Towards Robustness against Typographic Attack with Training-free Concept Localization*

Bohan Liu[1], Wenqian Ye[1], Guangzhi Xiong[1], Zhenghao He[1], Sanchit Sinha[1], and Aidong Zhang[1]

[1]University of Virginia, Charlottesville VA 22903, USA
{bohan, aidong}@virginia.edu

Abstract. Models trained via Contrastive Language-Image Pretraining (CLIP) serve as the foundational vision encoders for many modern Large Vision Language Models (LVLMs). Despite their widespread adoption, CLIP models exhibit a critical yet underexplored failure mode: irrelevant text appearing within images confounds visual representations, biasing them toward lexical meaning rather than true visual semantics. This robustness issue, commonly described as a Typographic Attack (TA), exposes a vulnerability that poses a significant risk to safety-critical applications such as autonomous driving. To achieve interpretable and effective robustness against TA, we propose a novel, training-free mechanistic interpretability method. Our method provides sampling-based interpretations of hidden state representations and quantitatively attributes semantic versus lexical focus to individual attention heads. Through probabilistic analysis and circuit mining, we isolate specific Vision Transformer (ViT) components that disproportionately encode lexical information, thereby identifying the mechanistic source of TA. We further show that simple interventions applied directly to the identified circuits, without any additional training, can substantially improve robustness against Typographic Attacks in object classification. These interventions, such as selective adjustment of attention weights, outperform both supervised and training-free defense methods. Our experiments also demonstrate that applying the proposed intervention to the vision encoders of several state-of-the-art LVLMs yields substantial gains in Visual Question Answering accuracy under Typographic Attack interference on RIO-Bench. These results confirm both the efficacy and the generalizability of our mechanistic approach. Code is released at https://github.com/Liu-524/SamplingTAR.

Keywords: Typographic Attack · Vision Language Models · Mechanistic Interpretability

1 Introduction

Contrastive Language-Image Pretraining (CLIP) [27] has become the cornerstone of modern visual perception systems. The simple contrastive objective of

* Accepted to the European Conference on Computer Vision (ECCV) 2026.

CLIP enables large-scale parametric learning and the alignment of vision and language. CLIP's strong out-of-distribution generalization enables its use as a zero-shot backbone for diverse visual recognition and perception tasks. Consequently, CLIP vision encoders have powered several state-of-the-art Large Vision Language Models (LVLMs), including LLaVA [24], Qwen-VL [2], and InternVL [5].

Despite their critical role in increasingly high-stakes LVLM applications [34, 39], our understanding of the internal mechanism of CLIP models remains limited. Given that most CLIP variants are fundamentally grounded in the Vision Transformer (ViT) architecture [10], investigating the ViT encoder is a central imperative [32]. Previous interpretability efforts have focused on residual stream decomposition [15, 16] or hidden space disentanglement with Sparse Dictionary Learning (SDL) [11, 15, 19, 40]. However, these approaches largely analyze the network at a macro level, treating layers as indivisible blocks. Few studies have directly probed the internal attention mechanisms of ViTs to uncover exactly how individual visual patches interact.

While highly capable of aligning vision and language, CLIP models have been shown to exhibit a distinct vulnerability, namely Typographic Attacks (TA) [18]. In contrast to conventional imperceptible adversarial perturbations, typographic attacks exploit the model's recognition capacity by imposing deceptive text onto an image. For instance, an image of the class "cat" with the injected word "Goose" usually leads to misclassification (see Fig. 1a). This vulnerability highlights a critical feature entanglement in the vision encoders, in which the encoders cannot effectively decouple the visual features from the semantics encoded in lexical shapes. Therefore, resolving this issue requires methods beyond standard black-box defenses. This motivates us to propose a mechanistic method to precisely determine where and how typographic features arise and interfere with visual semantic processing during encoding.

A recent study on the Linear Representation Hypothesis [12, 26] highlights that latent representations can be interpreted as linear combinations of concept directions. Building on this hypothesis, we introduce the stochastic sampling of pseudo-concept vectors as a training-free approach to interpret transformer modules. Additionally, we propose an attribution-based lexical-focus ranking mechanism to identify TA-vulnerable concept samples and pinpoint their host modules. We perform a systematic analysis of stochastic concept mining and leverage dimensionality reduction in Multi-Head Self-Attention (MHSA) projections to reduce the search space and mitigate concept entanglement. The resulting attribution-based label-free circuit mining method provides faithful and explainable TA robustness to CLIP ViT and LVLMs through simple and fast test-time circuit intervention. We summarize our contributions as follows:

- **Sampling-based Concept Mining.** Building on the Linear Representation Hypothesis [12, 26] and Sparse Autoencoders, we provide a training-free stochastic sampling method of concept mining and demonstrate that leveraging the natural decomposition of representation in Multi-Head Self-Attention modules provides a significant increase in the likelihood of a concept hit.

- **Model Vulnerability Explanation.** We propose a gradient-based attribution method that jointly considers the attention head "focus" and the concept "direction" to obtain a normalized attribution score that consistently identifies lexical-focusing concepts and attention modules.
- **Mechanistic Intervention for Improved Robustness.** Attention reweighting and ablation on harmful circuits significantly improve CLIP's object classification accuracy and LVLM's Visual Question-Answering (VQA) accuracy on typographic attack datasets. Our method is shown to outperform prior state-of-the-art methods and improve large LVLMs at minimal cost.

2 Related Works

2.1 Interpretability on Attention Heads

Multi-Head Self Attention (MHSA) [33] has become a fundamental building block of Transformers. Conceptually, MHSA creates parallel, lower-dimensional views of the same input data, applies separate attention operations to each view, and combines the views across heads. Despite its widespread adoption, the interaction between attention heads and their distinct behaviors remains underexplored. [23] applies attention-head pruning to reveal the importance of heads for specific tasks in a multi-task learning setup and exploits head-task affinity to improve multi-task learning. Specifically, in the context of CLIP model interpretability, [15] studies the decomposition of heads and layers in CLIP-ViT models by doing a greedy search on a set of text descriptions that maximizes the explained variance of each head's latent space. More recently, [30] exploits the output and value projections to the residual stream [13] of transformers to directly analyze the attention head's concept vectors in a learned dictionary. [22] extends self-supervised interpretation by directly applying a Sparse Dictionary Learning (SDL) to the attention layer outputs, with attributes based on the norms of the decoder channels. Based on the SDL intuition and the Linear Representation Hypothesis [12, 26], our work investigates stochastic sampling as a model interpretability tool when combined with mechanistic attribution and reduces the computational and data costs of parametric dictionary learning. Our method tailors the model interpretation process to the ViT architecture by jointly considering attention weighting and concept alignment and provides faithful attribution to TA-vulnerable modules based on the spatial information flow within ViT.

2.2 Typographic-Attack Robustness

Typographic-Attack Robustness has received increasing attention as the development and application of complex LVLMs have advanced, which rely on robust CLIP visual embeddings. Evaluations on TypoD [6] confirm the prevalence of TA vulnerability in CLIP ViT and LVLMs. [36] extends Typographic Attack to a multi-image setting. [7] applies typographic attack techniques to self-driving scenarios, and SceneTap [4] extends the scope of Typographic Attack

Table 1: Comparison of defense properties within the ViT paradigm. Our method leverages the specific architectural priors of Transformers to achieve robustness without requiring labeled greedy searches.

Method	Training-Free	Low Data	Interpretable	Intervention Cost
Defense-Prefix [1]	×	×	×	High
Dyslexify [20]	✓	×	✓	High (Iterative)
Ours	✓	✓	✓	**Low (Constant)**

in self-driving to the occurrence of realistic and coherent text distractors in images. Shown in Tab. 1, on the defense side, Defense-Prefix tuning [1] applies prompt token learning on annotated data to trigger lexical overlook behavior of CLIP models. Recently, advances in TA robustness [20] have shifted toward explainable, training-free interventions for vulnerable modules. Our work proposes optimization-free stochastic sampling as a replacement for expensive parametric dictionary learning and pushes the boundary of training-free intervention for TA robustness.

3 Method

In this section, we introduce and analyze our training-free model interpretation and robustness process, as shown in Fig. 1. Starting from the ViT architecture and the MHSA mechanism, we analyze the behavior of the concept vector sampling process, as visualized in Fig. 1a, under the Linear Representation Hypothesis. The circuit mining and intervention process outlined in Fig. 1b and Fig. 1c is then established based on gradient attribution and the ViT architecture.

3.1 ViT Architecture

To establish the foundation of our method, we briefly review the Multi-Head Self-Attention (MHSA) definition and the Residual Stream view of a Transformer [13]. Let the input to a specific layer of a ViT be $\mathbf{X} \in \mathbb{R}^{N \times d_{model}}$, where N is the sequence length, and d_{model} is the global residual stream dimension. The MHSA block decomposes the representation into H independent heads, projecting the data into a lower-dimensional subspace $d_{\text{head}} = d_{model}/H$. For a given head h, the computation is governed by the Query-Key (QK) routing circuit and the Output-Value (OV) feature circuit. The QK circuit determines the pre-softmax attention logits $s_{i,j}$ from destination token i to source token j:

$$s_{i,j} = \frac{(\mathbf{x}_i \mathbf{W}_Q^h)(\mathbf{x}_j \mathbf{W}_K^h)^T}{\sqrt{d_{\text{head}}}} \tag{1}$$

where $\mathbf{x}_i, \mathbf{x}_j$ are row vectors from $\mathbf{X}$, and $\mathbf{W}_Q^h, \mathbf{W}_K^h, \mathbf{W}_V^h \in \mathbb{R}^{d_{model} \times d_{\text{head}}}$ are the learned projection matrices. The routing probabilities are obtained via softmax:

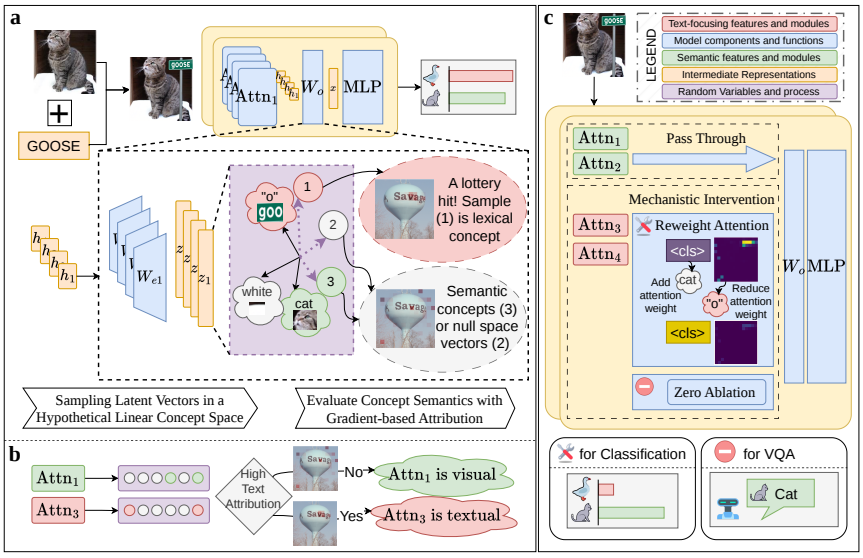

Fig. 1: An overview of the proposed method. a) shows the sampling-based mechanistic interpretability (the **Stochastic Lottery**) for lexical circuit mining. Latent samples (denoted by purple dashed arrows) on the hypothesized concept basis reveal distinct attribution patterns. **b)** illustrates the overall sampling and circuit-mining process using **gradient-based attribution**. **c)** shows the **mechanistic intervention** process. To improve TA robustness, we intervene in vulnerable attention modules via attention reweighting or zero-ablation.

$A_{i,j} = \frac{\exp(s_{i,j})}{\sum_k \exp(s_{i,k})}$. Concurrently, the OV circuit extracts the local semantic feature vector at patch j: $\mathbf{v}_j = \mathbf{x}_j \mathbf{W}_V^h$. The head output vector written to destination token i is $\mathbf{o}_i = \sum_{j=1}^{N} A_{i,j} \mathbf{v}_j$.

3.2 Mechanistic Defense Task: Typographic Attack Robustness

To motivate our interpretability framework, we formally define the task of mechanistic defense against typographic attacks in vision-language models like CLIP. A typographic attack occurs when a model is fooled by an adversarial text superimposed on an image (e.g., a cat labeled "Goose" being classified as a goose). Our goal is to localize and mitigate this vulnerability by decomposing the ViT as a collection of functional modules.

ViT as Nested Functional Modules. We formalize the ViT $\mathcal{M}$ as a set $\mathcal{B}$ comprising a nested hierarchy of functional modules $\mathcal{B} = \{c_1, c_2, \ldots, c_m\}$. These modules range from coarse structures (e.g., full residual blocks) to fine-grained components (e.g., attention modules and Multi-Layer Perceptrons (MLPs)). Given a dataset distribution of clean input images $\mathbf{x} \sim \mathcal{D}$ with true semantic labels y_{img}, a task model $\mathcal{F}_{\mathcal{M}}$ based on $\mathcal{M}$, and a typographic perturbation

function $\mathcal{T}(\mathbf{x}, y_{\text{text}})$ that superimposes the text distractor y_{text} onto $\mathbf{x}$, the vulnerability is defined by the model prioritizing the text over the visual semantic content across the distribution:

$$\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}\big(\mathcal{F}_{\mathcal{M}}(\mathcal{T}(\mathbf{x}, y_{\text{text}})) = y_{\text{text}}\big) \gg \mathbb{P}_{\mathbf{x}\sim\mathcal{D}}\big(\mathcal{F}_{\mathcal{M}}(\mathcal{T}(\mathbf{x}, y_{\text{text}})) = y_{\text{img}}\big) \tag{2}$$

Circuit Extraction and Intervention Objective. Mechanistic defense posits that this vulnerability is not distributed uniformly across the network but is localized within a specific "typographic reading circuit." The defense task is to extract a minimal subset of computational modules $\mathcal{C}_{\text{lex}} \subset \mathcal{B}$ that route and propagate the adversarial lexical signal. Let $\mathcal{I}$ be an intervention function applied specifically to the nodes in $\mathcal{C}_{\text{lex}}$ of ViT $\mathcal{M}$. Such interventions include activation ablation, attention reweighting, and others. The intervened model is denoted as $\mathcal{F}_{\mathcal{I}(\mathcal{M},\mathcal{C}_{\text{lex}})}$. The mechanistic defense objective is to find the optimal subset $\mathcal{C}^*_{\text{lex}}$ that maximizes robustness against the attack while minimizing the degradation of general capabilities on clean data, constrained by the sparsity of the localized circuit:

$$\mathcal{C}^*_{\text{lex}} = \arg\min_{\mathcal{C}_{\text{lex}}\subset\mathcal{B}} |\mathcal{C}_{\text{lex}}| \quad \text{s.t.} \quad \begin{cases} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\mathcal{L}\left(\mathcal{F}_{\mathcal{I}(\mathcal{M},\mathcal{C}_{\text{lex}})}(\mathcal{T}(\mathbf{x}, y_{\text{text}})), y_{\text{img}}\right)\right] < \epsilon_{robust} \\ \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\mathcal{L}\left(\mathcal{F}_{\mathcal{I}(\mathcal{M},\mathcal{C}_{\text{lex}})}(\mathbf{x}), y_{\text{img}}\right)\right] < \epsilon_{benign} \end{cases} \tag{3}$$

where $\mathcal{L}$ is the utility loss, e.g., Classification Error, ϵ_{robust} guarantees TA robustness, and ϵ_{benign} ensures the preservation of the model's core vision capabilities. To efficiently discover the members of $\mathcal{C}^*_{\text{lex}}$, we must systematically trace the attribution of the y_{text} signal back through the computational graph.

3.3 Information Flow Attribution

To trace the adversarial routing circuit, we attribute the flow of specific semantic concepts to the routing decisions made by the QK circuit. We probe the head's d_{head}-dimensional output space with a direction vector $\mathbf{u} \in \mathbb{R}^{d_{\text{head}}}$. We define the projected concept magnitude at source patch j as $V_j(\mathbf{u}) = \langle \mathbf{v}_j, \mathbf{u} \rangle$, and the concept strength aggregated at destination token i as:

$$F_i(\mathbf{u}) = \langle \mathbf{o}_i, \mathbf{u} \rangle = \sum_{j=1}^{N} A_{i,j} V_j(\mathbf{u}) \tag{4}$$

We formalize attribution as the partial derivative of the concept strength with respect to the pre-softmax logit $s_{i,j}$. Applying the chain rule through the softmax function yields the following attribution map:

$$\begin{aligned} \frac{\partial F_i(\mathbf{u})}{\partial s_{i,j}} &= \sum_k \frac{\partial A_{i,k}}{\partial s_{i,j}} V_k(\mathbf{u}) \\ &= A_{i,j}(1 - A_{i,j}) V_j(\mathbf{u}) - \sum_{k\neq j} A_{i,j} A_{i,k} V_k(\mathbf{u}) \\ &= A_{i,j}\big(V_j(\mathbf{u}) - F_i(\mathbf{u})\big) \end{aligned} \tag{5}$$

Interpretation. The gradient attribution of a patch's attention logit to the pseudo-concept vector $\mathbf{u}$ can be transformed into the product of the routing gate ($A_{i,j}$) and the marginal utility of the patch ($V_j(\mathbf{u}) - F_i(\mathbf{u})$). This formulation jointly considers the attention activation and concept alignment to ensure faithful attribution.

3.4 The Stochastic Lottery and Polysemantic Interference

A core difficulty in the efficient application of concept-based interpretation to deep neural networks is the acquisition of the concept vector $\mathbf{u}$ without computationally expensive and data-intensive dictionary learning. To relax this requirement and promote training-free model interpretability, we apply stochastic sampling to mechanistic interpretability research. Based on the Linear Representation Hypothesis [12, 26] that the dense hidden space representation is a superposition of several distinct concept vectors, we propose to sample and evaluate random hidden space vectors for interpreting neural network behaviors. This formulation intuitively connects to the Lottery Ticket Hypothesis [14]: By considering random vector samples as the linear probe initialization and the lexical information as the learning target, there exists a sparse subset of the model, in this case, a subset of sampled vectors, that preserves the model capacity to achieve the learning target. The primary mathematical threat to this stochastic lottery is polysemantic interference. For the analysis, we generate K independent random vectors per head, denoting each as $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \frac{1}{d_{\text{head}}}\mathbf{I})$. Let us isolate a target concept $\mathbf{c}_{\text{target}} \in \mathbb{R}^{d_{\text{head}}}$ ($\|\mathbf{c}_{\text{target}}\| = 1$). We orthogonally decompose a patch i's value vector $\mathbf{v}_i$:

$$\mathbf{v}_i = \alpha_i \mathbf{c}_{\text{target}} + \boldsymbol{\xi}_i, \tag{6}$$

where $\alpha_i \in \mathbb{R}$ is the true concept strength and $\boldsymbol{\xi}_i \in \mathbb{R}^{d_{\text{head}}}$ is the polysemantic interference vector, such that $\langle \mathbf{c}_{\text{target}}, \boldsymbol{\xi}_i \rangle = 0$. When projecting onto a random feature direction $\mathbf{u}$, the activation is $\langle \mathbf{v}_i, \mathbf{u} \rangle$:

$$Z_i \propto \langle \mathbf{v}_i, \mathbf{u} \rangle = \underbrace{\alpha_i \langle \mathbf{c}_{\text{target}}, \mathbf{u} \rangle}_{\text{Signal } S_{\mathbf{u},i}} + \underbrace{\langle \boldsymbol{\xi}_i, \mathbf{u} \rangle}_{\text{Interference } I_{\mathbf{u},i}}, \tag{7}$$

where $S_{\mathbf{u},i}$ is the true concept signal and $I_{\mathbf{u},i}$ denotes the polysemantic interference.

Bounding Interference via Concentration of Measure. By standard Gaussian tail bounds, the maximum interference across all N patches is bounded. Denoting $\|\boldsymbol{\xi}_{\max}\| = \max_{i \in \{1 \dots N\}} \|\boldsymbol{\xi}_i\|$ as the largest interference-vector norm across patches, the maximum error in estimating the true concept strength is bounded with high probability:

$$\max_{i \in \{1 \dots N\}} |I_{\mathbf{u},i}| \leq \|\boldsymbol{\xi}_{\max}\| \sqrt{\frac{4 \log N}{d_{\text{head}}}} \tag{8}$$

For a clean attribution map, the signal of the weakest on-concept patch $k = \arg\min_{i \in \text{on}} \alpha_i$ must dominate the maximum interference by a margin τ:

$$S_{\mathbf{u},k} > \max_i |I_{\mathbf{u},i}| + \tau. \tag{9}$$

The probability that a single random probe satisfies Eq. (9) is

$$p \approx \exp\left(-\frac{\left(\|\boldsymbol{\xi}_{\max}\|\sqrt{4\log N} + \tau\sqrt{d_{\text{head}}}\right)^2}{2(\alpha_k)^2} \right). \tag{10}$$

Since every on-concept patch carries the concept at least as strongly as α_k, a probe satisfying Eq. (9) resolves all of them simultaneously, yielding a concept-faithful attribution map. By construction of the text-injected inputs, the injected region carries the lexical concept strongly, so α_k is bounded well above zero. We defer the detailed derivation to the Appendix.

The Advantage of the MHSA Bottleneck. Considering a sample size of K, we define success as the event that at least one sampled $\mathbf{u}$ satisfies Eq. (9) with respect to $\mathbf{c}_{\text{target}}$. Then, to achieve a higher probability of finding clean feature maps (P_{success}), the sample-and-interpret process requires K lottery tickets per head:

$$K \geq \frac{\log(1 - P_{\text{success}})}{\log(1 - p)} \tag{11}$$

This inequality justifies operating in d_{head} rather than $d_{\text{model}} \gg d_{\text{head}}$. In the global stream, the interference norm $\|\boldsymbol{\xi}_{\max}\|^2$ contains the sum of all features across all heads, causing the required K to grow rapidly. Isolating the attention-head-specific subspace pushes head-excluded concepts into the null space, which significantly reduces $\|\boldsymbol{\xi}_{\max}\|^2$, increases p, and allows a computationally feasible K to reliably obtain concept basis vectors as winning tickets, yielding highly interpretable semantic vectors without parameter training.

3.5 Mechanistic Intervention

Given extracted lexical circuits consisting of attention heads with high text attribution, we perform attention reweighting on each attention head to simultaneously maintain representation consistency while reducing typographic information in the object classification task, and apply zero ablation to remove lexical distraction from all patch tokens under VQA evaluation, as shown in Fig. 1c.

Attention Reweighting. Following Dyslexify [20], for input sequence of length N, given the original attention map for the <cls> token $\mathbf{a}_{<\text{cls}>} = [A_{0,0}, A_{0,1}, A_{0,2}, \ldots A_{0,N-1}] \in \mathbb{R}^N$ and a control parameter $a \in [0, 1]$, the reweighted attention map $\mathbf{a}'_{<\text{cls}>} \in \mathbb{R}^N$ is

$$\mathbf{a}'_{<\text{cls}>}[i] = a \cdot \mathbb{I}(i = 0) + \mathbf{a}_{<\text{cls}>}[i] \frac{1 - a}{\sum_{j=1}^{N-1} \mathbf{a}_{<\text{cls}>}[j]} \cdot \mathbb{I}(i \neq 0) \tag{12}$$

where the i and j denote the token location in the attention maps, and the token order follows the ViT definition [10] with the first token being the <cls> token and the following being patch tokens. The reweighted attention map effectively reduces the information passed from textual patch tokens in a text-focused attention head, thereby suppressing the lexical circuit. Throughout our experiments, we set $a = 1$ to achieve maximal intervention.

Zero Ablation. Attention Reweighting does not apply to LVLMs, which often do not implement and train a dedicated class token and instead rely on all patch token embeddings for visual information processing. Therefore, to evaluate the quality and faithfulness of the extracted lexical circuits, we use simple zero ablation, in which a zero vector replaces the output of the vulnerable modules.

4 Experiments

To validate the efficacy of our method in extracting faithful typographic circuits and improving visual perception under lexical interference, we conduct object classification experiments using five ViT CLIP models of varying sizes pretrained on LAION-2B [28] and VQA experiments with several popular LVLMs.

4.1 Experimental Settings

Evaluation Datasets. Following prior works [1,20], we use RTA-100 [1], Disentangling [25], and PAINT [21] datasets for the object classification experiments. In addition, we construct a new dataset IN-100-Text from ImageNet-100 [31] by adding realistic, contextually coherent text distractors using Qwen-Image-Edit [38]. For each image, we sample a label from the IN-100 class vocabulary to create a semantic conflict between the visual content and the lexical distractor, and render the label in one of seven diverse visual styles detailed in the Appendix. The pixel-level edit area with a max-channel difference greater than 20/255 has a mean/median of 16.3%/11.5%. For VQA evaluation, we use RIO-Bench [37], a large VQA dataset consisting of TA images. We specifically evaluate our method on the attacked multiple-choice VQA, comprising easy, medium, and hard subsets.

Baseline Methods. We demonstrate the advantage of our proposed method in improving TA robustness and enhancing model performance under lexical interference by comparing to Defense-Prefix [1] and Dyslexify [20] for object classification. Defense-Prefix Training is a supervised method that learns a defense-prefix token embedding. Dyslexify [20] is a training-free method based on attention statistics and greedy circuit mining. To further demonstrate the efficacy of our method, we apply it to several LVLMs and report performance with and without our intervention.

Implementation Details. For object classification experiments, we test our method on ViT-B/16, ViT-L/14, ViT-H/14, ViT-g/14, and ViT-bigG/14 models pretrained on the LAION-2B [28] dataset. Based on prior observation that concept emerges in the late stages of ViTs [15], we systematically examine the last 20% of the transformer blocks in the ViTs. For each attention module, we sample 16 times the hidden dimension of random concept vectors and perform gradient-based attribution as described in Sec. 3.3. We use the ImageNet-1K [9] training set as the base dataset for text injection augmentation. Only 0.1% of the training set is used in each run. The one-time extraction runs on 1,280 images and takes **under one minute on a single A100** across all five backbones (ViT-B/16: 7.5 s; ViT-L/14: 14.0 s; ViT-H/14: 24.0 s; ViT-g/14: 32.1 s; ViT-bigG/14: 45.7 s). At test-time, the intervention incurs **near-zero overhead** because it operates on a fixed set of head indices.

4.2 Vulnerability Identification

Following Eq. (5), we consider in a practical setup with an input image I with known text location $M \subseteq \{0, \dots N-1\}$, the text-corresponding patch locations yield the text mask $\mathbf{m}^+ \in \mathbb{R}^N$ with $\mathbf{m}[p] = \mathbb{1}_M(p)$, where p is the token index and $\mathbb{1}_M$ is the indicator function given the set M. Then the inverse text mask $\mathbf{m}^- = \mathbf{1}_N - \mathbf{m}^+$, with the exception that the class token mask value is always zero. The Mask Attribution Score (MAS) and normalized Text Attribution Score (nTAS) with respect to the destination token index i are defined as:

$$\mathrm{MAS}_i(\mathbf{u}, \mathbf{m}) = \frac{1}{\langle \mathbf{1}_N, \mathbf{m} \rangle} \sum_{j=1}^{N} \mathrm{ReLU}\left(\frac{\partial F_i(\mathbf{u})}{\partial s_{i,j}}\right) \cdot \mathbf{m}[j], \tag{13}$$

$$\mathrm{nTAS}_i(\mathbf{u}, \mathbf{m}^+, \mathbf{m}^-) = \frac{\mathrm{MAS}_i(\mathbf{u}, \mathbf{m}^+)}{\mathrm{MAS}_i(\mathbf{u}, \mathbf{m}^+) + \mathrm{MAS}_i(\mathbf{u}, \mathbf{m}^-)}. \tag{14}$$

where $\mathbf{m}$ is a vector mask and $\mathbf{m}[\cdot]$ indexes the mask. In Eq. (13), the MAS considers only positive gradient attribution to the patch token sequence, consistent with the additive nature of the Linear Representation Hypothesis [12, 26]. Normalized TAS jointly accounts for object and lexical focus and scales the score to the $[0, 1]$ range. We take the mean of all sample vector $\mathbf{u}$ scores over an augmented unlabeled dataset as the score of the tested module. Attention head modules within a layer are compared with a simple z-test. The z-score threshold is either selected with a calibration dataset or defaults to one. This choice is not critical near the default: on ViT-H/14 (RTA-100), $z \in \{0.5, 1.0, 2.0\}$ selects 24/15/4 heads, yielding Object Classification Accuracy (OCA) = 75.9/76.3/66.5% and Text Confusion Rate (TCR) = 14.0/15.4/27.1%. The clean ImageNet-100 classification accuracy remains at 83.2/83.3/83.9%. Robustness and clean accuracy are stable around z = 1, and only at z = 2 do too few heads remain in the lexical circuit.

Table 2: Robustness improvement across ViT backbones on Zero-shot image classification. **OCA**: Object Classification Accuracy (↑), **TCR**: Text Confusion Rate (↓). The green colored numbers indicate the improvement in OCA, and the blue colored numbers indicate a decrease in TCR. **w. Int.** indicate *with intervention*.

Model	Method	RTA-100		Disentangling		PAINT		IN-100-Text	
		OCA(↑)	TCR(↓)	OCA(↑)	TCR(↓)	OCA(↑)	TCR(↓)	OCA(↑)	TCR(↓)
ViT-B/16	Baseline	56.3	30.8	52.2	44.4	60.2	33.0	54.6	33.0
	w. Int.	68.7 (+12.4)	12.6 (-18.2)	88.3 (+36.1)	11.7 (-32.7)	73.8 (+13.6)	16.5 (-16.5)	74.2 (+19.6)	7.1 (-25.9)
ViT-L/14	Baseline	54.6	39.0	51.7	47.8	61.2	33.0	58.2	32.9
	w. Int.	68.9 (+14.3)	21.2 (-17.8)	68.3 (+16.6)	31.1 (-16.7)	68.9 (+7.7)	22.3 (-10.7)	74.9 (+16.7)	12.1 (-20.8)
ViT-H/14	Baseline	53.4	42.0	46.1	53.3	49.5	46.6	56.6	36.9
	w. Int.	76.2 (+22.8)	14.4 (-27.6)	82.2 (+36.1)	17.2 (-36.1)	75.7 (+26.2)	14.6 (-32.0)	79.1 (+22.5)	9.5 (-27.4)
ViT-g/14	Baseline	50.3	45.8	58.3	41.1	53.4	39.8	57.0	36.4
	w. Int.	68.8 (+18.5)	23.4 (-22.4)	81.7 (+23.4)	17.8 (-23.3)	75.7 (+22.3)	18.4 (-21.4)	76.4 (+19.4)	12.7 (-23.7)
ViT-bigG/14	Baseline	61.0	32.5	48.3	51.1	49.5	38.8	62.3	31.0
	w. Int.	75.7 (+14.7)	15.3 (-17.2)	72.8 (+24.5)	26.7 (-24.4)	79.6 (+30.1)	10.7 (-28.1)	80.6 (+18.3)	8.9 (-22.1)

4.3 Improved TA Robustness

In Tab. 2, we demonstrate the efficacy of our method in improving the TA robustness of various CLIP ViT models with two metrics: object classification accuracy and text confusion rate. Object classification accuracy measures the classification accuracy on the target object depicted in the input image. Text confusion rate measures the accuracy of the models in predicting the text label of the text distractor, and a lower confusion rate indicates a better suppression of the lexical confusion signal. As reported in Tab. 2, our method improves the object classification accuracy of all five tested CLIP ViT models from the base model to the bigG variant. The object classification accuracy increases significantly, accompanied by a large decline in the text confusion rate. With minimal data and computational requirements, our method remains competitive with other methods. Notably, our method incurs an accuracy trade-off of less than 1% in exchange for a significant gain in robustness, as shown in the Appendix.

4.4 Object Classification Evaluation

To make a fair comparison with Dyslexify [20] and to fit the training-free test setup, we use the same 0.1% fraction of the ImageNet-1K training set as the base dataset to compute attention scores for its greedy circuit mining. As shown in Tab. 3, our method improves object classification accuracy across CLIP ViT variants and achieves a significant boost in average robustness over the supervised baseline. With concept-based mechanistic interpretability, our method achieves higher average robustness than the prior method [20]. Our joint modeling of the attention mechanism and concept distribution for circuit mining effectively isolates TA vulnerability in the transformer. This enables simple statistical vulnerability detection that requires minimal labeled data and provides greater TA robustness (an average OCA improvement of 6.1%) than a greedy search guided by large labeled image sets on noisy attention maps. We report detailed trade-

Table 3: Comparison across methods against typographic attacks. Numbers are zero-shot classification accuracy in percentage. * indicates results reported by Dyslexify [20], which requires iterative evaluation on full ImageNet-100 training set.

Method	Model	RTA-100	Disentangling	PAINT	IN-100-Text	IN-100
Defense-Prefix [1]	ViT-B/16	65.6	84.4	68.0	69.9	*75.4*
	ViT-L/14	62.9	*77.2*	71.8	71.6	*79.8*
	ViT-H/14	63.5	67.2	64.1	70.4	*83.4*
	ViT-g/14	58.1	60.0	66.0	66.7	83.2
	ViT-bigG/14	67.8	50.0	68.9	71.7	*85.4*
	Average	63.6	67.8	67.8	70.1	**81.4**
Dyslexify [20]	ViT-B/16	67.9	83.9	69.9	73.5	75.3
	ViT-L/14	66.6	68.9	*72.8*	74.4	79.5
	ViT-H/14	66.1	53.9	68.9	70.3	*83.4*
	ViT-g/14	67.0	75.6	77.7	75.0	83.0
	ViT-bigG/14	67.5	55.6	61.2	70.6	85.0
	Average	67.0	67.6	70.1	72.8	81.3
Dyslexify [20]*	ViT-B/16	68.3	85.0	72.7	-	75.0
	ViT-L/14	71.0	60.6	76.4	-	79.5
	ViT-H/14	68.3	72.2	70.9	-	*83.4*
	ViT-g/14	62.0	67.2	71.8	-	82.6
	ViT-bigG/14	72.9	68.3	69.1	-	84.7
	Average	68.5	70.7	72.2	-	81.0
Ours (nTAS)	ViT-B/16	*68.7*	*88.3*	*73.8*	*74.2*	74.2
	ViT-L/14	*68.9*	68.3	68.9	*74.9*	*79.8*
	ViT-H/14	*76.2*	*82.2*	*75.7*	*79.1*	82.9
	ViT-g/14	*68.8*	*81.7*	75.7	*76.4*	*83.3*
	ViT-bigG/14	*75.7*	*72.8*	*79.6*	*80.6*	84.6
	Average	**71.7**	**78.7**	**74.7**	**77.0**	81.0

offs on clean image perception in the Appendix. Our method incurs minimal accuracy trade-offs, consistent with other methods.

4.5 Application on LVLMs

In order to further demonstrate the efficacy and application of our method, we apply the circuit mining pipeline to the vision encoder of several latest LVLMs that follow the standard ViT-MLP-LLM architecture: Qwen3-VL [3], InternVL3.5 [35], and Gemma3 [17]. Given that many LVLMs do not implement a <cls> token, we perform attribution based on the first visual token. As ViTs are known to repurpose redundant patch tokens to process global information [8], the first visual token, fixed at the top-left corner, is highly likely to serve as an emergent <cls> surrogate. We compare our method with a vanilla LVLM without intervention. In Tab. 4, we show the VQA accuracy on RIO-Bench *obj-attack* split, which requires the models to ignore the text literal in the image and answer questions based on the visual cues. Our method improves the VQA accuracy across Qwen3-VL and Gemma3 model variants. The tradeoff on clean image VQA is further tested on RIO-Bench *clean* split. Our method incurs a minimal -0.3% to $+0.6\%$ effect on clean image VQA, with detailed numbers in the Appendix.

Table 4: VQA accuracy on RIO-Bench splits. Δ represents the improvement (Ours − Base). Improvements > 0.2 are highlighted in bold.

Model	Easy			Medium			Hard			Overall Average		
	Ours	Base	Δ ↑	Ours	Base	Δ ↑	Ours	Base	Δ ↑	Ours	Base	Δ ↑
Qwen3-VL-4B	69.71	68.29	**1.42**	66.52	65.67	**0.85**	55.53	54.90	**0.63**	63.92	62.95	**0.97**
Qwen3-VL-8B	75.81	74.19	**1.62**	73.34	71.23	**2.11**	65.92	64.91	**1.01**	71.69	70.11	**1.58**
Qwen3-VL-30B-A3B	69.11	67.34	**1.77**	66.87	66.71	0.16	62.35	60.90	**1.45**	66.11	64.98	**1.13**
InternVL3.5-8B	63.96	63.87	0.09	61.18	60.77	**0.41**	50.28	50.09	0.19	58.47	58.24	**0.23**
InternVL3.5-14B	57.01	56.95	0.06	55.37	55.65	-0.28	46.87	46.62	**0.25**	53.08	53.07	0.01
Gemma3-4B	49.72	47.98	**1.74**	47.98	46.46	**1.52**	40.52	38.60	**1.92**	46.07	44.35	**1.73**
Gemma3-12B	52.27	49.91	**2.36**	49.84	48.74	**1.10**	47.22	46.08	**1.14**	49.78	48.24	**1.53**

Fig. 2: Attribution Map of Aligned and Unaligned Concept Vectors. Left: Image patches are attributed with an unaligned concept vector. The resulting attribution map does not create an interpretable region. **Right:** The image patches are attributed with an aligned vector that produces the nTAS score of 0.8980. The resulting attribution map is highly concentrated in the lexical shapes in the images.

5 Discussion and Analysis

5.1 Qualitative Study

Fig. 2 shows that nTAS is robust against unaligned concept directions, as attributing an off-concept vector produces a diffuse, noisy attribution map without semantic focus. Beyond this null-case check, we further validate the extracted circuits and the capacity of the sampling and ranking process in Fig. 3, which shows the attribution maps generated by selected concept vectors on their activating images: when contrasting high-scoring nTAS cohorts with their low-scoring counterparts, a positive association is observed between nTAS magnitude and the degree of lexical focus. This supports our design of nTAS for TA-vulnerable circuit mining.

5.2 Ablation and Variation Studies

We evaluate the stability of our method by repeating the object classification experiments with 4 random seeds and expansion ratios of $4, 8, 16, 32, 64$.

Consistency over Random States. To showcase the stability of our method, we repeat the object classification experiment with 4 random seeds and report the results in Tab. 5. Our method consistently extracts faithful lexical circuits and effectively promotes TA robustness across ViT variants and datasets, as evidenced by the low standard deviation of OCA and TCR.

Fig. 3: Concept Localization. Attribution map of selected random concept vectors with high text focus indicated by a high nTAS (Left), and with low text focus indicated by a low nTAS (Right). High attribution-score patches are indicated by a red gradient.

5.3 Influence of the Number of Samples.

As shown in Fig. 4, the standard deviation of the module score across random states decreases as the number of samples increases, which demonstrates that our training-free model explanation stabilizes with a large number of samples.

6 Conclusion

We tackle the challenge of interpreting hidden concept directions in ViT intermediate states using unlabeled images and efficiently extract causal circuits responsible for typographic-attack vulnerability. Our method treats concept discovery as a stochastic lottery within the lower-dimensional MHSA head subspace, grounded in the Linear Representation Hypothesis [12,26] and the Lottery Ticket Hypothesis [14], and uses a gradient-based attribution score to mine the lexical circuits. The proposed circuit mining pipeline requires only a diverse, unlabeled image dataset and minimal validation data for threshold hyperparameter search. It improves robustness to typographic attacks across model scales and datasets, surpassing prior supervised and training-free defenses. Extension of our method

Table 5: Robustness to random seed variations. We report the mean and standard deviation for object classification accuracy (OCA) % and text confusion rate (TCR) % at an expansion ratio of 16. The low standard deviations demonstrate the stability of our method.

Model	Metric	RTA-100	Disentangling	PAINT	IN-100-Text
ViT-B/16	OCA	68.7 ± 0.0	88.3 ± 0.0	73.8 ± 0.0	74.2 ± 0.0
	TCR	12.6 ± 0.0	11.7 ± 0.0	16.5 ± 0.0	7.1 ± 0.0
ViT-L/14	OCA	68.8 ± 0.1	68.3 ± 0.0	68.9 ± 0.0	74.7 ± 0.3
	TCR	21.3 ± 0.2	31.1 ± 0.0	22.3 ± 0.0	12.2 ± 0.3
ViT-H/14	OCA	76.2 ± 0.1	83.2 ± 1.9	77.2 ± 2.9	79.2 ± 0.1
	TCR	14.1 ± 0.6	16.3 ± 1.9	13.6 ± 1.9	9.2 ± 0.5
ViT-g/14	OCA	68.8 ± 0.0	83.2 ± 3.1	75.7 ± 0.0	76.4 ± 0.1
	TCR	23.4 ± 0.1	16.3 ± 3.1	18.2 ± 0.5	12.6 ± 0.3
ViT-bigG/14	OCA	75.4 ± 0.2	70.6 ± 2.6	75.0 ± 3.7	80.0 ± 0.4
	TCR	15.7 ± 0.6	28.9 ± 2.6	12.9 ± 2.0	9.6 ± 0.6

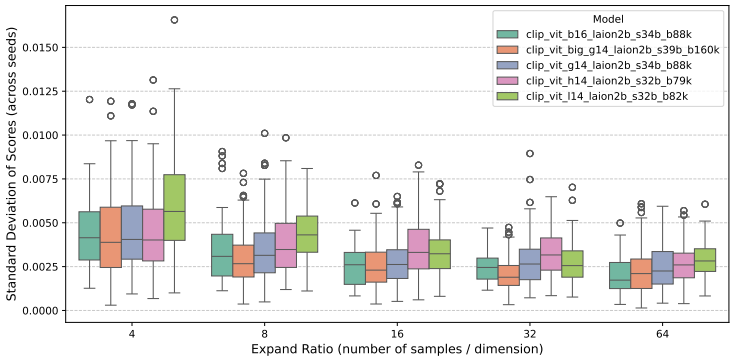

Fig. 4: Impact of the number of samples (parametrized by expansion ratio) on the method stability (in the standard deviation of the module score across random seeds). Our method stabilizes at 8 samples per dimension across ViT sizes and continues to stabilize as the number of samples increases.

to LVLM demonstrates the ability to improve the VQA accuracy of LVLMs under lexical distraction. Finally, the qualitative analysis of the neuron's activations verifies the causal relationship between the attention heads' focus on text and their correspondence in the lexical circuits. We hope this research will advance the frontier of mechanistic solutions to VLM robustness.

Acknowledgment

This work is supported in part by the US National Science Foundation (NSF) under grants IIS-2538206, IIS-2529378, IIS-2500341, CCF-2217071, CNS-2213700, and OAC-2530655. Any recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF.

References

1. Azuma, H., Matsui, Y.: Defense-prefix for preventing typographic attacks on CLIP. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3644–3653 (2023)
2. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Cao, Y., Xing, Y., Zhang, J., Lin, D., Zhang, T., Tsang, I., Liu, Y., Guo, Q.: Scenetap: Scene-coherent typographic adversarial planner against vision-language models in real-world environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 25050–25059 (June 2025)
5. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24185–24198. IEEE (2024)
6. Cheng, H., Xiao, E., Gu, J., Yang, L., Duan, J., Zhang, J., Cao, J., Xu, K., Xu, R.: Unveiling typographic deceptions: Insights of the typographic vulnerability in large vision-language models. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LIX. p. 179–196. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-73202-7_11, https://doi.org/10.1007/978-3-031-73202-7_11
7. Chung, N., Gao, S., Vu, T.A., Zhang, J., Liu, A., Lin, Y., Dong, J.S., Guo, Q.: Towards transferable attacks against vision-llms in autonomous driving with typography. arXiv preprint arXiv:2405.14169 (2024)
8. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=2dnO3LLiJ1
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=YicbFdNTTy
11. Dreyer, M., Hufe, L., Berend, J., Wiegand, T., Lapuschkin, S., Samek, W.: From what to how: Attributing CLIP's latent components reveals unexpected semantic reliance. arXiv preprint arXiv:2505.20229 (2025)
12. Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., Olah, C.: Toy models of superposition. Transformer Circuits Thread (2022), https://transformer-circuits.pub/2022/toy_model/index.html

13. Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., Olah, C.: A mathematical framework for transformer circuits. Transformer Circuits Thread (2021), https://transformer-circuits.pub/2021/framework/index.html
14. Frankle, J., Carbin, M.: The lottery ticket hypothesis: Finding sparse, trainable neural networks. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=rJl-b3RcF7
15. Gandelsman, Y., Efros, A.A., Steinhardt, J.: Interpreting CLIP's image representation via text-based decomposition. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=5Ca9sSzuDp
16. Gandelsman, Y., Efros, A.A., Steinhardt, J.: Interpreting the second-order effects of neurons in CLIP. arXiv preprint arXiv:2406.04341 (2024), https://arxiv.org/abs/2406.04341
17. Gemma Team, Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., Grill, J.B., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., Liu, G., Visin, F., Kenealy, K., Beyer, L., Zhai, X., Tsitsulin, A., Busa-Fekete, R., Feng, A., Sachdeva, N., Coleman, B., Gao, Y., Mustafa, B., Barr, I., Parisotto, E., Tian, D., Eyal, M., Cherry, C., Peter, J.T., Sinopalnikov, D., Bhupatiraju, S., Agarwal, R., Kazemi, M., Malkin, D., Kumar, R., Vilar, D., Brusilovsky, I., Luo, J., Steiner, A., Friesen, A., Sharma, A., Sharma, A., Gilady, A.M., Goedeckemeyer, A., Saade, A., Feng, A., Kolesnikov, A., Bendebury, A., Abdagic, A., Vadi, A., György, A., Pinto, A.S., Das, A., Bapna, A., Miech, A., Yang, A., Paterson, A., Shenoy, A., Chakrabarti, A., Piot, B., Wu, B., Shahriari, B., Petrini, B., Chen, C., Lan, C.L., Choquette-Choo, C.A., Carey, C.J., Brick, C., Deutsch, D., Eisenbud, D., Cattle, D., Cheng, D., Paparas, D., Sreepathihalli, D.S., Reid, D., Tran, D., Zelle, D., Noland, E., Huizenga, E., Kharitonov, E., Liu, F., Amirkhanyan, G., Cameron, G., Hashemi, H., Klimczak-Plucińska, H., Singh, H., Mehta, H., Lehri, H.T., Hazimeh, H., Ballantyne, I., Szpektor, I., Nardini, I., Pouget-Abadie, J., Chan, J., Stanton, J., Wieting, J., Lai, J., Orbay, J., Fernandez, J., Newlan, J., Ji, J.Y., Singh, J., Black, K., Yu, K., Hui, K., Vodrahalli, K., Greff, K., Qiu, L., Valentine, M., Coelho, M., Ritter, M., Hoffman, M., Watson, M., Chaturvedi, M., Moynihan, M., Ma, M., Babar, N., Noy, N., Byrd, N., Roy, N., Momchev, N., Chauhan, N., Sachdeva, N., Bunyan, O., Botarda, P., Caron, P., Rubenstein, P.K., Culliton, P., Schmid, P., Sessa, P.G., Xu, P., Stanczyk, P., Tafti, P., Shivanna, R., Wu, R., Pan, R., Rokni, R., Willoughby, R., Vallu, R., Mullins, R., Jerome, S., Smoot, S., Girgin, S., Iqbal, S., Reddy, S., Sheth, S., Põder, S., Bhatnagar, S., Panyam, S.R., Eiger, S., Zhang, S., Liu, T., Yacovone, T., Liechty, T., Kalra, U., Evci, U., Misra, V., Roseberry, V., Feinberg, V., Kolesnikov, V., Han, W., Kwon, W., Chen, X., Chow, Y., Zhu, Y., Wei, Z., Egyed, Z., Cotruta, V., Giang, M., Kirk, P., Rao, A., Black, K., Babar, N., Lo, J., Moreira, E., Martins, L.G., Sanseviero, O., Gonzalez, L., Gleicher, Z., Warkentin, T., Mirrokni, V., Senter, E., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Matias, Y., Sculley, D., Petrov, S., Fiedel, N., Shazeer, N., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Alayrac, J.B., Anil, R., Dmitry, Lepikhin, Borgeaud, S., Bachem, O., Joulin, A., Andreev, A., Hardin, C., Dadashi, R., Hussenot, L.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (Mar 2025)

18. Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., Radford, A., Olah, C.: Multimodal neurons in artificial neural networks. Distill **6**(3), e30 (2021)
19. Huben, R., Cunningham, H., Smith, L.R., Ewart, A., Sharkey, L.: Sparse autoencoders find highly interpretable features in language models. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=F76bwRSLeK
20. Hufe, L., Venhoff, C., Dreyer, M., Purelku, E., Lapuschkin, S., Samek, W.: Dyslexify: A mechanistic defense against typographic attacks in CLIP. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=UI7mbsIZeN
21. Ilharco, G., Wortsman, M., Gadre, S.Y., Song, S., Hajishirzi, H., Kornblith, S., Farhadi, A., Schmidt, L.: Patching open-vocabulary models by interpolating weights. Advances in Neural Information Processing Systems **35**, 29262–29277 (2022)
22. Kissane, C., Krzyzanowski, R., Bloom, J.I., Conmy, A., Nanda, N.: Interpreting attention layer outputs with sparse autoencoders. arXiv preprint arXiv:2406.17759 (2024), https://arxiv.org/abs/2406.17759
23. Li, C., Wang, S., Zhang, Y., Zhang, J., Zong, C.: Interpreting and exploiting functional specialization in multi-head attention under multi-task learning. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 16460–16476. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-main.1026, https://aclanthology.org/2023.emnlp-main.1026/
24. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
25. Materzyńska, J., Torralba, A., Bau, D.: Disentangling visual and written concepts in CLIP. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16410–16419 (June 2022)
26. Park, K., Choe, Y.J., Veitch, V.: The linear representation hypothesis and the geometry of large language models. In: Proceedings of the 41st International Conference on Machine Learning. ICML'24, JMLR.org (2024)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/radford21a.html
28. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: LAION-5B: An open large-scale dataset for training next generation image-text models. Advances in neural information processing systems **35**, 25278–25294 (2022)
29. Shekhar, A.: ImageNet100. Kaggle (2021), https://www.kaggle.com/datasets/ambityga/imagenet100, accessed: 2026-01-16
30. Su, J., Kempe, J., Ullrich, K.: From concepts to components: Concept-agnostic attention module discovery in transformers. arXiv preprint arXiv:2506.17052 (2025)
31. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI. p. 776–794. Springer-Verlag, Berlin, Heidelberg (2020).

https://doi.org/10.1007/978-3-030-58621-8_45, https://doi.org/10.1007/978-3-030-58621-8_45

32. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9568–9578 (2024)
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
34. Wang, G., Bai, L., Nah, W.J., Wang, J., Zhang, Z., Chen, Z., Wu, J., Islam, M., Liu, H., Ren, H.: Surgical-LVLM: Learning to adapt large vision-language model for grounded visual question answering in robotic surgery. In: ICLR 2025 Workshop on Foundation Models in the Wild (2025), https://openreview.net/forum?id=U4z69U9m9t
35. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
36. Wang, X., Zhao, Z., Larson, M.: Typographic attacks in a multi-image setting. In: Chiruzzo, L., Ritter, A., Wang, L. (eds.) Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 12594–12604. Association for Computational Linguistics, Albuquerque, New Mexico (Apr 2025). https://doi.org/10.18653/v1/2025.naacl-long.626, https://aclanthology.org/2025.naacl-long.626/
37. Waseda, F., Yamabe, S., Shiono, D., Sasaki, K., Takahashi, T.: Read or ignore? A unified benchmark for typographic-attack robustness and text recognition in vision-language models. CoRR **abs/2512.11899** (2025). https://doi.org/10.48550/ARXIV.2512.11899, https://doi.org/10.48550/arXiv.2512.11899
38. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-Image technical report. arXiv preprint arXiv:2508.02324 (2025)
39. Yang, Z., Jia, X., Li, H., Yan, J.: LLM4Drive: A survey of large language models for autonomous driving. arXiv preprint arXiv:2311.01043 (2023)
40. Zaigrajew, V., Baniecki, H., Biecek, P.: Interpreting CLIP with hierarchical sparse autoencoders. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=5MQQsenQBm

A Appendix

A.1 Dataset Details

Circuit Mining Dataset. Our training-free concept mining process takes a text-injected image dataset without labels. In our experiments, we use a uniformly sampled 0.1% subset of ImageNet [9] as the base dataset. The texts are randomly rendered in various colors and fonts at the image margins, spanning 20% of the image width, as shown in Fig. 5. The proposed method requires only the text-injected image and the text location, provided by the augmentation.

Fig. 5: Example of text-augmented images. Texts with various colors and fonts occur at each border with equal likelihood.

Evaluation Datasets. We evaluate the defense methods on the following datasets:

- **RTA-100** [1] consists of 100 categories and a 1000-item mixture of synthetic and real-world typographic attack images. The images are drawn from 10 regular image datasets, 2 typographic attack datasets, and additional images collected from the literature.
- **Disentangling** [25] consists of 19 categories and 171 images, constructed with daily objects and attack text written on sticky notes.
- **PAINT** [21] contains 110 images of daily objects and attack-text sticky notes.
- **IN-100-Text** is constructed from CMC's dataset [31], a 100-class subset of ImageNet [9]. We use the validation set of 5000 images to augment the data with Qwen-Image-Edit [38]. We consider seven common text types that often co-occur with objects across varying scenarios and are visualized in ways not covered by previous datasets: neon sign, comic chat balloon, street sign, billboard, engraved metal plate, sculpted 3D CGI text, and obvious watermark. The prompt for generating the text-injected images is shown in Fig. 6, and a visualized subset of all images in the generated dataset is given in Fig. 7
- **ImageNet-100** [29] is a subset of ImageNet consisting of 100 classes. We use it to evaluate defense trade-offs in terms of standard image classification accuracy relative to other methods.

User Prompt

<image> Add a {text_type} displaying the text "{text_word}" next to the {gt_text}.

Fig. 6: Prompt for Typographic Image Editing. text_type refers to one of the seven text types we defined. text_word is chosen randomly from the class label set, excluding the ground truth label of the input image. gt_text refers to the ground truth label of the input image.

A.2 Derivation of P_{success}

Bounding Interference via Concentration of Measure. We establish a high-probability bound on the maximum polysemantic interference across N sequence patches. Let the random probe $\mathbf{u}$ be drawn from an isotropic Gaussian scaled by the head dimension:

$$\mathbf{u} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{d_{\text{head}}}\mathbf{I}\right). \tag{15}$$

For patch i, let $\boldsymbol{\xi}_i \in \mathbb{R}^{d_{\text{head}}}$ be the deterministic interference vector orthogonal to $\mathbf{c}_{\text{target}}$. Its projection onto the probe is a zero-mean Gaussian:

$$I_{\mathbf{u},i} = \langle \boldsymbol{\xi}_i, \mathbf{u} \rangle \sim \mathcal{N}\left(0, \frac{\|\boldsymbol{\xi}_i\|^2}{d_{\text{head}}}\right). \tag{16}$$

For a Gaussian $Z \sim \mathcal{N}(0, \sigma^2)$, the sub-Gaussian tail bound gives $\mathbb{P}(|Z| \geq t) \leq 2\exp(-t^2/2\sigma^2)$. Applied to the interference projection and combined with the union bound over all N patches, with $\|\boldsymbol{\xi}_{\max}\| = \max_i \|\boldsymbol{\xi}_i\|$:

$$\mathbb{P}\left(\max_{i\in\{1\ldots N\}} |I_{\mathbf{u},i}| \geq t\right) \leq 2N \exp\left(-\frac{t^2 d_{\text{head}}}{2\|\boldsymbol{\xi}_{\max}\|^2}\right). \tag{17}$$

We set the threshold to $t = \|\boldsymbol{\xi}_{\max}\|\sqrt{\frac{2(1+\delta)\log N}{d_{\text{head}}}}$ for a slack parameter $\delta > 0$. The exponent becomes $-(1+\delta)\log N$, so the exponential contributes $N^{-(1+\delta)}$ and the union-bound factor N is overpowered:

$$\mathbb{P}\left(\max_i |I_{\mathbf{u},i}| \geq \|\boldsymbol{\xi}_{\max}\|\sqrt{\frac{2(1+\delta)\log N}{d_{\text{head}}}}\right) \leq 2N \cdot N^{-(1+\delta)} = 2N^{-\delta}. \tag{18}$$

The failure probability $2N^{-\delta}$ decays polynomially in N for any $\delta > 0$, with δ trading bound tightness ($2N^{-\delta}$) against the separation threshold ($\propto \sqrt{1+\delta}$). We set $\delta = 1$ to balance the two, yielding a failure probability of $2/N$ and the high-probability bound

$$\max_{i\in\{1\ldots N\}} |I_{\mathbf{u},i}| \leq \|\boldsymbol{\xi}_{\max}\|\sqrt{\frac{4\log N}{d_{\text{head}}}}. \tag{19}$$

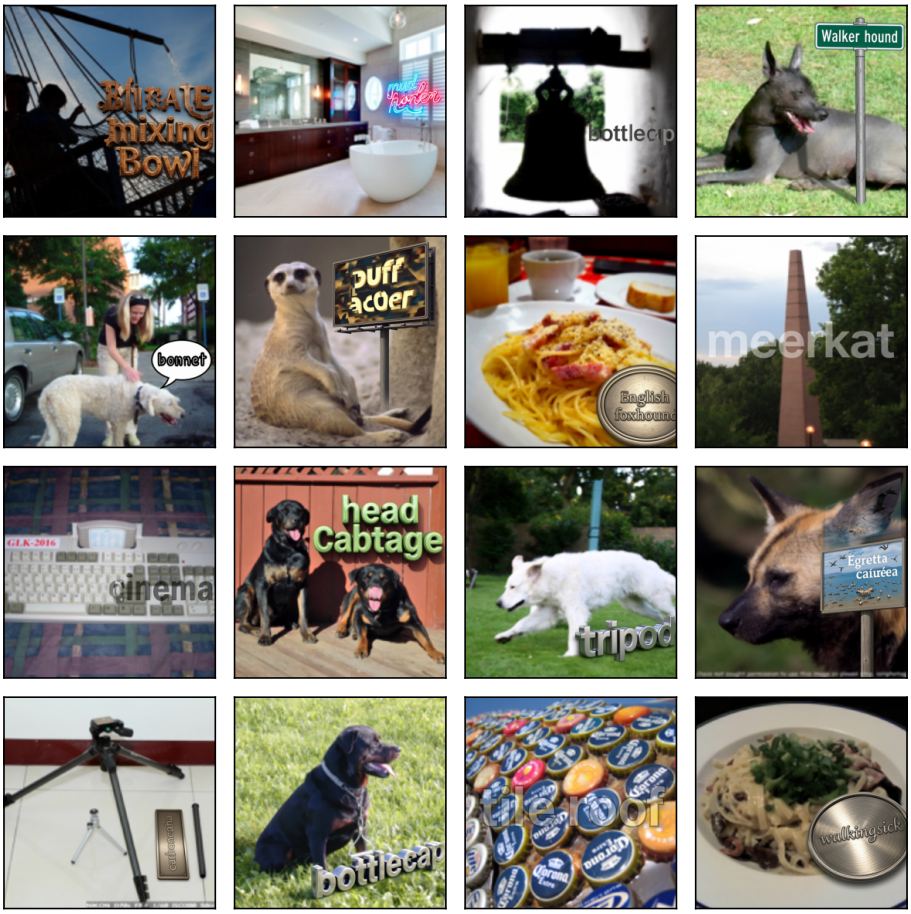

Fig. 7: Example data from IN-100-Text.

Derivation of the Single Success Probability (p). Next, we derive the probability p that a single random probe $\mathbf{u}$ resolves the weakest on-concept patch $k = \arg\min_{i \in \text{on}} \alpha_i$ above the background interference. Since every on-concept patch carries the concept at least as strongly as α_k, resolving patch k resolves all on-concept patches simultaneously. The alignment of the probe with the concept direction is $A_{\mathbf{u}} = \langle \mathbf{c}_{\text{target}}, \mathbf{u} \rangle$, with

$$A_{\mathbf{u}} \sim \mathcal{N}\left(0, \frac{1}{d_{\text{head}}}\right). \tag{20}$$

The signal of the weakest on-concept patch is $S_{\mathbf{u},k} = \alpha_k A_{\mathbf{u}}$, distributed as

$$S_{\mathbf{u},k} \sim \mathcal{N}\left(0, \frac{\alpha_k^2}{d_{\text{head}}}\right). \tag{21}$$

To utilize standard normal bounds, we scale $S_{\mathbf{u},k}$ to a standard normal variable $Z \sim \mathcal{N}(0,1)$:

$$Z = \frac{S_{\mathbf{u},k}\sqrt{d_{\text{head}}}}{\alpha_k} \sim \mathcal{N}(0,1). \tag{22}$$

For successful separation, the signal of patch k must exceed the sum of τ and the maximum interference bound derived in Section 3.4:

$$S_{\mathbf{u},k} > \|\boldsymbol{\xi}_{\max}\|\sqrt{\frac{4\log N}{d_{\text{head}}}} + \tau. \tag{23}$$

Multiplying both sides by $\sqrt{d_{\text{head}}}$ translates this condition into our standard normal variable Z:

$$Z > \frac{\|\boldsymbol{\xi}_{\max}\|\sqrt{4\log N} + \tau\sqrt{d_{\text{head}}}}{\alpha_k}. \tag{24}$$

Using the Gaussian tail approximation $\mathbb{P}(Z > z) \approx \exp(-z^2/2)$ for large z, we obtain the estimated probability p of drawing a successful separation vector:

$$p \approx \exp\left(-\frac{\left(\|\boldsymbol{\xi}_{\max}\|\sqrt{4\log N} + \tau\sqrt{d_{\text{head}}}\right)^2}{2\alpha_k^2}\right) \tag{25}$$

Derivation of the Required Sample Size (K). Finally, we calculate the number of random vectors K required in our Stochastic Lottery to guarantee finding at least one clean feature map with a desired confidence level, P_{success}.

Given that each of the K probes is drawn independently, the probability that a single probe fails to isolate the concept is $(1-p)$. The probability that **all** K probes fail is the product of their individual failure probabilities:

$$\mathbb{P}(\text{all fail}) = (1-p)^K \tag{26}$$

The probability of obtaining at least one successful probe is the complement of total failure. We require this to be greater than or equal to P_{success}:

$$1 - (1-p)^K \geq P_{\text{success}} \tag{27}$$

We rearrange this inequality to solve for K:

$$(1-p)^K \leq 1 - P_{\text{success}} \tag{28}$$

Taking the natural logarithm of both sides yields:

$$K\log(1-p) \leq \log(1 - P_{\text{success}}) \tag{29}$$

Because p is a probability between 0 and 1, the term $(1-p)$ is less than 1, making its logarithm strictly negative. Dividing both sides by a negative number flips the inequality sign, resulting in the final lower bound for K:

$$K \geq \frac{\log(1 - P_{\text{success}})}{\log(1-p)} \tag{30}$$

This completes the derivation, justifying the benefit of dimension reduction from operating within the d_{head} subspace.

Table 6: Performance Tradeoff. We present the performance trade-off of the compared methods in terms of classification accuracy on the ImageNet-100 validation set.

Method	ViT-B/16	ViT-L/14	ViT-H/14	ViT-g/14	ViT-bigG/14	Average
Vanilla ViT	74.8	80.0	83.8	83.8	85.3	81.5
Defense-Prefix	75.4 (+**0.6**)	79.8 (**-0.2**)	83.4 (**-0.4**)	83.2 (-0.6)	85.4 (+**0.1**)	81.4 (**-0.14**)
Dyslexify	75.3 (+0.5)	79.5 (-0.5)	83.4 (**-0.4**)	83.0 (-0.8)	85.0 (-0.3)	81.3 (-0.24)
Dyslexify*	75.0 (+0.2)	79.5 (-0.5)	83.4 (**-0.4**)	82.6 (-1.2)	84.7 (-0.6)	81.0 (-0.54)
Ours (nTAS)	74.2 (-0.6)	79.8 (**-0.2**)	82.9 (-0.9)	83.3 (**-0.5**)	84.6 (-0.7)	81.0 (-0.54)

A.3 nTAS Distribution over ViT Model Depth

In Fig. 8, we show the per-attention-module nTAS across the last half of the ViT-B/16 and the ViT-H/14 models. Our method can produce distinct lexical-focused attention modules across ViT layers, as highlighted in Fig. 8. The comparison between ViT-B/16 and ViT-H/14 reveals a potential discrepancy in model-level internal behavior related to model size: ViT-B/16 has more evenly distributed lexical attention modules, whereas ViT-H/14 exhibits a "dark zone" from layer 22 to layer 26. Such an observation sparks further mechanistic interpretability research into shifts in ViT model behavior with respect to model size.

A.4 More Results on General Performance Trade-off

As shown in Tab. 6, our method consistently yields competitive general classification results compared with the methods we compare against. The supervised optimization-based Defense-Prefix tuning yields the least trade-off in overall performance, as the optimization process preserves model behavior on normal inputs. Notably, our method achieves comparable or better-than-supervised general classification accuracy across two of the five ViT variants tested.

Fig. 9 shows the robustness-capacity tradeoff by plotting the typographic-attack dataset accuracies and the ImageNet-100 dataset accuracies. Our method consistently yields the best results among large ViT models, benefiting from the model-agnostic nature of the top-down attribution-based circuit mining process.

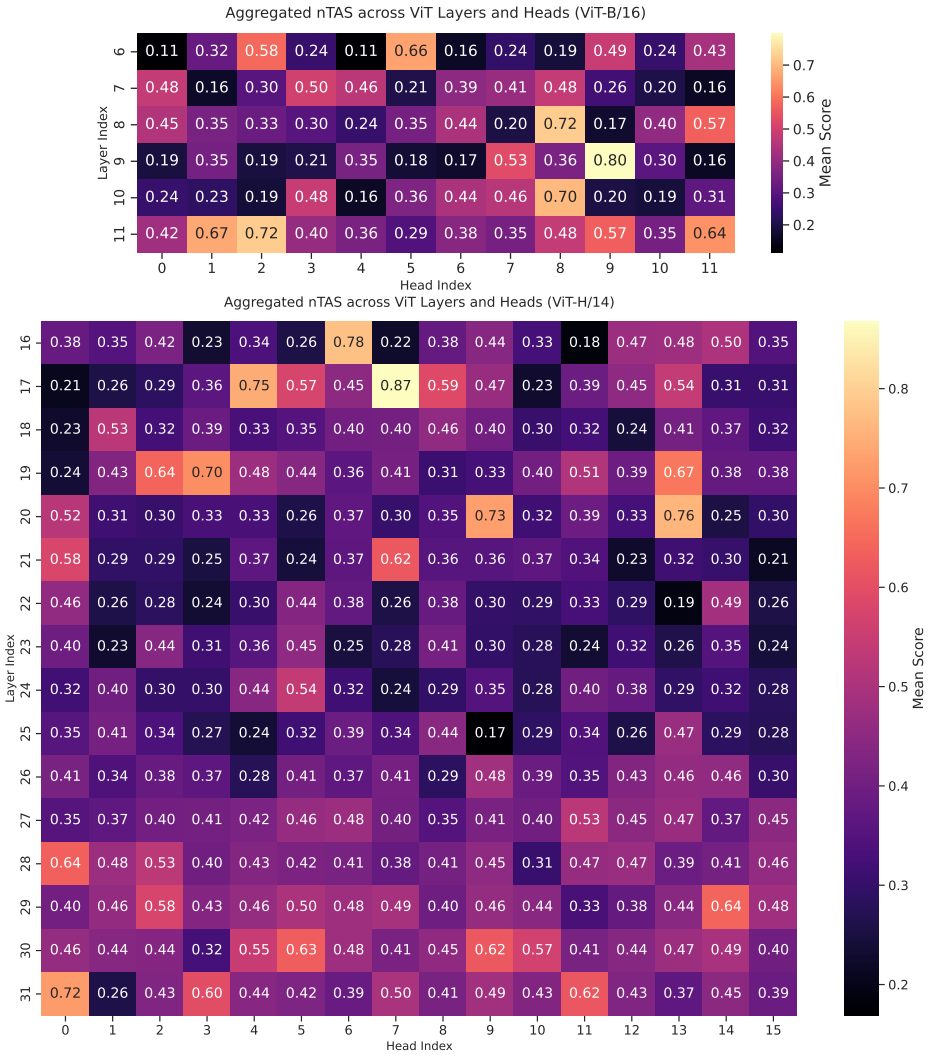

Fig. 8: Mean nTAS of Attention Heads across ViT layers.

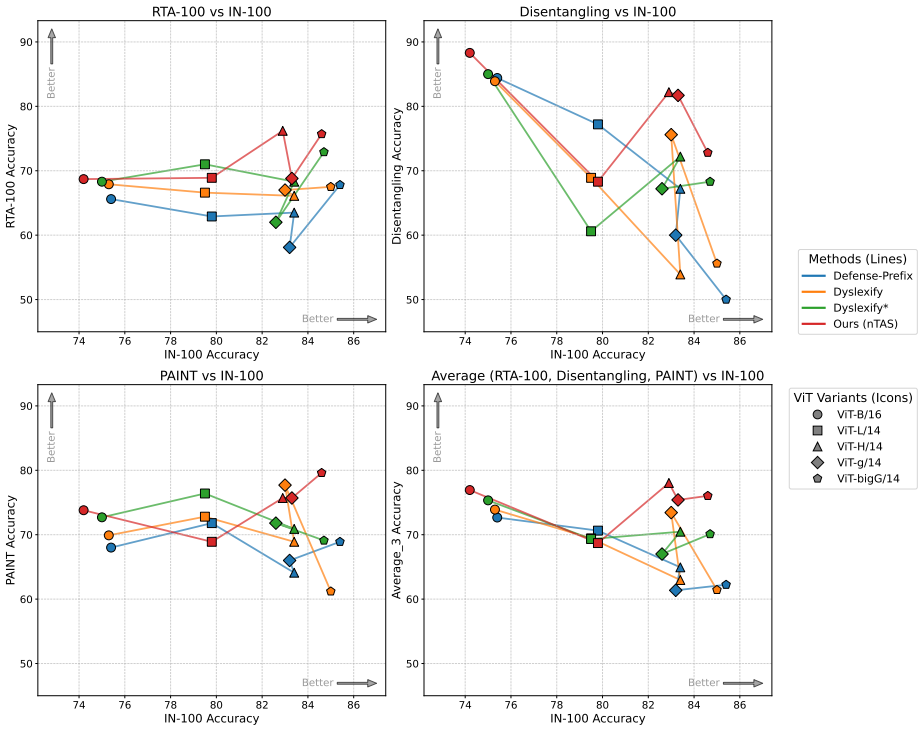

Fig. 9: Classification Accuracy on Typographic-Attack Datasets and ImageNet-100 Dataset. The joint plot shows the overall robustness trade-off of each tested method. Our method consistently produces strong robustness with minimal trade-off, especially in large and complex ViT variants.