), while others contribute weakly but broadly across many superclasses.

Emergent Functional Niches. Clustering heads by their specialization profiles yields interpretable functional niches, including animal specialists, vehicle specialists, plant specialists, object specialists, and generalists. These niches emerge without explicit architectural constraints or supervision, suggesting that CDHS exposes latent modular organization shaped by evolutionary pressure on model fitness.

Fitness Distribution by Niche. Figure 12 bottom shows the distribution of fitness scores across specialization groups. Highly specialized heads exhibit the largest average fitness and relatively low variance, indicating consistent causal fitness. Generalist heads, while more numerous, have lower average fitness and are substantially more robust to elimination, consistent with their redundant or supportive role.

Causal Head Specialization Emerges During Training To test whether head importance is merely a post hoc ranking property of the final model, or instead reflects an organization that emerges during learning, we track causal head fitness throughout training on CIFAR-100. At each checkpoint, we compute a fitness score for every attention head using the same causal intervention used in our

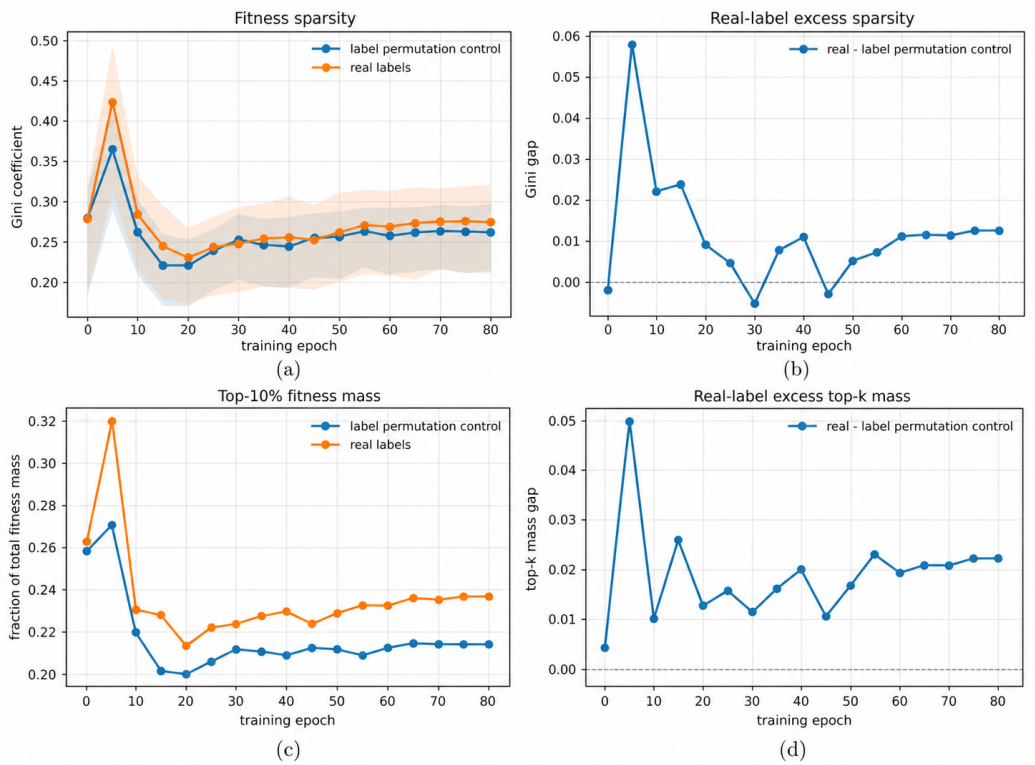


Figure 13: Training-time emergence of causal head specialization on CIFAR-100. We compare real-label training with a label-permutation control using the same architecture, data distribution, optimizer, and training schedule. At each checkpoint, we compute causal head fitness and summarize its concentration using the Gini coefficient and the top-10% fitness mass. Real-label training yields higher head-fitness sparsity and a larger fraction of fitness mass concentrated in the top heads than the permutation control. The real-minus-control excess is positive for most of training, indicating that causal head specialization is not merely a post-hoc ordering of the final model, but progressively emerges during learning under meaningful supervision.

main analysis, and summarize the resulting distribution by two complementary statistics: the Gini coefficient of head fitness, which measures global sparsity, and the fraction of total fitness mass assigned to the top 10% of heads, which measures concentration among the most causally influential heads. As a control, we train the same architecture with the same optimizer, schedule, and input distribution, but under a label-permutation condition. This control preserves the image statistics and optimization dynamics while removing the semantic structure of the supervision signal.

Figure 13 shows that real-label training consistently induces a more concentrated causal-fitness distribution than the label-permutation control. In early training, both conditions exhibit a transient increase in sparsity, suggesting that some degree of head imbalance can arise from generic optimization dynamics. However, the real-label model displays a substantially larger early peak and, more importantly, maintains a persistent positive excess over the control throughout most of training. The same pattern appears in both the Gini coefficient and the top-10% fitness mass: under real labels, a small subset of heads accounts for a larger fraction of the total causal contribution than in the permutation baseline. The excess plots further show that this difference is not confined to the final checkpoint, but is already visible during training and remains positive after the initial transient phase.

These results indicate that causal head specialization is a training-time phenomenon. Since the comparison isolates the effect of meaningful label structure from architecture- and optimizer-induced sparsity, the additional concentration observed under real labels suggests that supervised learning progressively organizes attention heads into a sparse causal structure. Thus, the most important heads in the final model are not simply identified by sorting a static set of heads after training; rather, their

disproportionate causal influence emerges and stabilizes as the model learns task-relevant structure from CIFAR-100.

Across all CIFAR-100 experiments, Causal Darwinian Head Selection reveals a sparse and hierarchical organization of attention heads. Model performance is dominated by a small number of high-fitness, highly specialized heads, while the majority occupy redundant or low-impact niches. Compared to CIFAR-10, this structure is even more pronounced, reflecting the increased demands of fine-grained visual recognition. These findings support a mechanistic, evolutionary interpretation of attention heads as differentiated computational units subject to strong selection effects within the Transformer architecture.

A.2.4 More Experiments on Tiny-ImageNet

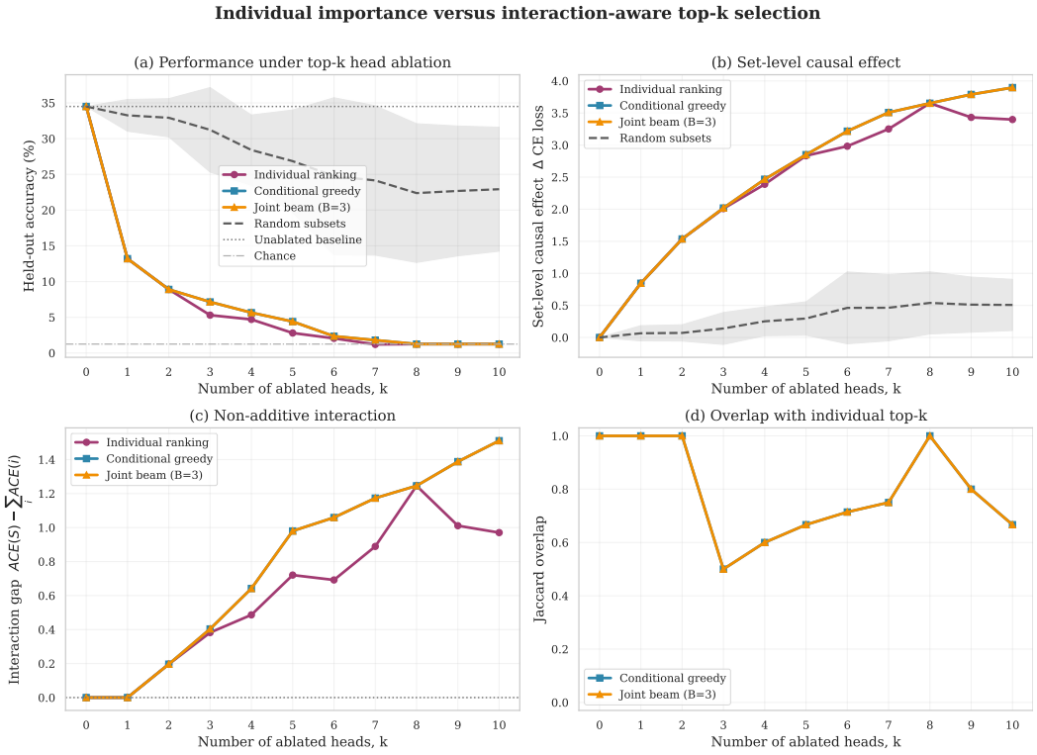


Figure 14: Individual importance versus interaction-aware top-k head selection on Tiny-ImageNet. We compare individual ranking, conditional greedy selection, and joint beam search with beam width $B = 3$ as the number of ablated heads k increases. (a) Held-out classification accuracy after ablation. The dotted horizontal line denotes the unablated baseline, and the dash-dotted line denotes the plotted chance reference. (b) Set-level ablation effect, $\mathrm{ACE}(S) = \mathcal{L}\mathrm{CE}(f^{(-S)}) - \mathcal{L}\mathrm{CE}(f)$, measured as the increase in cross-entropy relative to the unablated model. (c) Non-additive interaction gap, $\mathrm{ACE}(S) - \sum_{i \in S} \mathrm{ACE}(i)$; positive values indicate super-additive loss increases under joint ablation. (d) Jaccard overlap between each interaction-aware subset and the individual top-k subset, with the empty-set overlap at $k = 0$ assigned a value of one. Gray dashed curves and shaded bands show the random-subset reference and its variability. Conditional greedy selection and joint beam search exhibit nearly overlapping curves.

Individual Importance and Interaction-Aware Selection To complement the head-ranking analysis in Section 4.1, we investigate whether a fixed ranking of individual units adequately characterizes the functional effects of ablating them jointly. Figure 14 compares individual ranking, conditional greedy selection, and joint beam search with beam width $B = 3$ in a Tiny-ImageNet head-ablation experiment. Random subsets of the same cardinality serve as a reference. Let S denote a set of ablated heads and $f^{(-S)}$ the model under the corresponding intervention. We quantify the set-level ablation effect using the increase in evaluation cross-entropy:

$$\mathrm{ACE}(S) = \mathcal{L}_{\mathrm{CE}}\Big(f^{(-S)}\Big) - \mathcal{L}_{\mathrm{CE}}(f).$$

To assess departures from additivity, we further measure

$$\Delta_{\mathrm{int}}(S) = \mathrm{ACE}(S) - \sum_{i \in S} \mathrm{ACE}(\{i\}).$$

A positive interaction gap indicates that joint ablation increases the loss more than the sum of the corresponding single-head ablation effects. Figure 14(a) shows that all three targeted selection strategies produce substantially greater accuracy degradation than random ablation, indicating that they identify functionally influential heads. The accuracy curves nevertheless provide limited separation among the targeted strategies once accuracy becomes very low. The loss-based measurements reveal additional differences. At $k = 10$, conditional greedy selection and joint beam search yield a set-level effect of approximately 3.9, compared with 3.4 for individual ranking (Figure 14(b)). Their interaction gaps are approximately 1.5, compared with 1.0 for individual ranking (Figure 14(c)), indicating stronger departures from an additive account of the ablation effects. Figure 14(d) further shows that interaction-aware selection can identify different subsets: its Jaccard overlap with the individual top-k set is approximately 0.50 at $k = 3$ and 0.67 at $k = 10$. Conditional greedy selection and beam search produce nearly coincident curves across the reported measurements. These observations show that individual importance rankings can identify influential units while leaving aspects of their joint contribution unaccounted for. For the Darwinian framework, this motivates examining how unit-level fitness translates into the functional contribution of groups of units. Here, larger ablation effects indicate greater disruption of the evaluated model; improved accuracy-preserving pruning would require a separate evaluation.

Cross-Seed Consistency of Functional Niches To examine whether the functional organization described in Section 4.2.4 recurs across independently trained models, Figure 15 compares five consensus clusters (C0–C4) with their corresponding confirmation centroids from five training seeds (0–4) on Tiny-ImageNet. The analysis summarizes the distribution of positive class-conditional average causal effects (ACE), consistent with the specialization measure in Section 3.4. Panel A presents the class-level profiles, while Panel B reports complementary criteria measuring single-class concentration, semantic-domain concentration, and the breadth of class-level contributions. The three semantic specialist clusters exhibit consistent domain preferences across seeds. The consensus centroids of C0, C1, and C2 assign positive ACE mass of 0.77, 0.62, and 0.72 to animal, vehicle, and household classes, respectively. Across the five seeds, the corresponding domain masses range from 0.73 to 0.83, 0.57 to 0.69, and 0.64 to 0.82. These results indicate that broad semantic preferences recur across training runs despite variation in the contributions of individual classes. The two clusters labeled as generalists exhibit different degrees of class selectivity. C3 has the highest mean normalized class-profile entropy, with a consensus value of 0.81 and seed-specific values between 0.77 and 0.85. C4 has a lower consensus entropy of 0.60, with values between 0.58 and 0.62 across seeds; its generalist label therefore should not be interpreted as evidence of equally diffuse contributions. Moreover, the fine-grained specialization criterion remains low across all displayed rows (0.01–0.05), providing limited evidence for a recurring, strongly single-class-selective cluster in this analysis. Overall, these observations support the recurrence of broad semantic specialization at the cluster level. They do not establish a one-to-one correspondence between individual heads across runs or the reproducibility of every niche category described in the main text.

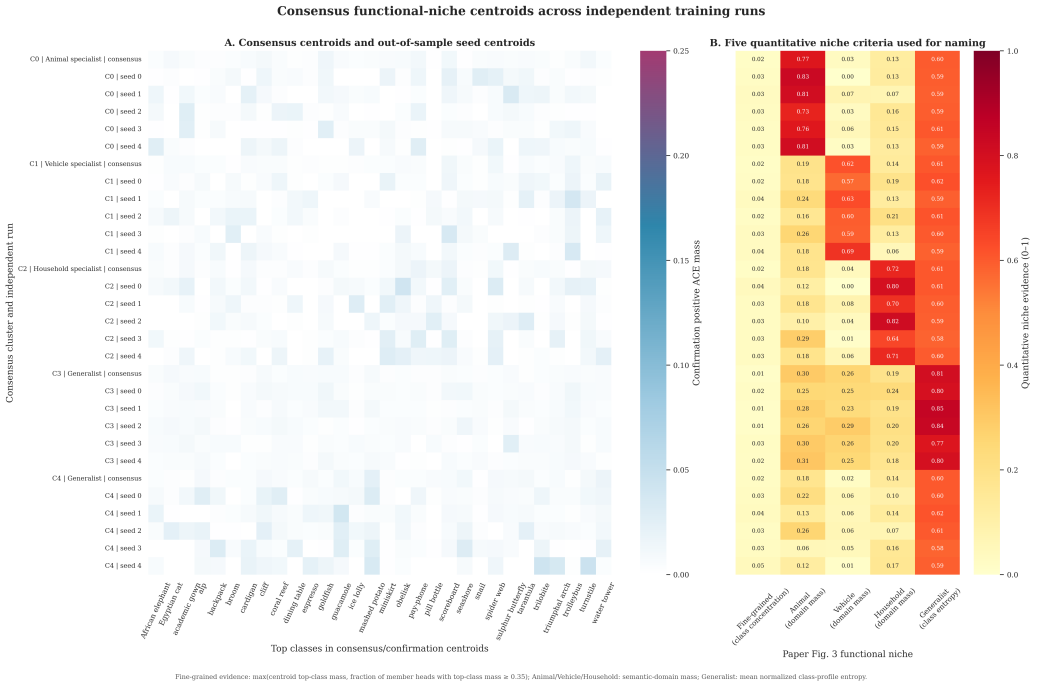


Figure 15: Consensus functional-niche centroids across independent training runs on Tiny-ImageNet. (A) Class-level profiles for five consensus clusters (C0–C4). Each six-row block contains the consensus centroid followed by the corresponding confirmation centroids from training seeds 0–4. Columns show selected classes, and color intensity denotes normalized positive class-conditional average causal effect (ACE) mass. (B) Quantitative criteria used to interpret the clusters, with rows ordered as in (A). Fine-grained evidence is the maximum of the centroid's largest class mass and the fraction of member heads whose largest class mass is at least 0.35. Animal, vehicle, and household evidence measure positive ACE mass within the respective semantic domains. Generalist evidence is the mean normalized entropy of member-head class profiles. These criteria measure different properties and do not form a probability distribution over niche types. C0–C2 retain their respective semantic-domain preferences across seeds, whereas the two clusters labeled as generalists differ in profile entropy.