ASYMMETRIC TRAINING WITH HETEROGENEOUS LOSSES: A PROBE INTO ARCHITECTURAL RESONANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Is deep learning robustness necessarily rooted in optimizing a single objective? We explore an alternative view: adaptive generalization may emerge from structured interactions among heterogeneous objectives during training. We propose an Asymmetric Training Paradigm that temporarily introduces non-competitive, per-class supervision (sigmoid losses) into networks optimized with competitive softmax objectives. This is realized through orthogonally initialized auxiliary pathways, modulated by a scalar coefficient α and present only during training. This controlled form of temporary topological redundancy creates an ideal probe for studying objective interactions. Our mechanistic analysis shows that such redundancy consistently smooths the initial loss landscape, but its performance impact follows a Principle of Architectural Resonance: auxiliary signals benefit models only when aligned with architectural inductive biases. A 6-block Vision Transformer (ViT-6L) exhibits constructive gradient alignment (cosine similarity +0.19), yielding up to 25% accuracy gains on CIFAR-100 with $20\times$ redundancy; by contrast, a CNN shows destructive conflicts (cosine similarity -0.26), leading to degradation. These findings challenge the view of auxiliary supervision as a universal regularizer. Instead, they reveal robustness as an outcome of structured internal dialogues between objectives, opening a path toward the design of multi-objective training systems tuned to architectural biases.

1 INTRODUCTION

A fundamental challenge in deep learning is understanding the complex interplay between a model's architectural inductive biases and the training strategies it is subjected to. While auxiliary supervision is a widely adopted technique for improving model performance(Szegedy et al., 2015; Lee et al., 2015; Caruana, 1997; Ruder, 2017), its application has been predominantly homogeneous, using objectives conceptually aligned with the main task. This raises a critical and largely unexplored question: what happens when auxiliary signals are fundamentally heterogeneous? Specifically, how does a system designed for "winner-takes-all" competition (via softmax) react to signals that encourage "feature coexistence" (via sigmoid)?

To investigate this, we propose the Asymmetric Training Paradigm, a framework designed as a precise scientific probe. It temporarily introduces non-competitive, sigmoid-based supervision into a network through orthogonally initialized pathways, allowing us to systematically study the resulting internal dynamics. Our investigation reveals a striking phenomenon that challenges conventional wisdom, which we term the Principle of Architectural Resonance. On CIFAR-100, this single paradigm produces radically divergent outcomes: Vision Transformers achieve a remarkable +25.4% performance improvement, driven by a sustained, constructive gradient synergy (cosine similarity +0.19), while Convolutional Neural Networks (CNNs) suffer a severe degradation (-22.0%), caused by a persistent, destructive gradient conflict (cosine similarity -0.26).

Our work makes three key contributions. First, we introduce the Asymmetric Training Paradigm as a novel and controllable platform for analyzing architecture-objective interactions. Second, using this probe, we discover and empirically validate the Principle of Architectural Resonance, providing multi-dimensional evidence that the efficacy of auxiliary supervision is highly architecture-

dependent. Third, to the best of our knowledge, we provide the first quantitative characterization of these gradient dynamics, establishing a direct link between the nature of the internal signal dialogue and the final generalization performance.

2 Related Work

2.1 Auxiliary Supervision and Multi-Task Learning

The use of intermediate supervision, or "deep supervision," is a well-established technique, originally pioneered in networks like GoogLeNet to combat vanishing gradients in deep architectures (Szegedy et al., 2015; Lee et al., 2015). Modern applications leverage auxiliary tasks for representation learning, notably in self-supervised learning (Gidaris et al., 2018; Chen et al., 2020) and Multi-Task Learning (MTL) (Caruana, 1997; Ruder, 2017; Kendall et al., 2018). However, a common thread in these approaches is the use of homogeneous or synergistic tasks. Our work diverges by using a deliberately heterogeneous signal (non-competitive vs. competitive) not merely for performance, but as a scientific probe to understand a system's response to conflicting objectives. This contrasts with recent trends that leverage auxiliary tasks primarily for representational consistency and robustness enhancement within homogeneous objective families (Navon et al., 2022; Shamsian et al., 2023). While these approaches demonstrate effectiveness in traditional multi-task scenarios, they do not explore the fundamental architectural response to qualitatively different supervisory signals, which constitutes the core contribution of our study.

2.2 Architectural Inductive Biases

Our analysis is grounded in the distinct inductive biases of different architectures(Goyal & Bengio, 2022). CNNs enforce strong priors on spatial locality and translation equivariance through weight-sharing kernels (LeCun et al., 1989; Cohen & Welling, 2016). In contrast, ViTs have weaker spatial priors, relying on self-attention to dynamically learn global relationships from data (Dosovitskiy et al., 2021; Vaswani et al., 2017). MLPs, with the weakest biases, serve as a reference (Tolstikhin et al., 2021). While these individual biases are well-studied, how they govern a model's response to heterogeneous supervisory signals remains largely unexplored. This knowledge gap persists despite recent advances in understanding the subtle differences between CNN and ViT architectures in terms of optimization landscapes, feature geometries, and inductive bias mechanisms (Lu et al., 2022; Tuli et al., 2021). Our work directly addresses this unexplored interaction by treating the supervisory signal as a controlled variable and the architecture as the primary subject of investigation.

2.3 The Interplay of Architectures and Training Objectives

Our research intersects with fields that study the co-design of architectures and training, such as Neural Architecture Search (NAS) (Zoph & Le, 2016; Liu et al., 2019; Mellor et al., 2021) and studies on architecture-aware regularization (Srivastava et al., 2014; Ioffe & Szegedy, 2015). However, our methodology differs fundamentally. Rather than searching for an optimal architecture-objective pair to maximize performance, we employ an "experimental physics" approach: we fix the architectures and systematically vary the properties of the external signal (e.g., strength α and redundancy N_{branches}) to map their interaction landscape. This paradigm enables us to uncover a general principle—Architectural Resonance—rather than a task-specific optimal configuration. To our knowledge, this is the first study of heterogeneous supervision and architectural bias interaction with the explicit goal of discovering a fundamental principle that governs this interplay.

3 Methodology

3.1 The Asymmetric Training Paradigm

This paper introduces the Asymmetric Training Paradigm, a novel framework designed to enrich a model's learning signals by introducing temporary, learnable auxiliary structures exclusively during the training phase. This approach aims to improve generalization and reveal underlying learning mechanisms without incurring any additional inference cost. The paradigm is founded on three core

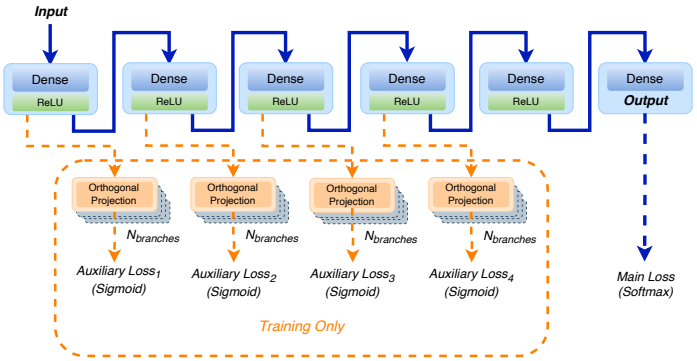


Figure 1: The Asymmetric Training Paradigm.

design principles: **Asymmetry** (different topologies for training vs. inference), **Heterogeneity** (dissimilar primary and auxiliary learning objectives), and controllable **Redundancy** (scalable auxiliary pathways). As illustrated in Figure 1, our framework serves not only as a performance enhancement technique but, more importantly, as a principled instrument to uncover a key mechanism for robust generalization: structured multi-objective dialogue.

3.2 Core Hypothesis: The Principle of Architectural Resonance

We propose the Principle of Architectural Resonance, which posits that a network's generalization ability originates not just from optimizing a single objective, but from the structured interaction between heterogeneous learning objectives and the architecture's intrinsic inductive biases. Specifically, we hypothesize that: (1) When an auxiliary signal is synergistic with an architecture's inductive bias (e.g., a ViT's global relational capacity), it produces constructive gradient alignment and improves generalization. (2) When the two conflict (e.g., a CNN's strong spatial locality), it leads to destructive gradient conflicts and impedes learning. Our paradigm provides an ideal testbed for empirically testing this principle, as we will demonstrate in Section 4.

3.3 Core Mechanisms and Design Principles

3.3.1 Architectural Design

To test our hypothesis in a controlled environment, our study is conducted on the CIFAR-10/100 datasets. We employ three lightweight backbone architectures representing a spectrum of inductive biases: a 6-layer MLP, a 6-conv-layer CNN with spatial downsampling, and a 6-block Vision Transformer (ViT-6L). All auxiliary branches consist of a single linear layer and are attached at intermediate points of the backbone. To ensure training stability, all auxiliary branch weights are initialized using Orthogonal Initialization (Saxe et al., 2014). All models are trained using the AdamW optimizer (Kingma & Ba, 2014; Loshchilov & Hutter, 2019). Specific architectural configurations are detailed in Appendix C.3.

3.3.2 Key Variables: Redundancy and Dialogue Strength

Our paradigm features two core controllable variables:

- **Architectural Redundancy (N_{branches})**: The number of parallel auxiliary branches at each attachment point. For example, a redundancy of $20\times$ ($N_{\text{branches}} = 20$) means that at each point where an auxiliary signal is injected, we create 20 separate, parallel linear layers. Each of these 20 branches takes the exact same intermediate representation as input, and each is intended to produce an auxiliary loss, though only one is active during backpropagation due to our "single-activation" strategy. In our primary performance experiments, we evaluate redundancy levels of $N_{\text{branches}} \in \{1, 7, 10, 20\}$. For specific mechanistic analysis, such as the initial loss landscape A.1, our exploration includes levels of $\{1, 20, 300\}$. A

supplementary ablation on an alternative hyperparameter search strategy for $N_{\text{branches}} = 3$ is provided in Appendix Table 16.

- **Dialogue Strength** (α): A scalar hyperparameter that balances the primary and auxiliary losses. The total training objective is:

$$L_{\text{total}} = L_{\text{main}} + \alpha \sum_{k=1}^{K} L_{\text{aux}}^{(k)} \tag{1}$$

3.3.3 Auxiliary Branch Activation: A Deterministic Probe for Landscape Effects

A critical design choice in our paradigm is how the multiple auxiliary branches are utilized. Our central hypothesis is that the initial geometry of the loss landscape, modulated by topological redundancy, is a key determinant of the final generalization performance. To create the cleanest possible testbed for this hypothesis, we employ a deterministic fixed-path activation strategy.

Implementation Details. At each of the K attachment points, we initialize N_{branches} parallel auxiliary branches. However, during the entire training process, for every forward and backward pass, we consistently and exclusively select only the last branch (the N_{branches}-th one) to be active. This means the first $N_{\text{branches}} - 1$ branches are never activated or trained, and their weights remain fixed at their initial values.

Design Rationale. The rationale for this deterministic choice is to isolate a single key variable: the impact of the initial loss landscape's geometry. A core premise of our approach is that increasing the number of static branches reliably smooths the initial loss landscape, a phenomenon we systematically validate in Appendix A.1. For instance, our analysis shows that for the MLP architecture, increasing redundancy from $1\times$ to $300\times$ reduces the standard deviation of the loss surface by over 90% Table 5 in Appendix. Our fixed-path design uses the single active branch as a constant, unchanging probe to measure how different architectures navigate these systematically altered landscapes.

An alternative, such as stochastically activating different branches, would have introduced a powerful confounding variable—randomized regularization—making it impossible to disentangle the effects of the landscape's geometry from the effects of the stochastic training signal. Our deterministic approach intentionally removes this randomness. It allows us to ask a precise question: Does the initial smoothness provided by static redundancy directly translate into a dynamic optimization advantage?

Interpreting the Effects. The observed performance divergence must be interpreted through this lens. The starkly different outcomes for CNNs and ViTs are a direct answer to our research question. The results suggest that Architectural Resonance is a fundamental principle governing how an architecture's inductive bias determines its ability to exploit the properties of a given optimization landscape. For ViTs, initial smoothness is a benefit they can leverage; for CNNs, under the same conditions, it becomes a detriment. This finding highlights a deeper, more geometric level of interaction between model architecture and the optimization process.

3.3.4 Heterogeneous Learning Objectives

The core of our paradigm is an "internal dialogue" between two qualitatively different objectives:

- **Primary Objective (L_{main})**: The standard, class-competitive softmax cross-entropy loss.

$$L_{\text{main}} = -\sum_{i=1}^{C} y_i \log(\text{softmax}(\mathbf{z}_{\text{final}})_i) \tag{2}$$

- **Auxiliary Objective (L_{aux})**: An independent, non-competitive sigmoid-based binary cross-entropy loss applied per class.

$$L_{\text{aux}} = -\frac{1}{C} \sum_{i=1}^{C} [y_i \log(\sigma(z_i)) + (1 - y_i) \log(1 - \sigma(z_i))] \tag{3}$$

3.4 Methodological Rigor: The Principle of Fair Comparison

Our experiments are predicated on the principle of fair comparison. For each baseline (Plain) model, we first perform a comprehensive grid search to identify the best-performing learning rate and weight decay on the validation set. We then freeze these hyperparameters and conduct a wide-range search for the best α for our Asymmetric variant. We term this the "Pragmatic Gold Standard" strategy, as it isolates the specific effect of our module and mirrors a realistic plug-and-play application scenario.

4 Experiments and Analysis

This section provides a rigorous empirical validation of the Principle of Architectural Resonance. Our analysis follows a clear logical chain: we first establish a universal geometric effect (Section 4.2), then present the core puzzle of performance divergence (Section 4.3), unpack the puzzle with multi-faceted mechanistic analysis (Section 4.4), conduct a quantitative dose-response analysis (Section 5), and finally, explore the principle's applicability boundaries (Section 5.1).

4.1 Experimental Setup

Datasets and Architectures. Our experiments are conducted on CIFAR-10 and CIFAR-100. We employ the three architectures (MLP, CNN, and ViT) as specified in Section 3.3.1. The 50,000 training images are split into an 80%/20% ratio for training and validation subsets.

Implementation Details. For all comparisons, we follow the "Pragmatic Gold Standard" hyperparameter search strategy detailed in Section 3.4. Further details on the specific hyperparameter search spaces are provided in Appendix C.1. To ensure reliability, all reported results are the mean $\pm$ standard deviation over 10 independent runs. Statistical significance is evaluated using a two-tailed paired t-test ($p < 0.05$). Experiments were conducted on a server with 8 NVIDIA A100 GPUs. Our code will be made public upon acceptance.

4.2 Foundational Phenomenon: Universal Smoothing Effect of Architectural Redundancy

Before investigating final performance, we analyzed the impact of our paradigm on the initial geometry of the loss landscape. As detailed in Table 5 and 6, we observe a universal trend across all architectures: as the degree of temporary topological redundancy increases, the initial loss landscape becomes demonstrably smoother 6. This indicates that our paradigm provides a better-structured starting point for the optimizer, consistent with literature suggesting that smoother landscapes facilitate better generalization (Li et al., 2018; Keskar et al., 2017; Garipov et al., 2018).

Table 1: Architecture performance comparison across different configurations on CIFAR-100

$N_{branches}$	Arch	Plain	Asymmetric	Improvement	p-value
1×	MLP	0.232 ± 0.003	0.243 ± 0.003 ($\alpha = 100.0$)	+4.9%	0.0001
	CNN	0.395 ± 0.004	0.308 ± 0.003 ($\alpha = 100.0$)	-22.0%	0.0000
	ViT-6L	0.358 ± 0.010	0.359 ± 0.005 ($\alpha = 4.642$)	+0.3%	0.7773
7×	MLP	0.234±0.005	0.221±0.003 ($\alpha = 0.1$)	-5.5%	0.0001
	CNN	0.398 ± 0.006	0.317 ± 0.029 ($\alpha = 0.1$)	-20.2%	0.0000
	ViT-6L	0.359 ± 0.008	0.400 ± 0.021 ($\alpha = 4.642$)	+11.3%	0.0000
10×	MLP	0.233 ± 0.003	0.217 ± 0.002 ($\alpha = 0.1$)	-6.9%	0.0000
	CNN	0.397 ± 0.006	0.2431 ± 0.021 ($\alpha = 1.0$)	-38.7%	0.0000
	ViT-6L	0.360 ± 0.009	0.418 ± 0.022 ($\alpha = 4.642$)	+15.9%	0.0000
20×	MLP	0.232 ± 0.003	0.211 ± 0.003 ($\alpha = 0.1$)	-9.0%	0.0000
	CNN	0.397 ± 0.006	0.336 ± 0.012 ($\alpha = 0.1$)	-15.3%	0.0000
	ViT-6L	0.361 ± 0.010	0.453 ± 0.013 ($\alpha = 1.0$)	+25.4%	0.0000

4.3 Core Finding: Performance Divergence on a Smoothed Landscape

However, on this consistently simplified optimization environment, we observe a surprising, architecture-dependent performance divergence. As presented in Table 1, on the challenging CIFAR-100 dataset, the ViT's performance scales positively with redundancy, achieving a gain of up to **+25.4%** ($p < 0.001$). In stark contrast, the CNN's performance catastrophically degrades by **-22.0%** ($p < 0.001$) with just a single auxiliary branch. The MLP exhibits a complex, non-monotonic pattern. This stark divergence strongly suggests that an architecture's intrinsic properties, not landscape smoothness alone, fundamentally determine the utility of the auxiliary signal.

4.4 Mechanistic Analysis: Multi-layered Evidence for Architectural Resonance

To unravel this puzzle, we conduct a multi-faceted mechanistic analysis.

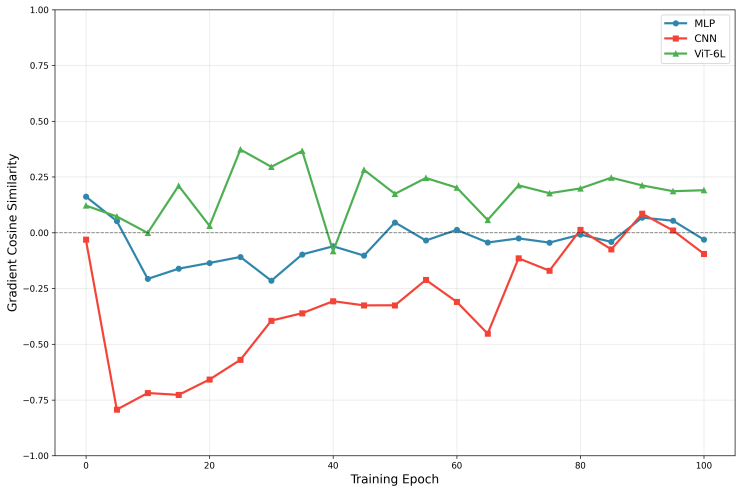


Figure 2: Gradient conflict evolution across architectures during training

Table 2: Detailed Gradient Conflict Analysis

Architecture	Final Sim	Avg Sim	Min Sim	Max Sim	Gradient Interaction
MLP	-0.0309	-0.0801	-0.3422	0.0701	Neutral/weak interaction
CNN	-0.1926	-0.2574	-0.8210	0.0464	Strong conflict detected
ViT	0.2631	0.1870	-0.1845	0.3654	Constructive synergy

4.4.1 Mathematical Level: Gradient Dynamics

To diagnose the mathematical origins of this performance divergence, we directly measured the cosine similarity between the primary (g_{main}) and auxiliary (g_{aux}) gradients throughout training. The evolution of this gradient alignment, visualized in Figure 2, reveals immediate and persistent architectural signatures. The ViT's alignment (green curve) consistently remains in positive territory, indicating a constructive dialogue. Conversely, the CNN's alignment (red curve) immediately plunges into and remains in negative territory, signifying a sustained conflict. The MLP (blue curve) hovers around zero, suggesting a largely uncorrelated or directionless interaction.

To precisely quantify these visual patterns, we provide a detailed statistical summary in Table 2. The data confirms the visual narrative with high fidelity. The ViT maintains a healthy average positive similarity of +0.1870, confirming the presence of a "Constructive synergy." In stark contrast, the CNN exhibits a strong average negative similarity of -0.2574. More tellingly, the conflict in the CNN can be extreme, reaching a minimum similarity of -0.8210, which indicates moments of near-perfect gradient opposition. This finding strongly supports the table's qualitative conclusion of a "Strong conflict detected."

This direct mathematical evidence provides a clear mechanistic explanation for the divergent learning outcomes. The success of the ViT and the failure of the CNN are not accidental; they are a direct reflection of the persistent synergistic or conflicting nature of their internal gradient dialogues.

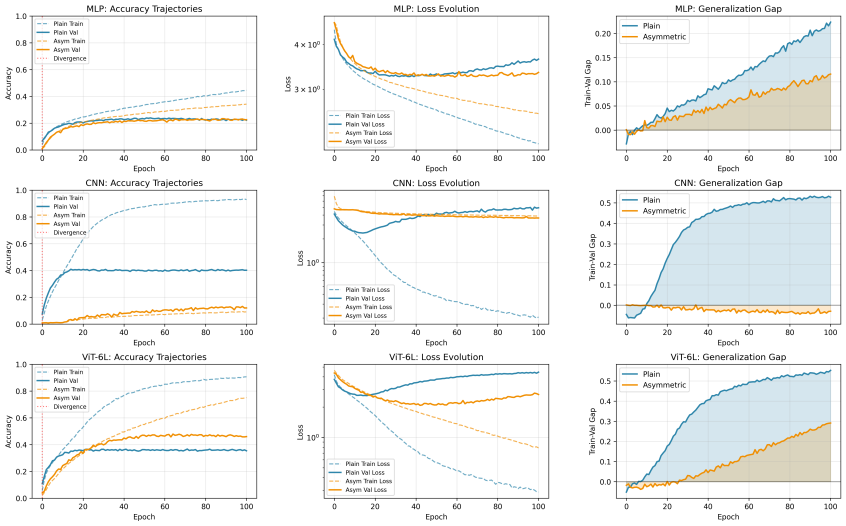


Figure 3: Architecture-specific learning dynamics reveal the Principle of Architectural Resonance. Comprehensive learning trajectories comparing Plain (baseline) and Asymmetric training across three architectures on CIFAR-100. Left column shows accuracy evolution (dashed: training, solid: validation); middle column displays loss curves; right column presents generalization gaps (train-val accuracy difference). CNN exhibits catastrophic degradation with massive overfitting under asymmetric training. MLP demonstrates effective regularization with reduced generalization gap but limited accuracy gains. ViT achieves substantial performance improvements with superior generalization. The divergence points (red vertical lines) mark early onset of architecture-dependent responses to heterogeneous supervision, empirically validating our core hypothesis that auxiliary signal efficacy depends fundamentally on architectural inductive biases.

Table 3: Architecture-dependent convergence patterns and performance outcomes.

	Convergence (Epoch)		**Final Val Acc**	
Architecture	Plain	Asymmetric	Plain	Asymmetric
MLP	15	21	0.2220	0.2305
CNN	20	5	0.4034	0.1171
ViT-6L	19	30	0.3498	0.4568

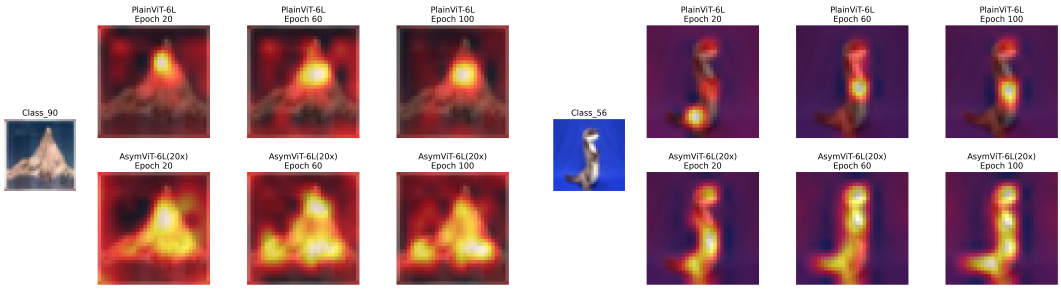
Table 4: Generalization Gap Evolution Across Training Phases

	Early(epoch10)		**Middle**(epoch50)		**Late**(epoch100)	
Architecture	Plain	Asymmetric	Plain	Asymmetric	Plain	Asymmetric
MLP	0.0098	0.0049	0.1039	0.0566	0.2234	0.1225
CNN	0.0073	0.0002	0.4665	-0.0242	0.5253	-0.0365
ViT-6L	0.0334	-0.0270	0.4717	0.0838	0.5543	0.2934

4.4.2 Process Level: Learning Trajectories

This microscopic gradient behavior directly translates into dramatically different macroscopic learning dynamics, as evidenced by our comprehensive analysis across Figure 3 and Tables 3-4. For CNN, the persistent gradient conflict drives a catastrophic optimization collapse—the model converges

prematurely in just 5 epochs to a drastically inferior solution with 71% performance degradation. More tellingly, the generalization gap becomes negative by epoch 50 (-0.0242), indicating the model performs better on validation than training data—a clear symptom of learning failure. In contrast, ViT's constructive gradient synergy guides the optimization along a more exploratory but ultimately superior trajectory, requiring 11 additional epochs but achieving both 30.6% higher validation accuracy and 47% better generalization (gap reduction from 0.5543 to 0.2934). This demonstrates that beneficial gradient alignment not only improves final performance but fundamentally enhances the learning process itself.



(a) (b)

Figure 4: Comparison of attention evolution

4.4.3 Representation Level: Attention Patterns

For the ViT, the synergistic effect manifests in the learned representations. As shown in Figure 4, the Asymmetric model learns a more holistic attention pattern. Quantitative analysis of attention patterns on 100 randomly sampled CIFAR-100 test images reveals that the ViT-6L(20×) achieves significantly higher average object coverage (93.7%) compared to the Plain baseline (87.9%), visually confirming a more comprehensive understanding of the input (Figure 4).

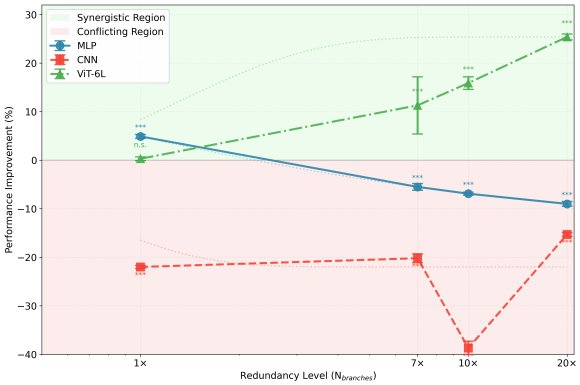


Figure 5: Architecture-dependent dose-response to auxiliary supervision. Each architecture shows distinct sensitivity patterns: ViT (synergistic), CNN (destructive), MLP (orthogonal). Performance improvement plotted against branch redundancy (N_{branches}) on CIFAR-100. Error bars represent standard error (n=10).

5 Dose-Response Analysis: Differential Modulation by Redundancy

To systematically quantify the interaction between our paradigm and architectural biases, we performed a dose-response analysis, treating the level of architectural redundancy (N_{branches}) as the "dose" and the resulting performance gain as the "response." The results provide compelling quan-

titative evidence for the Principle of Architectural Resonance, with detailed performance metrics available in Table 1 and the overall trends visualized in Figure 5, revealing three distinct patterns.

First, the ViT-6L exhibits a clear synergistic resonance. Its performance scales positively and monotonically with increasing redundancy, culminating in a remarkable +25.4% improvement at $20\times$ redundancy (Table 1). This synergistic trend is clearly depicted by the green curve in Figure 5.

In stark contrast, the CNN demonstrates a destructive conflict. Even a single auxiliary branch ($1\times$) causes a severe performance degradation of -22.0% (Table 1). As shown by the red curve in Figure 5, the overall effect remains strongly negative across all redundancy levels, indicating a fundamental incompatibility between the CNN's strong spatial priors and the nature of the auxiliary objective.

Finally, the MLP, possessing the weakest inductive bias, shows a complex, non-monotonic profile, peaking at a +4.9% gain before declining to -9.0% (Table 1). It gains a small benefit at low redundancy (+4.9% at $1\times$), but this quickly turns into a performance loss as redundancy increases, declining to -9.0% at $20\times$. This declining trend, also visible in Figure 5, suggests that while a small amount of signal diversity can initially help, the MLP's architecture lacks the structural capacity of a ViT to productively organize a large volume of heterogeneous signals.

Taken together, these divergent curves form the empirical bedrock of the Architectural Resonance principle, directly linking an architecture's intrinsic properties to its ability to benefit from multi-objective dialogues.

5.1 Boundary Conditions: Contextual Dependencies

Finally, we conducted preliminary investigations into the principle's applicability under different conditions, including few-shot and noisy-label scenarios. Initial results (see Appendix B) suggest the effect may be context-dependent, with dataset properties potentially modulating the resonance effect. However, comprehensive characterization of these boundary conditions remains an important direction for future work.

6 Discussion and Conclusion

This work identifies an underlying principle we term Architectural Resonance: the efficacy of an auxiliary training signal appears to be heavily influenced by its compatibility with a model's intrinsic inductive biases. We provided strong empirical evidence for this principle through our Asymmetric Training Paradigm. This framework revealed a stark performance divergence under identical conditions: a +25.4% gain for Vision Transformers, versus a -22.0% degradation for CNNs, a discrepancy we traced to their opposing gradient dynamics.

These findings invite a deeper interpretation. The ViT, with its flexible self-attention mechanism, appears to leverage the auxiliary signal to learn a more comprehensive feature set. In contrast, the CNN, with its strong spatial priors, experiences a destructive conflict. The MLP's declining response is characteristic of its weak inductive bias—while initially benefiting from minimal signal diversity (+4.9%), it becomes increasingly overwhelmed as redundancy grows, lacking the structural capacity to productively organize large volumes of heterogeneous signals. Together, these results suggest a more nuanced perspective than viewing auxiliary supervision as a universal regularizer, pointing towards the value of co-designing architecture-aware training strategies.

Our study, however, has its boundaries. The empirical validation was conducted in a controlled setting on CIFAR datasets with lightweight models, a deliberate choice to enable the rigorous mechanistic analysis necessary to first identify the principle. A critical open question is the scalability of our findings to larger benchmarks like ImageNet and standard, pre-trained architectures. Exploring the manifestations of Architectural Resonance in other domains, such as NLP and graph learning, presents another valuable direction for future work.

In conclusion, our work points towards a promising avenue for improving model training. It suggests that rather than pursuing a single, universally optimal objective, a valuable research direction is the principled engineering of structured, internal dialogues that are carefully tuned to the resonant properties of each unique architecture.

REFERENCES

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, pp. 2990–2999. PMLR, 2016.

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

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in Neural Information Processing Systems*, 31, 2018.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.

Anirudh Goyal and Yoshua Bengio. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478(2266):20210068, 2022.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pp. 448–456. PMLR, 2015.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491, 2018.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. doi: 10.1162/neco.1989.1.4.541.

Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pp. 562–570. PMLR, 2015.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in Neural Information Processing Systems*, 31, 2018.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. In *International Conference on Learning Representations*, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Zhiying Lu, Hongtao Xie, Chuanbin Liu, and Yongdong Zhang. Bridging the gap between vision transformers and convolutional neural networks on small datasets, 2022.

Joseph Mellor, Jack Turner, Amos Storkey, and Elliot J. Crowley. Neural architecture search without training, 2021.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16428–16446. PMLR, 17–23 Jul 2022.

Sebastian Ruder. An overview of multi-task learning in deep neural networks. *ArXiv*, abs/1706.05098, 2017.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

Aviv Shamsian, Aviv Navon, Neta Glazer, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Auxiliary learning as an asymmetric bargaining game, 2023.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34:24261–24272, 2021.

Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L. Griffiths. Are convolutional neural networks or transformers more like human vision?, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning. *CoRR*, abs/1611.01578, 2016.

A Appendix

A.1 Loss Landscape Analysis

We conducted systematic loss landscape analysis across MLP, CNN, and ViT-6L architectures on CIFAR-10 and CIFAR-100. Following established protocols (Li et al., 2018), we visualized the loss surfaces using a 51×51 grid centered at the initialization point, with directions determined by random Gaussian perturbations normalized to unit variance. To balance computational efficiency with statistical robustness, we randomly sampled 500 training instances for loss evaluation at each grid point. This sampling size provides sufficient statistical power while remaining computationally tractable for systematic analysis across multiple architectures and redundancy levels. The resulting visualizations reveal distinct architectural signatures in terms of loss surface smoothness and optimization landscape complexity. (Figure 6, Table 5 and 6)

B Boundary Conditions and Extended Analysis

B.1 Few-Shot Learning Robustness

To assess the generalizability of our architectural resonance findings under data-scarce conditions, we conducted systematic few-shot learning experiments on CIFAR-10 and CIFAR-100. We hypothesize that asymmetric training benefits should be amplified in low-data regimes, where auxiliary

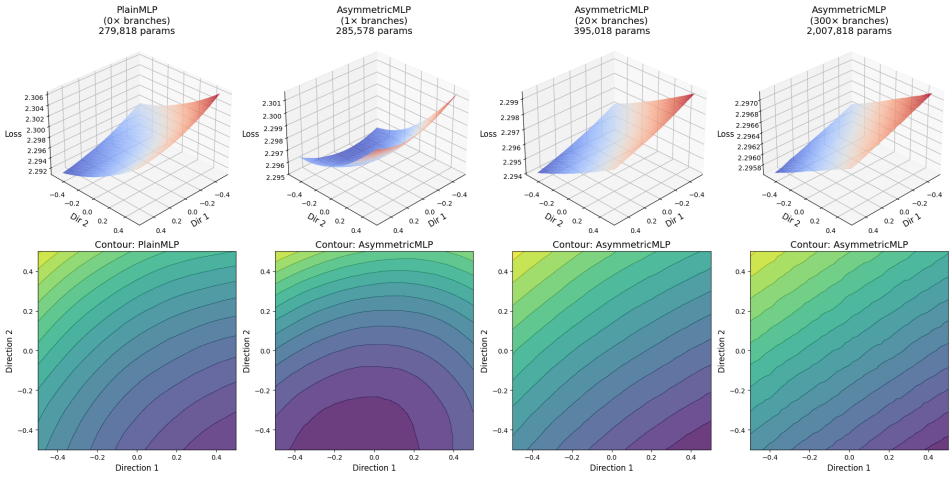


Figure 6: Loss landscape visualization for MLP on CIFAR-10

Table 5: Progressive loss landscape smoothing in MLP architecture on CIFAR-10. Standard deviation (Std), range, and mean gradient magnitude all decrease systematically with increased redundancy, demonstrating that topological modifications consistently flatten the optimization surface independent of final performance outcomes. Percentages indicate relative change from baseline.

Model	Params	Std(Loss)	Range(Loss)	Mean(Grad)
Plain	0.28M	0.0032	0.0152	0.0002
Asymmetric(1×)	0.29M	0.0015 (-52.3%)	0.0064 (-57.6%)	0.0001 (-49.8%)
Asymmetric(20×)	0.40M	0.0011 (-64.4%)	0.0053 (-64.9%)	0.0001 (-65.0%)
Asymmetric(300×)	2.01M	0.0003 (-90.2%)	0.0015 (-90.3%)	0.0000 (-90.4%)

supervision can provide crucial structural guidance when primary signals are sparse (Tables 7, 8, and 9).

Experimental Design. We systematically varied the number of training samples per class from 5 to 5000, creating a comprehensive data scarcity spectrum. For each data regime, we maintained the original test set size to ensure consistent evaluation conditions. All experiments were repeated across 10 random seeds with stratified sampling to ensure class balance. Statistical significance was assessed using two-tailed paired t-tests.

Theoretical Motivation. Under data scarcity, the auxiliary sigmoid branches should provide particularly valuable regularization, as the primary softmax objective becomes increasingly prone to overfitting. This effect should be most pronounced in architectures that exhibit gradient synergy rather than conflict.

B.2 Noise Robustness Analysis

We evaluated model robustness under label noise by corrupting a fraction of training labels and measuring performance degradation. Label noise was introduced by randomly flipping labels with probabilities ranging from 10% to 90%, while maintaining the original test set for consistent evaluation (Tables 11 and 12).

Table 6: Cross-architecture comparison of loss landscape smoothing on CIFAR-100. Despite universal landscape flattening effects (up to 90% reduction in surface roughness), architectural differences emerge: CNN shows the most dramatic smoothing with minimal parameter increase, while ViT-6L exhibits more modest but consistent improvements. These results demonstrate that landscape conditioning is architecture-agnostic, yet performance benefits depend critically on architectural resonance with auxiliary signals.

Model	Params	Std(Loss)	Range(Loss)	Mean(Grad)
CNN				
Plain	2.43M	0.0008	0.0040	0.0001
Asymmetric(1×)	2.51M	0.0007 (-16.3%)	0.0033 (-17.3%)	0.0000 (-15.9%)
Asymmetric(20×)	4.09M	0.0003 (-66.7%)	0.0013 (-68.0%)	0.0000 (-66.9%)
Asymmetric(300×)	27.4M	0.0001 (-90.2%)	0.0004 (-90.3%)	0.0000 (-90.2%)
ViT-6L				
Plain	1.22M	0.0045	0.0212	0.0003
Asymmetric(1×)	1.29M	0.0043 (-3.7%)	0.0204 (-3.7%)	0.0003 (-3.6%)
Asymmetric(20×)	2.75M	0.0028 (-37.7%)	0.0132 (-38.0%)	0.0002 (-37.1%)
Asymmetric(300×)	24.3M	0.0009 (-80.1%)	0.0044 (-79.3%)	0.0001 (-77.6%)

Table 7: Few-shot learning performance of MLP on CIFAR-10.

Samples/Class	PlainMLP	AsymmetricMLP	Improvement	p-value
10	0.2290 ± 0.0167	0.2299 ± 0.0171	+0.38%	0.8082
50	0.2870 ± 0.0100	0.2897 ± 0.0082	+0.94%	0.4346
100	0.3199 ± 0.0118	0.3330 ± 0.0093	+4.09%	0.0122
500	0.3893 ± 0.0117	0.3944 ± 0.0105	+1.31%	0.3087
1000	0.4313 ± 0.0039	0.4400 ± 0.0043	+2.02%	0.0043
5000	0.5357 ± 0.0039	0.5374 ± 0.0056	+0.32%	0.4824

C Detailed Experimental Configuration

C.1 Hyperparameter Settings

All hyperparameters were determined through systematic grid search following our "Pragmatic Gold Standard" strategy to ensure fair comparison. This three-stage optimization process isolates the effect of our asymmetric training paradigm while maintaining scientific rigor.

C.1.1 Optimization Strategy

For each architecture, we employed a principled three-stage hyperparameter search:

Stage 1: Learning Rate Optimization We fixed weight decay at 10^{-4} and conducted grid search over learning rates $\{10^{-4}, 3 \times 10^{-4}, 10^{-3}, 3 \times 10^{-3}, 10^{-2}\}$ for the baseline Plain model, training for 150 epochs and selecting the configuration yielding highest validation accuracy.

Stage 2: Weight Decay Refinement Using the optimal learning rate from Stage 1, we searched over weight decay values $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ for the Plain model, again training for 150 epochs.

Stage 3: Auxiliary Weight Search With optimal learning rate and weight decay fixed, we searched for the optimal auxiliary weight α using logarithmic spacing: $\alpha \in \{0.1, 0.215, 0.464, 1.0, 2.154, 4.642, 10.0, 21.544, 46.416, 100.0\}$ for the Asymmetric model. For CIFAR-10 MLP specifically, we employed linear spacing $\alpha \in [0, 50]$ to accommodate its distinct optimization characteristics.

Table 8: Few-shot learning performance of MLP on CIFAR-100.

Samples/Class	PlainMLP	AsymmetricMLP	Improvement	p-value
5	0.0406 ± 0.0064	0.0522 ± 0.0025	+28.50%	0.0002
10	0.0727 ± 0.0038	0.0697 ± 0.0039	-4.17%	0.0619
20	0.0936 ± 0.0036	0.0949 ± 0.0029	+1.35%	0.2965
50	0.1298 ± 0.0045	0.1398 ± 0.0048	+7.72%	0.0041
100	0.1595 ± 0.0028	0.1697 ± 0.0029	+6.37%	0.0000
200	0.2131 ± 0.0044	0.2225 ± 0.0022	+4.43%	0.0004
500	0.2594 ± 0.0031	0.2665 ± 0.0032	+2.77%	0.0007

Table 9: Few-shot learning performance of ViT-6L(20×) on CIFAR-100.

Samples/Class	PlainViT-6L	AsymmetricViT-6L	Improvement (%)	p-value
5	0.0510 ± 0.0050	0.0559 ± 0.0033	+9.61	0.0022
10	0.0792 ± 0.0052	0.0879 ± 0.0049	+11.06	0.0008
20	0.0967 ± 0.0049	0.1066 ± 0.0068	+10.30	0.0052
50	0.1561 ± 0.0079	0.1558 ± 0.0051	-0.22	0.8687
100	0.2029 ± 0.0058	0.2316 ± 0.0078	+14.13	0.0001
200	0.2800 ± 0.0062	0.3223 ± 0.0179	+15.10	0.0011
500	0.4353 ± 0.0145	0.5328 ± 0.0065	+22.41	0.0000

C.1.2 Final Hyperparameter Configurations

The optimal hyperparameters determined through our systematic search are:

C.2 Training Configuration

Training Duration: All final results were obtained using 200 epochs.

Statistical Validation: Each configuration was evaluated across 10 independent runs with different random seeds (42-51) to ensure statistical robustness. Performance comparisons used two-tailed paired t-tests.

Hardware: All experiments were conducted on NVIDIA RTX 3090 GPUs with consistent computational environments to ensure reproducibility.

C.3 Architecture-Specific Details

MLP: 6 linear layers with ReLU activations. Auxiliary branches attached after the first 4 ReLU activations.

CNN: 6 convolutional layers, 2 MaxPooling layers, 1 Dropout layer, and 3 linear layers. Auxiliary branches are strategically placed after ReLU activations in convolutional blocks. When a convolutional layer is immediately followed by max-pooling, the auxiliary branch is placed after the max-pooling operation to maintain spatial coherence.

ViT-6L: 6 Transformer blocks with 4 attention heads each and embedding dimension of 128. Auxiliary branches attached after each Transformer block output.

All auxiliary branches consist of a single linear layer with output dimension equal to the number of classes, initialized using orthogonal initialization for training stability.

D Attention Pattern Evolution Analysis

D.1 Detailed Attention Visualization

(7)

Table 10: Architecture Performance Comparison (CIFAR-10)

Architecture	Baseline	Asymmetric(1×)	Improvement	P-Value
MLP	0.495 ± 0.003	0.506 ± 0.004	+2.3%	0.0001
CNN	0.770 ± 0.006	0.772 ± 0.010	+0.3%	0.5371
ViT-6L	0.625 ± 0.007	0.636 ± 0.004	+1.7%	0.0068

Table 11: MLP noise robustness on CIFAR-10

Noise Level	PlainMLP	AsymmetricMLP	Improvement	P-Value
0.0%	0.5352 ± 0.0053	0.5374 ± 0.0036	+0.42%	0.2025
10.0%	0.5182 ± 0.0047	0.5208 ± 0.0042	+0.50%	0.1785
20.0%	0.4992 ± 0.0055	0.5022 ± 0.0070	+0.59%	0.3667
30.0%	0.4769 ± 0.0036	0.4842 ± 0.0045	+1.54%	0.0007
40.0%	0.4496 ± 0.0062	0.4590 ± 0.0085	+2.09%	0.0067
50.0%	0.4213 ± 0.0109	0.4266 ± 0.0052	+1.24%	0.2633
60.0%	0.3835 ± 0.0057	0.3923 ± 0.0070	+2.29%	0.0530
70.0%	0.3294 ± 0.0085	0.3205 ± 0.0093	-2.70%	0.0291
80.0%	0.2369 ± 0.0161	0.2294 ± 0.0094	-3.17%	0.1580
90.0%	0.1043 ± 0.0053	0.1013 ± 0.0061	-2.87%	0.1081

D.2 Quantitative Attention Analysis

We measured attention pattern quality using several metrics:

Key findings:

- **Peak Strength**: Asymmetric training produces more diffuse attention patterns (lower peak values)
- **Map Entropy**: Higher entropy indicates more distributed attention across spatial locations
- **Sparsity**: Lower Gini coefficient suggests more egalitarian attention distribution
- **Object Coverage**: Asymmetric models achieve near-optimal object coverage much earlier (Epoch 20 vs 100)

D.3 Layer-wise Attention Development

The layer-wise analysis reveals that asymmetric training guides the development of hierarchical attention patterns: - **Early layers (L1)**: Both variants show similar low-level feature attention - **Middle layers (L3)**: Asymmetric variant begins showing more structured patterns - **Late layers (L6)**: Clear differentiation—asymmetric model develops coherent object-level attention while plain model remains diffuse

E Statistical Validation

All reported results were validated using appropriate statistical tests. For performance comparisons, we used paired t-tests with Bonferroni correction for multiple comparisons. Effect sizes were calculated using Cohen's d, with the following interpretations: small (0.2), medium (0.5), large (0.8).

All main results show statistical significance (p ¡ 0.001) with large effect sizes, confirming the robustness of our findings.

To investigate the formation process of the final attention patterns, we visualized the evolution of attention across different training stages (e.g., 20, 60, 100 epochs), as shown in Figure 7. We observe that the attention patterns of the Asymmetric model gradually become more holistic and compre-

Table 12: MLP noise robustness on CIFAR-100

Noise Level	PlainMLP	AsymmetricMLP	Improvement	P-Value
0.0%	0.2564 ± 0.0029	0.2600 ± 0.0034	+1.41%	0.0040
10.0%	0.2476 ± 0.0036	0.2536 ± 0.0047	+2.42%	0.0104
20.0%	0.2383 ± 0.0034	0.2429 ± 0.0038	+1.92%	0.0014
30.0%	0.2254 ± 0.0025	0.2309 ± 0.0036	+2.45%	0.0006
40.0%	0.2097 ± 0.0042	0.2182 ± 0.0036	+4.03%	0.0023
50.0%	0.1908 ± 0.0043	0.2015 ± 0.0028	+5.63%	0.0000
60.0%	0.1652 ± 0.0039	0.1797 ± 0.0044	+8.81%	0.0000
70.0%	0.1276 ± 0.0063	0.1496 ± 0.0058	+17.23%	0.0000
80.0%	0.0816 ± 0.0073	0.0956 ± 0.0049	+17.14%	0.0005
90.0%	0.0360 ± 0.0054	0.0349 ± 0.0043	-3.00%	0.4902

Table 13: Optimal hyperparameters for Table 1

$N_{branches}$	Architecture	Learning Rate	Weight Decay	α
1×	MLP	0.0001	0.001	100.0
	CNN	0.0003	0.01	100.0
	ViT-6L	0.001	0.1	4.642
7×	MLP	0.0001	0.001	0.1
	CNN	0.0003	0.001	0.1
	ViT-6L	0.001	0.01	4.642
10×	MLP	0.0001	0.001	0.1
	CNN	0.0003	0.001	1.0
	ViT-6L	0.001	0.01	4.642
20×	MLP	0.0001	0.001	0.1
	CNN	0.0003	0.001	0.1
	ViT-6L	0.001	0.01	1.0

hensive as training progresses. In contrast, the attention of the Plain model saturates earlier and consistently focuses more on local textures.

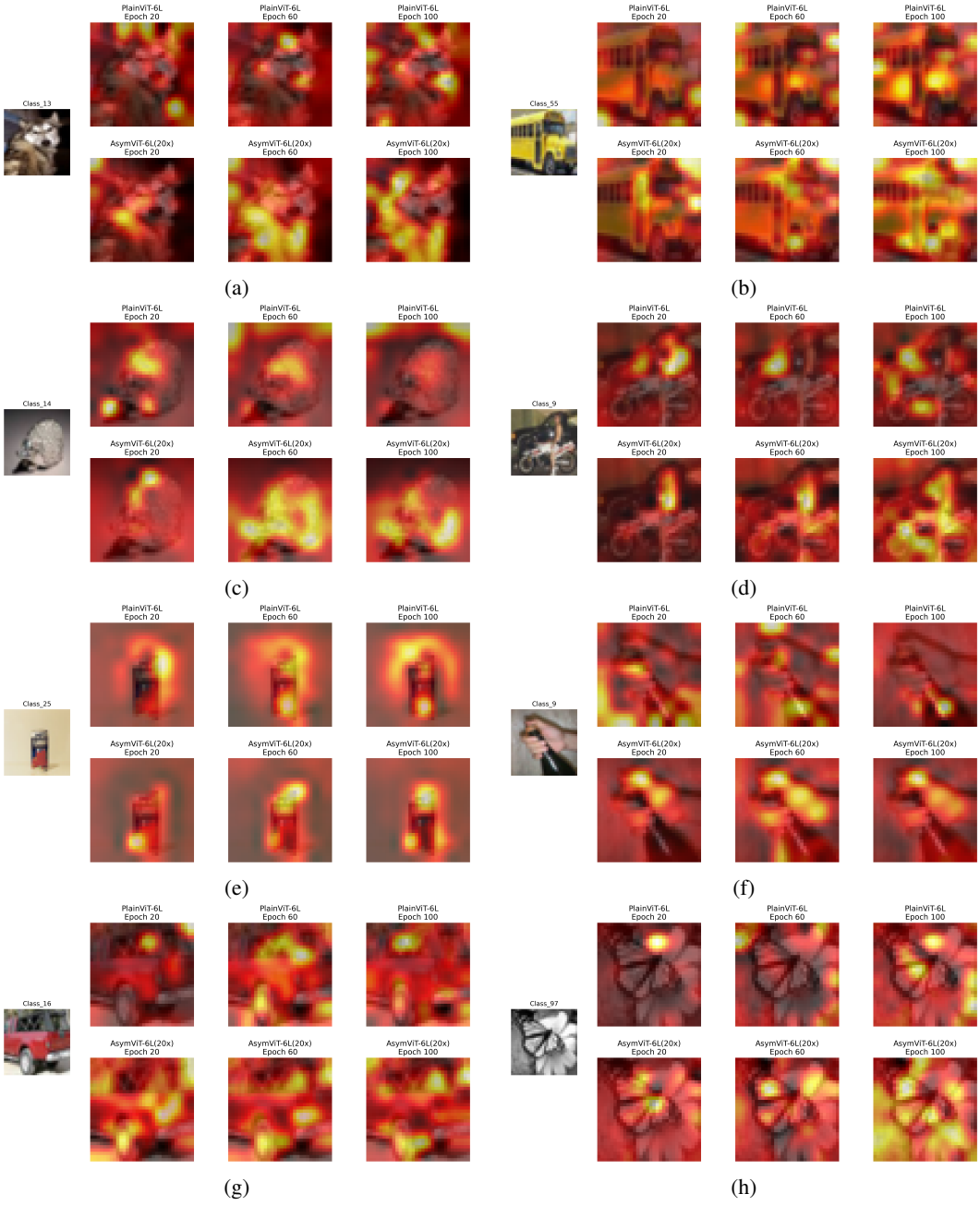


Figure 7: Comprehensive attention pattern evolution for ViT-6L on CIFAR-100.

Table 14: Attention pattern quality metrics across training epochs

Metric	Plain ViT-6L			Asymmetric ViT-6L		
	Epoch 20	Epoch 60	Epoch 100	Epoch 20	Epoch 60	Epoch 100
Peak Strength	0.092	0.063	0.051	0.053	0.049	0.045
Map Entropy	3.73	3.89	3.97	4.02	4.03	4.05
Sparsity (Gini)	0.471	0.415	0.374	0.254	0.287	0.302
Object Coverage	0.58	0.70	0.84	0.96	0.95	0.94

Table 15: Gradient Conflict Analysis. Cosine similarity analysis between main (softmax) and auxiliary (sigmoid) gradients during training. ViT-6L shows consistent positive similarity (synergy), while CNN exhibits strong negative similarity (conflict), and MLP demonstrates near-orthogonal gradients with slight conflict tendency.

Architecture	Final Similarity	Average Similarity	Min Similarity	Max Similarity
MLP	-0.0309	-0.0801	-0.3422	0.0701
CNN	-0.1926	-0.2574	-0.8210	0.0464
ViT-6L	0.2631	0.1870	-0.1845	0.3654

Table 16: To explore the impact of different hyperparameter search strategies, we conducted a supplementary experiment for the $N_{\text{branches}} = 3$ configuration, where hyperparameters (learning rate and weight decay) were independently optimized for both baseline and asymmetric models. The results are shown in the table, where "all active" indicates that all three auxiliary branches at each connection point participate in backpropagation, while "one active" means only one auxiliary branch per connection point is activated during backpropagation. Although this "dual optimization" strategy can yield benefits in certain cases, we consistently adopted the "Pragmatic Gold Standard" strategy throughout the main text to isolate the pure effect of our paradigm.

$N_{branches}$	Architecture	Plain	Asymmetric	Improvement	p-value
3×branches (all active)	MLP	0.246 ± 0.003	0.207 ± 0.005	-16.0%	0.0000
	CNN	0.399 ± 0.006	0.362 ± 0.006	-9.3%	0.0000
	ViT-6L	0.362 ± 0.006	0.421 ± 0.027	+16.2%	0.0002
3×branches (one active)	MLP	0.246 ± 0.003	0.237 ± 0.006	-3.8%	0.0008
	CNN	0.400 ± 0.008	0.339 ± 0.009	-15.1%	0.0000
	ViT-6L	0.365 ± 0.007	0.440 ± 0.059	+20.7%	0.0045

Table 17: Architecture performance comparison with asymmetric training on CIFAR-10. Results show differential architectural responses to auxiliary supervision, with statistical significance assessed using two-tailed paired t-tests across 10 independent runs.

Architecture	Baseline	Asymmetric	Improvement	p-value	Significant
MLP	0.495 ± 0.003	0.506 ± 0.004	+2.3%	0.0001	Yes
CNN	0.770 ± 0.006	0.772 ± 0.010	+0.3%	0.5371	No
ViT-6L	0.625 ± 0.007	0.636 ± 0.004	+1.7%	0.0068	Yes

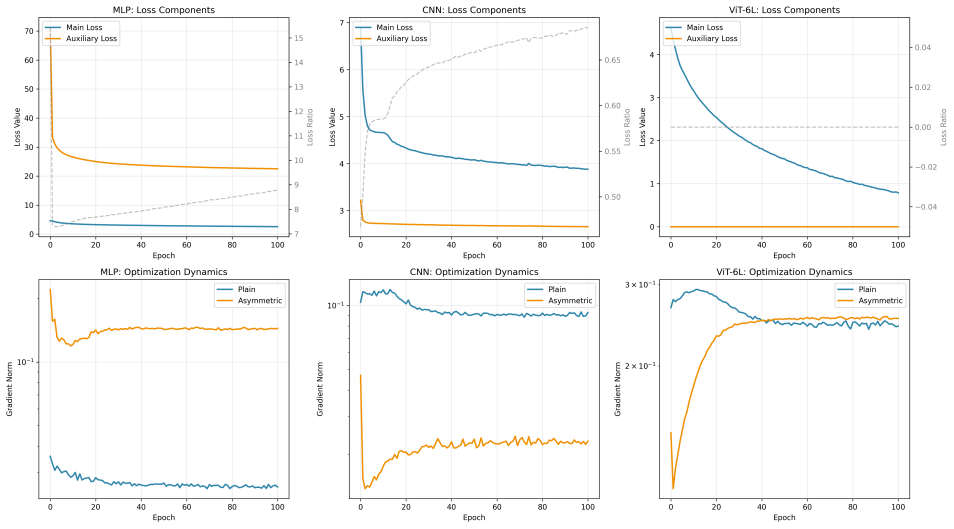


Figure 8: Training dynamics comparison across architectures showing loss components and optimization trajectories. **Top row**: Evolution of main loss (softmax) and auxiliary loss (sigmoid) during training. **Bottom row**: Gradient norm dynamics for plain and asymmetric variants. Asymmetric training exhibits architecture-specific patterns: MLP shows stable auxiliary loss with reduced gradient norms, CNN demonstrates auxiliary loss divergence with increased gradient instability, while ViT-6L displays rapid auxiliary loss convergence with improved optimization stability.

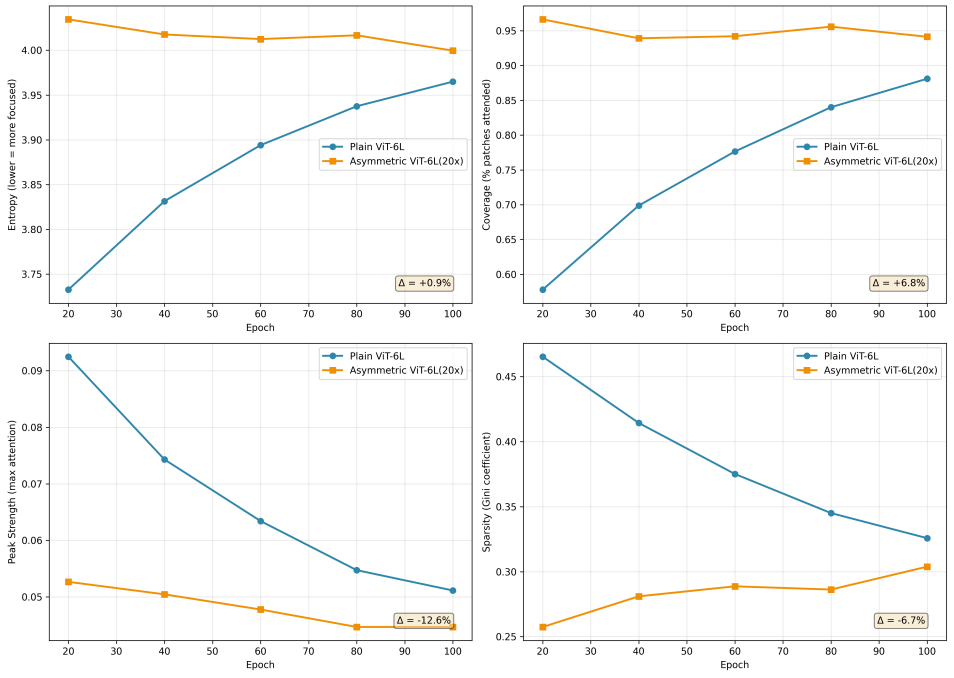


Figure 9: ViT-6L attention mechanism analysis comparing plain and asymmetric training. **Top left**: Entropy evolution showing asymmetric training maintains higher attention diversity. **Top right**: Coverage percentage demonstrating improved spatial attention coverage (+6.8%). **Bottom left**: Peak strength indicating more focused attention patterns (-12.6%). **Bottom right**: Sparsity coefficient revealing attention distribution characteristics (-6.7%). Results suggest asymmetric training promotes more comprehensive yet focused attention patterns.