Towards Mechanistic Defenses Against Typographic Attacks in CLIP

Anonymous Author(s)
Affiliation
Address
email

Abstract

Typographic attacks exploit multi-modal systems by injecting text into images, leading to targeted misclassifications, malicious content generation and even Vision-Language Model jailbreaks. In this work, we analyze how CLIP vision encoders behave under typographic attacks, locating specialized attention heads in the latter half of the model's layers that causally extract and transmit typographic information to the cls token. Building on these insights, we introduce a method to defend CLIP models against typographic attacks by selectively ablating a typographic circuit, consisting of attention heads. Without requiring finetuning, our method improves performance by up to 19.6% on a typographic variant of ImageNet-100, while reducing standard ImageNet-100 accuracy by less than 1%. Notably, our training-free approach remains competitive with current state-of-the-art typographic defenses that rely on finetuning. To this end, we release a family of dyslexic CLIP models which are significantly more robust against typographic attacks. These models serve as suitable drop-in replacements for a broad range of safety-critical applications, where the risks of text-based manipulation outweigh the utility of text recognition.

1 Introduction

CLIP models are increasingly adopted as general-purpose vision–language representations, enabling applications in zero-shot classification, retrieval, diffusion-based generative models, and large-scale vision–language models (VLMs). Their versatility has further driven adoption in safety-relevant domains such as healthcare [1, 2, 3], remote sensing [4, 5, 6], and content moderation [7, 8, 9]. However, despite their widespread use, CLIP models remain vulnerable to typographic attacks: inserting text into an image can mislead classification, trigger malicious generations, or even jailbreak multi-modal systems (see Fig. 1).

Existing defenses against typographic attack require gradient-based optimization. While effective to some extent, these methods require substantial computational resources and lack interpretability into the mechanisms underlying CLIP's behaivor.

In this work, we propose a training-free defense that directly targets the circuits responsible for typographic vulnerability. By identifying and ablating a set of causal attention heads, we construct *dyslexic* CLIP models that are substantially more robust to typographic attacks. Our method scales seamlessly to billion-parameter models, making it applicable to state-of-the-art multi-modal systems. Beyond improving robustness this approach also enhances interpretability of CLIP models, enabling targeted intervention that are computationally efficient and easily integrated into existing pipelines without additional training overhead.

The contributions of this work are:

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

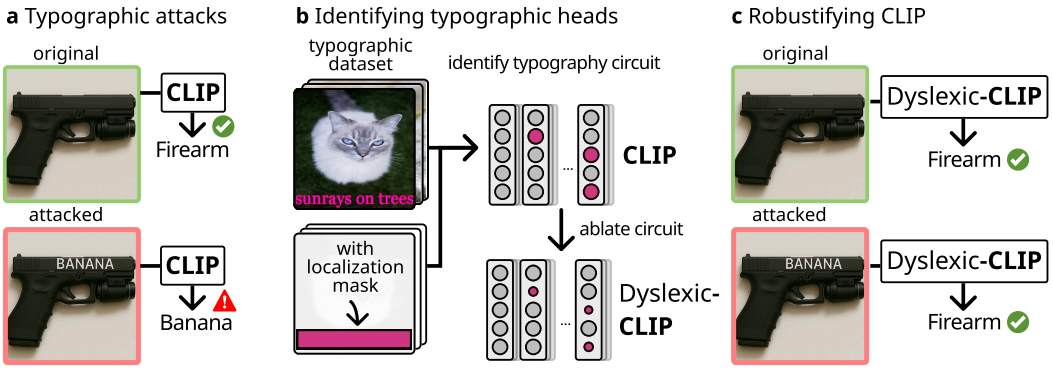

Figure 1: Defending CLIP against typographic attacks a) Adversarial text in images can dominate CLIP's representation and lead to misclassification. b) We construct a circuit of attention heads responsible for transmitting typographic information. c) By suppresses the typographic circuit, we defend against typographic attacks without a single gradient step.

- **Mechanistic Understanding:** We present the Typograhic Attention Score to locate specialized typographic attention heads and demonstrate their causal role in typographic attacks within CLIP models through controlled interventions.
- **Training-Free Defense:** We introduce a method that utilizes circuit ablation to effectively defend against typographic attacks, while maintaining general visual capabilities. Due to its training-free nature our method seamlessly scales to billion-parameter models on consumer grade hardware.
- **Empirical Validation:** We validate our method across a diverse set of zero-shot classification tasks, demonstrating that our approach improves robustness to typographic attacks by up to 19.6% on a typographic version of Imagenet-100 while maintaining high accuracy on non-typographic benchmarks.
- **Model Release:** To facilitate safer deployment, we release a family of dislexic CLIP models with reduced typographic sensitivity, suitable for use in safety-critical applications.

Our approach provides a practical, interpretable, and computationally efficient typographic defense, paving the way for safer multimodal systems without the need for fine-tuning or model retraining.

2 Related work

CLIP models [10] are pretrained on large-scale image–text datasets such as Laion-5b [7], aligning global image features with textual descriptions for strong zero-shot transfer. This reliance on textual supervision also makes them vulnerable to typographic attacks.

Typographic attacks [11] insert written text into an image to maliciously alter a model's behavior. Recent work demonstrates that typographic attacks can degrade model performance, bypass safety filters (jailbreaking) and hijack goals in Vision Language Models (VLMs) [12, 13, 14, 15, 16], trigger harmful content generation in image-to-image pipelines [17], and cause targeted misclassification in object detection and zero-shot classification settings [18, 19, 20, 16, 21].

While several defenses have been proposed, they rely either on fine-tuning the model [19], learning a projection matrix [18], incorporating a learnable text-token called Defense-Prefix [20], or employing Sparse Autoencoders [22]. Crucially, none of these approaches offer a training-free method for mitigating typographic attacks. In contrast, our work introduces a controllable intervention at inference time by directly locating and suppressing the components responsible for typographic sensitivity, without requiring additional training or fine-tuning. This makes our approach both interpretable and applicable, enabling safer deployment in scenarios prone to typographic attacks.

3 Datasets

To present our method we conduct experiments on two datasets: ImageNet-100-Typo and Unsplash-Typo. We use Unsplash-Typo to calculate the Typographic Attention Score, and we utilize ImageNet-100-Typo during circuit construction.

Unsplash-Typo: The Unsplash-Typo dataset consists of 10,000 natural images from Unsplash [23], originally containing minimal typographic content. We synthetically introduce typographic content at the bottom center of the image. The fixed text position enables efficient attention patterns analysis.

ImageNet-100-Typo: ImageNet-100 [24] is a commonly used subset of ImageNet [25], consisting of 100 classes. We construct Imagenet-100-Typo by assigning each image x with label y_{image} an additional random class label y_{typo}. Then we overlay the text description of y_{typo} on a random location int the image x, following Azuma et al. [20].

Furthermore we utilize the datasets RTA-100 [20], Disentangling [18] and PAINT-DS [19] of real world typographic attacks to benchmark the effectiveness of our typographic defense. Additionally we utilize the Aircraft [26] and Food-101 [27] datasets to benchmark the zero-shot image classification performance.

4 Locating layers of typographic understanding

Typographic attacks have been shown to be an effective attack vector on CLIP models. To defend against those attacks we investigate which layers and which components are responsible for typographic understanding, using linear probes.

We train linear probes on the cls token embedding at each layer of the OpenCLIP models of scales ViT-B to ViT-Big-G [28, 29], using the ImageNet-100-typo dataset.

More formally, we train a set of linear probes P_ℓ for each layer ℓ, where the probe predicts:

$$\hat{y}_\ell(x) = w^\top h_\ell(x) + b \tag{1}$$

Here, $h_\ell(x) \in \mathbb{R}^d$ denotes the activation of the model at layer ℓ for an input sample x, w a weight vector, and b a bias term. We train two types of probes: $P_{\text{image},\ell}$, which predicts the object label y_{image} and $P_{\text{typo},\ell}$, which predicts the typographic label y_{typo}. We write $\text{Acc}(P)$ to note the accuracy of a probe.

Fig. 2b shows that $\text{Acc}(P_{typo,\ell}) > 0.99$ accuracy for at the final layer for all tested models, indicating that CLIP models have significant typographic understanding. Furthermore it shows, that $\text{Acc}(P_{\text{typo},\ell})$ is low in early layers, but exhibit a sharp increase in the latter half of the model.

Fig. 2c shows that $\text{Acc}(P_{image,\ell})$ improves gradually over the layers, which is not the case for $\text{Acc}(P_{typo,\ell})$. The object probes show a gradual increase in performance throughout the model depth, while the typographic probes display a sharp performance rise around second half of the models layers.

Fig. 2d highlights the effect of the attention and the MLP blocks onto $\text{Acc}(P_{image,\ell})$ and $\text{Acc}(P_{typo,\ell})$. Attention layers consistently improve accuracy indicating that they add linearly decodable information to the cls token. In contrast, the MLP layers tend to reduce accuracy. We show in App. A, that the Intrinsic Dimensionality (ID) of the signal decreases after the MLP blocks, suggesting that the MLP compresses or discards information.

5 Method

To defend CLIP against typographic attacks we present a framework for detecting typographic circuits. Based on the results in Sec. 4 we only consider attention heads for the circuit construction. We define a circuit as a subset $\mathcal{C} \subseteq \Psi$, where Ψ denotes the set of attention heads in CLIP

$$\Psi = \{\mathcal{H}_{i,\ell} \mid i \in \{0, \ldots, I\},\ \ell \in \{0, \ldots, L\}\} \tag{2}$$

with I heads per layer and L layers in total.

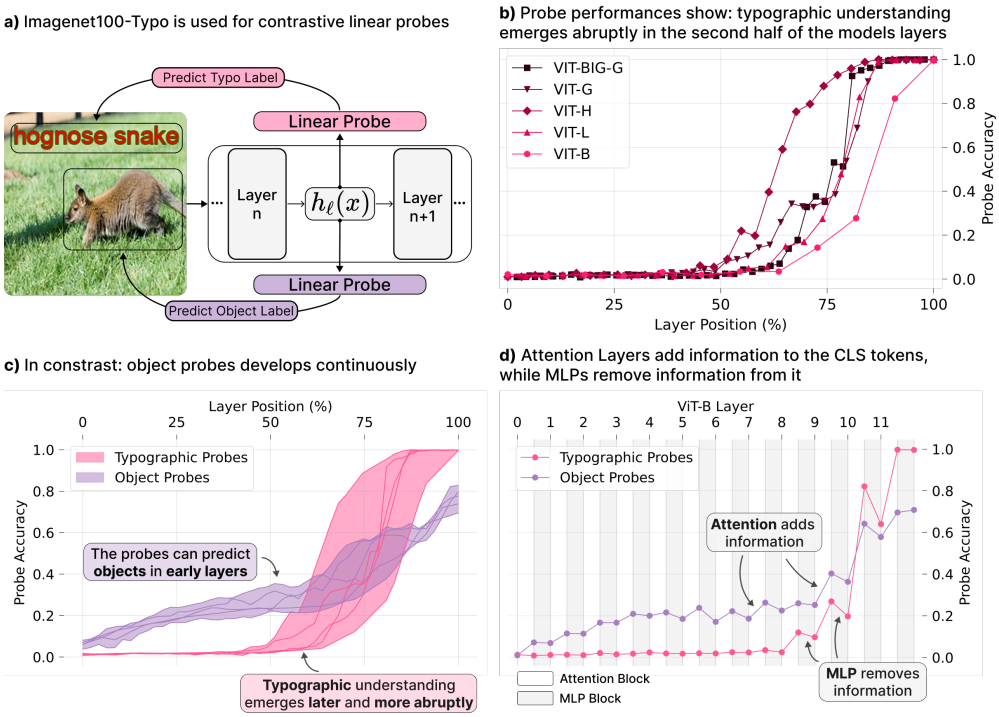

Figure 2: **a)** We train two linear probes on all layers of each CLIP model. $P_{\text{image},\ell}$ is used to predict the text label of each sample while $P_{\text{typo},\ell}$ is trained to predict the typographic class. **b)** $P_{\text{typo},\ell}$ shows a consistent pattern across all model sizes: typographic information emerges abruptly in the second half of the models layers. **c)** This trend is not true for the object probes $P_{\text{image},\ell}$. Object specific information builds gradually over the layers. Each line in the shaded area represents one CLIP model. **d)** While attention layers add linearly decodable information to the cls token, MLP layers removes linearly decodable information.

To defend a model $\mathcal{M}$ against typographic attacks, we conduct *circuit-ablation* of a typographic circuit $\mathcal{C}$. Circuit-ablation modifies only the residual stream of the cls token, leaving all other computations intact. Specifically, the residual update of the cls token is given

$$h^{\ell}_{\text{cls}} = h^{\ell-1}_{\text{cls}} + \sum_{i=1}^{H} \mathcal{H}_{i,\ell}(x)_{\text{cls}}, \tag{3}$$

where $\mathcal{H}_{i,\ell}(x)_{\text{cls}}$ is the contribution of head $\mathcal{H}_{i,\ell}$ to the cls token. For a circuit $\mathcal{C}$, we define circuit-ablation as

$$\mathcal{H}_{i,\ell}(x)_{\text{cls}} \mapsto 0 \quad \text{for all } \mathcal{H}_{i,\ell} \in \mathcal{C}, \tag{4}$$

while leaving all spatial contributions unchanged. We use $\mathcal{M}_{\mathcal{C}}$ to denote a model $\mathcal{M}$ in which $\mathcal{C}$ is circuit-ablated.

5.1 Typographic attention score

Building on Hung et al. [30] we introduce the Typographic Attention Score $T_{i,\ell}$ to guide the circuit construction.

Given a head $\mathcal{H}_{i,\ell}$ and an input x, we write $A_{i,\ell}(x) \in [0,1]^{T+1}$ to denote the attention pattern of the cls token, where T is the number of spatial tokens and the additional entry corresponds to the cls token. We further define the spatial cls-attention pattern $A^*_{i,\ell}(x) \in [0,1]^{T}$ which excludes the cls-to-cls entry, such that

$$A_{i,\ell}(x) = \big(A_{i,\ell}(x)_{\text{cls}},\ A^*_{i,\ell}(x)\big). \tag{5}$$

Formally the score is given by:

$$T_{i,\ell} = \sum_{x \in \mathcal{D}} \sum_{t \in \{1,\dots,T\}} \frac{\mathbb{1}(t)\, A^*_{i,\ell}(x)_t}{A^*_{i,\ell}(x)_t} \tag{6}$$

Where x is a data point, $\mathcal{D}$ is the dataset, and $\mathbb{1}$ is an indicator function so that $\mathbb{1}(t) = 1$ if the token at index t contains typographic content and is zero otherwise.

5.2 Typographic Circuit Construction

To defend against typographic attacks without degrading zero-shot classification, we construct a typographic circuit $\mathcal{C}$. The circuit is built iteratively while monitoring the accuracy of the circuit-ablated model $\mathcal{M}_{\mathcal{C}}$ on an image classification benchmark D_{image}.

Our procedure follows two steps: (i) rank all attention heads $\mathcal{H}_{i,\ell}$ by their typographic score $T_{i,\ell}$; (ii) add heads to $\mathcal{C}$ in descending order of $T_{i,\ell}$, evaluating accuracy after each addition. Let $\text{Acc}(\mathcal{M})$ denote accuracy on D_{image}. For each candidate head, we compute

$$\Delta\text{Acc} = \text{Acc}(\mathcal{M}) - \text{Acc}(\mathcal{M}_{\mathcal{C}}). \tag{7}$$

If $\Delta\text{Acc} < \epsilon$, the head is retained in $\mathcal{C}$; otherwise the head is discarded and the algorithm terminates.

We refer to the resulting model $\mathcal{M}_{\mathcal{C}}$ with the final circuit $\mathcal{C}$ as the *dyslexic* model.

Algorithm 1 Typographic Circuit Construction

```
Initialize circuit C ← ∅.
Rank all heads H_{i,ℓ} by score T_{i,ℓ}.
for head H_{i,ℓ} in descending order of T_{i,ℓ} do
    Tentatively add H_{i,ℓ} to C.
    Compute ΔAcc = Acc(M) − Acc(M_C).
    if ΔAcc ≥ ε then
        Remove H_{i,ℓ} and break.
    end if
end for
Return final circuit C.
```

6 Experiments

6.1 Evaluating the typographic attention score

For each attention head $\mathcal{H}_{i,\ell}$ we extract the attention pattern $A^*_{i,\ell}(x)$ over the Unsplash-Typo dataset. We use the known spatial bias of Unsplash-Typo to define the indicator mask $\mathbb{1}(t) \in \{0, 1\}$, where $\mathbb{1}(t) = 1$ at the the typography region in the bottom of the image and $\mathbb{1}(t) = 0$ elsewhere. Given $\mathcal{H}^{\text{cls}}_{i,\ell}$ and $\mathbb{1}$, we compute the typographic attention score $T_{i,\ell} \in [0, 1]$

Fig. 3 shows ViT-B's $T_{i,\ell}$ on UnsplashTypo. The empirical mean $\bar{T} = (26.9 \pm 14.6)\% = \frac{1}{I \cdot L} \sum_{I,L} T_{i,\ell}$ is close to the expected value under uniform attention, $\mathbb{E}[T_{i,\ell}]_{A^* \text{ uniform}} \approx 0.214$. This indicates that most heads do not disproportionately attend to the typographic region of the image. However, a small subset of heads show high scores of up to $T_{i,\ell} \geq 3 \times \bar{T}$, revealing a strong spatial bias towards typography. Furthermore we observe, that the spike in $\text{Acc}(P_{typo,\ell})$ only occurs after layers with high $T_{i,\ell}$ heads.

6.2 Evaluating the typographic circuits

We construct the circuits $\mathcal{C}$, we choose $\epsilon = 0.01$ by and set D_{image} to be a 5% class balanced subset of the non typographic Imagenet-100 train split. Tab. 1 shows that the resulting circuits are sparse, covering a maximum of 10.2% of the total number of heads in Ψ. Fig. 4 show the accuracy curves of $\text{Acc}(\mathcal{M}_{\mathcal{C}})$ per added head over the a Imagenet-100 and its typographic version Imagenet-100-typo.

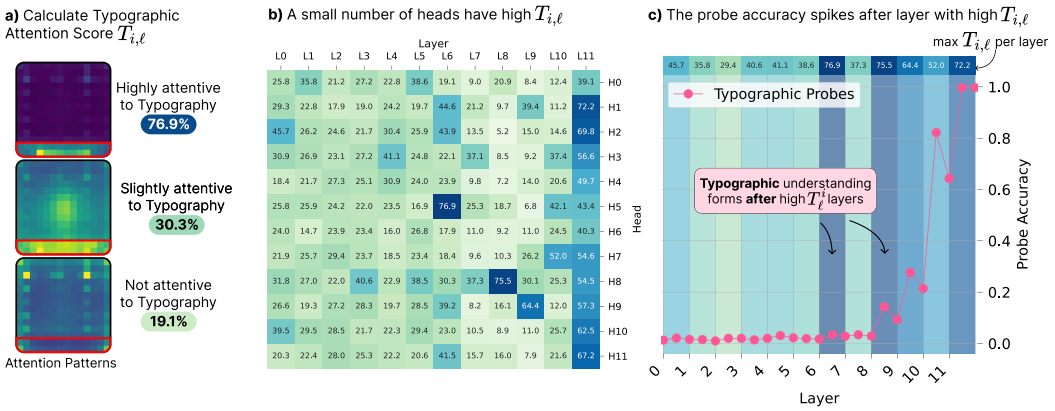

Figure 3: **a)** For each head in the model we calculate the Typographic Attention Score $T_{i,\ell}$, utilizing the spatial bias in the Unsplash-typo dataset. **b)** Depiction of ViT-B's $T_{i,\ell}$ scores. While most attention heads do not show any spatial bias in their attention patterns, a few attention heads indicate $T_{i,\ell}$ scores up to $\times 3$ higher than the mean score. Those heads only occur in the second half of the models layers **c)** Extracting the maximum $T_{i,\ell}$ score of the layer and overlaying these scores with the linear probes, highlights an interesting correlation. Only after the attention heads with exceptionally high $T_{i,\ell}$ scores were infered by the model the accuracy of $P_{\text{typo},\ell}$ begins to increase significantly.

Table 1: Number of heads $\mathcal{H}_{i,\ell}$ in $\mathcal{C}$ per Model

Model	Selected	Total	Percentage (%)
B	6	144	4.2
L	13	288	4.5
H	30	384	7.8
G	27	480	5.6
Big-G	59	576	10.2

6.2.1 Demonstrating causality of typographic circuits

We observed that attention heads in $\mathcal{C}$, utilized their CLS-to-CLS attention as attention sinks [31], which are utilized strongly when no typographic content is present in the image. More information on this can be found in App. B.

To demonstrate causal dependencies in the model's typographic vulnerability, we manipulate attention patterns of the circuit $\mathcal{C}$. Specifically we construct $A^{\alpha}_{i,\ell} \in [0,1]^{w \times h+1}$ as follows

$$A^{\alpha}_{i,\ell} = [\alpha, A^{*}_{i,\ell} \cdot (1-\alpha)/\|A^{*}_{i,\ell}\|] \tag{8}$$

where choose $\alpha \in \{0.0, 0.1, \ldots, 0.9, 1.0\}$, and ensure that the attention pattern sums to 1 via the scaling factor $(1-\alpha)/\|A^{*}_{i,\ell}\|$.

We measure the effectiveness of the typogrpahic attacks, when manipulating the attention pattern of the circuit, by tracing the probabilities labels $p(y_{\text{text}})$ and $p(y_{\text{typo}})$ depending on α on the Imagenet-100-typo dataset.

Results: Fig. 5 shows that increasing α causally lowers the typographic attacks effectiveness, while decreasing α increases the effectiveness. This demonstrates the causal link: as the circuit heads attend stronger to their attention-sink, the typographic attack becomes less effective. Conversely, when α is low and the norm of the spatial attention is high $p(y_{typo})$ raises, indicating that typographic information is being transferred from the spatial to the cls token.

6.3 Evaluating the dislexic models

To evaluate our method we construct dyslexic OpenClip model variants: ViT-B, L, H, G, and BigG and conduct zero-shot classification experiments. Concretely we first evaluate the effectiveness of our defense against typographic attacks and secondly measure the zero-shot object classification

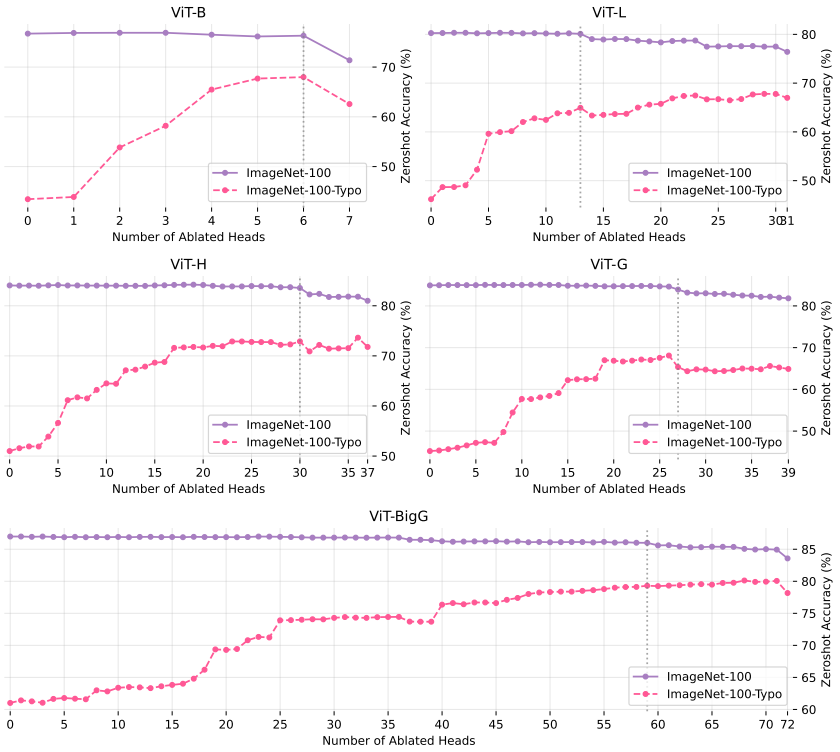

Figure 4: Tradeoff between general accuracy and typographic robustness as a function of the number of ablated heads. Ablations are applied in decreasing order of $T_{i,\ell}$. The dotted vertical line indicates the point at which further head removal is halted, based on a maximum allowed ΔAcc of $\epsilon = 0.01$.

capabilities on non-typographic datasets. For each model, we record the accuracy difference between the original model and dyslexic model.

Results: Tab. 2 shows that our method consistently improves typographic robustness across a diverse set of real world typographic attack datasets, including RTA-100, Disentangling, PAINT as well as our synthetic ImageNet-100-Typo dataset. The results in Tab. 2 further demonstrate that our method generalizes effectively across a wide range of model sizes, indicating that the existence of specialized typographic circuits in large models.

Tab. 3 confirms that our method does not significantly degrade performance on standard vision datasets. In most cases, the change in zero-shot accuracy is within $\pm 1\%$, with ViT-B even showing a slight improvement ($+1.59\%$) on the Aircraft dataset. The strongest performance decline shows ViT-G with -1.06% on ImageNet-100, which is close to the maximally allowed drop of $\epsilon = 1\%$.

Table 2: Comparison of dyslexic models performance on datasets of typographic attacks across model sizes, showing accuracy changes relative to the base model, with improvements (↑) or declines (↓).

Model	RTA-100 [20]	Disentangling[18]	PAINT-DS[19]	ImageNet-100-Typo
B	67.70%↑8.50%	88.33%↑32.78%	65.45%↑6.36%	66.50%↑19.60%
L	61.60%↑6.10%	65.00%↑13.89%	69.09%↑9.09%	65.54%↑13.14%
H	67.40%↑11.60%	73.33%↑27.22%	70.91%↑20.00%	72.26%↑18.84%
G	62.30%↑10.00%	57.22%↑5.56%	71.82%↑16.36%	64.64%↑17.28%
Big-G	72.60%↑11.50%	68.33%↑20.00%	71.82%↑21.82%	77.36%↑15.52%

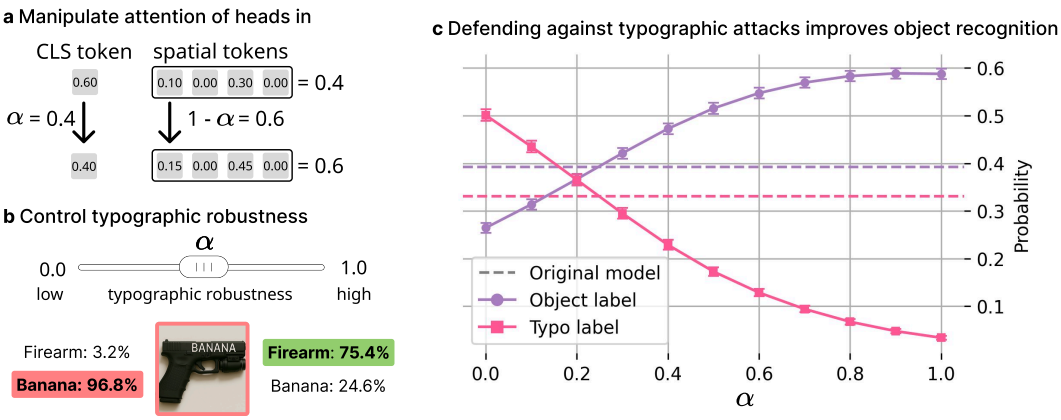

Figure 5: Controlling typographic understanding by manipulating attention patterns in typographic attention heads. a) We set the cls tokent attention to a specified value of $1 - \alpha$, and rescale spatial token attentions to sum up to α. b) By increasing (or decreasing) α, we can increase (or decrease) the amount of typographic understanding in images. c) Reducing typographic understanding causes a probability increase for predicting the true object class in images.

Table 3: Comparison of dyslexic model performance on non-typographic datasets across model sizes, showing accuracy changes relative to the base model, with improvements (↑) or declines (↓).

Model	Aircraft[26]	Food-101[27]	ImageNet-100 [25]
B	27.48%↑1.59%	86.23%=0.00%	75.26%↑0.50%
L	34.38%↑0.15%	90.67%↑0.06%	79.60%↓0.34%
H	40.35%↓0.15%	92.32%↓0.42%	83.00%↓0.80%
G	44.22%↑0.18%	91.81%↓0.36%	82.70%↓1.06%
Big-G	47.88%↑0.45%	92.96%↓0.08%	84.90%↓0.48%

6.4 Comparing to baselines

We compare our training-free method to Defense-Prefix (DP) [20], which introduces a learnable prefix token for CLIP's language transformer on OpenCLIP ViT-L without fine-tuning the full ViT. Following their setup, we train on the ImageNet-100 training split for 6 epochs with a learning rate of 0.002, batch size of 64, and hyperparameters $\gamma = 3.0$ and $\eta = 1.0$.

Tab. 4 shows that DP achieves stronger performance on most typographic benchmarks. However, our method consistently outperforms DP on non-typographic benchmarks, with the exception of ImageNet-100. This highlights a key advantage of our approach. Since DP is optimized on ImageNet-100-Typo, it yields a modest improvement on the corresponding non-typographic ImageNet-100 set, but at the cost of reduced performance on other non-typographic datasets. We hypothesize that the black-box optimization in DP captures features relevant not only to typographic defense but also to ImageNet-100 classification, thereby limiting its generalization. In contrast, our method focuses on causal mechanisms of typographic vulnerability, leading to more robust transfer across non-typographic benchmarks.

7 Discussion

Limitations and future work: Our methods enhances the typographic robustness of the cls token by preventing specialized typographic attention heads from writing to it. However, many multimodal applications, such as LLaVA and IP adapters [32, 33, 34], leverage not only the cls token but also spatial tokens, allowing typographic information to propagate into downstream tasks. This might limit the impact of our method on improving robustness in applications, and calls for further investigation into its generalizability to VLM setups.

Table 4: Performance comparison of our method and Defense-Prefix (DP) on ViT-L across typographic and non-typographic datasets. For each method we show the accuracy followed by the derivation from the baseline, (↑) for improvement, (↓) for decline.

	Typographic			Training	Non-Typographic	
Method	RTA-100	Disentangling	PAINT	ImageNet-100	Food-101	Aircraft
Baseline	55.50%	51.11%	60.00%	79.94%	90.61%	34.11%
DP	62.20%↑ 6.70%	82.78%↑31.67%	71.82%↑11.82%	81.70%↑ 1.76%	89.83%↓ 0.78%	32.94%↓ 1.17%
Ours	61.60%↑ 6.10%	65.00%↑13.89%	69.09%↑ 9.09%	79.60%↓ 0.34%	90.67%↑ 0.05%	34.38%↑ 0.27%

Misuse Potential: While we aim to enhance the safety of multimodal systems, we acknowledge that our insights into CLIP's behaivor under typographic attacks could be exploited by attackers. Specifically, adversarial inputs might be crafted to increase the spatial attention of heads in $\mathcal{C}$, making typographic attacks even more effective.

8 Conclusion

We present a mechanistic defense against typographic attack in CLIP using an interpretability first approach. We reveal that a small number of attention heads located in the later layers of the vision encoder are casual for the effectiveness of typographic attacks. By selectively ablating a typographic circuit, we are able defend CLIP against typographic attacks without requiring fine-tuning of any kind, offering a practical and interpretable method for controlling model behavior. To our knowledge, this is the first work to address typographic vulnerabilities in CLIP through causal interventions. Our method, demonstrates that fine-grained control over model capabilities is achievable through targeted architectural manipulations *without* retraining, and thus paves the way for more robust and modular deployment of multi-modal models. We believe this work motivates a broader shift toward mechanistic interpretability as a tool not only for understanding, but for controlling safety relevant behaviors in deep transformer models. Finally we release a family of dyslexic CLIP models that are significantly more robust against typographic attacks. These models serve as drop-in replacements for safety-critical applications where the risks posed by adversarial text manipulation outweigh the benefits of typographic understanding.

References

[1] Yue Yang, Mona Gandhi, Yufei Wang, Yifan Wu, Michael Yao, Chris Callison-Burch, James Gee, and Mark Yatskar. A textbook remedy for domain shifts: Knowledge priors for medical image analysis. *Advances in neural information processing systems*, 37:90683–90713, 2024.

[2] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, page 3876, 2022.

[3] Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1151–1163, 2023.

[4] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024.

[5] Vicente Vivanco, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. In *Advances in Neural Information Processing Systems*, 2023.

[6] Xiang Li, Congcong Wen, Yuan Hu, and Nan Zhou. Rs-clip: Zero shot remote sensing scene classification via contrastive vision-language supervision. *International Journal of Applied Earth Observation and Geoinformation*, 124:103497, 2023.

[7] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, 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.

[8] Mingrui Liu, Sixiao Zhang, and Cheng Long. Wukong framework for not safe for work detection in text-to-image systems. *arXiv preprint arXiv:2508.00591*, 2025.

[9] Camilo Carvajal Reyes, Joaquín Fontbona, and Felipe Tobar. Towards sfw sampling for diffusion models via external conditioning. *arXiv preprint arXiv:2505.08817*, 2025.

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[11] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021. https://distill.pub/2021/multimodal-neurons.

[12] Maan Qraitem, Nazia Tasnim, Piotr Teterwak, Kate Saenko, and Bryan A Plummer. Vision-llms can fool themselves with self-generated typographic attacks. *arXiv preprint arXiv:2402.00626*, 2024.

[13] Subaru Kimura, Ryota Tanaka, Shumpei Miyawaki, Jun Suzuki, and Keisuke Sakaguchi. Empirical analysis of large vision-language models against goal hijacking via visual prompt injection. *arXiv preprint arXiv:2408.03554*, 2024.

[14] Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. Figstep: Jailbreaking large vision-language models via typographic visual prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23951–23959, 2025.

[15] Yue Cao, Yun Xing, Jie Zhang, Di Lin, Tianwei Zhang, Ivor Tsang, Yang Liu, and Qing Guo. Scenetap: Scene-coherent typographic adversarial planner against vision-language models in real-world environments. *arXiv preprint arXiv:2412.00114*, 2024.

[16] Justus Westerhoff, Erblina Purellku, Jakob Hackstein, Leo Pinetzki, and Lorenz Hufe. Scam: A real-world typographic robustness evaluation for multimodal foundation models. *arXiv preprint arXiv:2504.04893*, 2025.

[17] Hao Cheng, Erjia Xiao, Jiayan Yang, Jiahang Cao, Qiang Zhang, Jize Zhang, Kaidi Xu, Jindong Gu, and Renjing Xu. Uncovering vision modality threats in image-to-image tasks. *arXiv preprint arXiv:2412.05538*, 2024.

[18] Joanna Materzyńska, Antonio Torralba, and David Bau. Disentangling visual and written concepts in clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16410–16419, 2022.

[19] Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights. *Advances in Neural Information Processing Systems*, 35:29262–29277, 2022.

[20] Hiroki Azuma and Yusuke Matsui. Defense-prefix for preventing typographic attacks on clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3644–3653, 2023.

[21] Maximilian Dreyer, Lorenz Hufe, Jim Berend, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. From what to how: Attributing clip's latent components reveals unexpected semantic reliance. *arXiv preprint arXiv:2505.20229*, 2025.

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

[23] wtcherr (Hugging Face user). unsplash_10k_canny (2hugging face dataset). https://huggingface.co/datasets/wtcherr/unsplash_10k_canny, 2023.

[24] ambityga (Kaggle user). Imagenet100 (kaggle dataset). https://www.kaggle.com/datasets/ambityga/imagenet100, 2023. Accessed: 2025-04-21.

[25] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[26] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

[27] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014.

[28] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021.

[29] Sonia Joseph, Praneet Suresh, Lorenz Hufe, Edward Stevinson, Robert Graham, Yash Vadi, Danilo Bzdok, Sebastian Lapuschkin, Lee Sharkey, and Blake Aaron Richards. Prisma: An open source toolkit for mechanistic interpretability in vision and video. *arXiv preprint arXiv:2504.19475*, 2025.

[30] Kuo-Han Hung, Ching-Yun Ko, Ambrish Rawat, I Chung, Winston H Hsu, Pin-Yu Chen, et al. Attention tracker: Detecting prompt injection attacks in llms. *arXiv preprint arXiv:2411.00348*, 2024.

[31] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

[32] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[34] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[35] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023.

[36] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[38] Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. *arXiv preprint arXiv:2310.03502*, 2023.

[39] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[40] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[41] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[42] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.

[43] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

[44] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36:8958–8974, 2023.

[45] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

[46] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[47] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[48] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

[49] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[50] Christoph Schuhmann, Romain Beaumont, and Nousr y. CLIP-based-NSFW-Detector. https://github.com/LAION-AI/CLIP-based-NSFW-Detector, 2022. GitHub repository, MIT License.

A MLP layers compress information

In Section 4, we observe that $\text{Acc}(P_{\text{image},\ell})$ and $\text{Acc}(P_{\text{typo},\ell})$ increases after attention layers, but consistently drops following MLP blocks.

Thus we estimate the intrinsic dimensionality (ID) of `cls`-token representations across layers using PCA. From a 5% split of the ImageNet-100-Typo training set, we extract residual stream activations for each layer. For each embedding $h_\ell \in \mathbb{R}^d$, we fit PCA and define ID as the smallest number of principal components k such that the cumulative explained variance exceeds 95%:

$$\text{ID} = \min \left\{ k : \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{d} \lambda_j} \geq 0.95 \right\},$$

where λ_j are the PCA eigenvalues. A larger ID indicates higher representational complexity. We report ID across layers and token types to analyze how typographic inputs affect the geometry of CLIP representations.

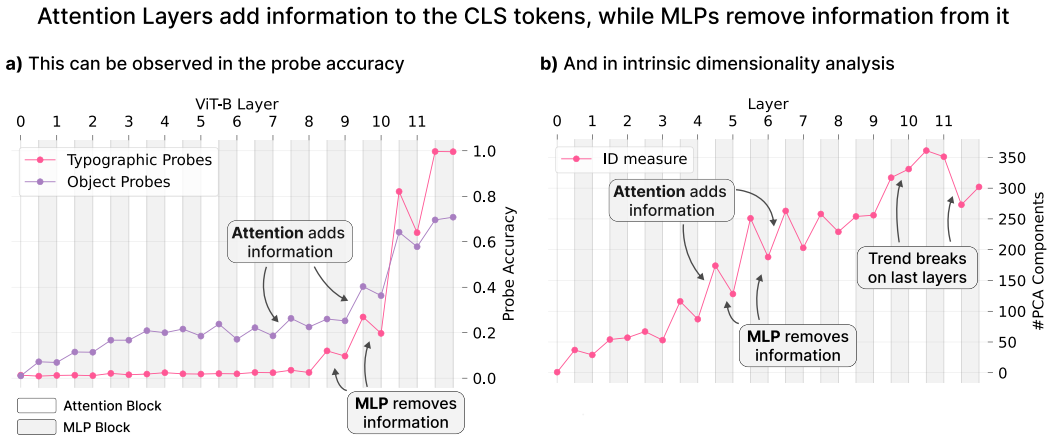

Figure 6: Attention layers increase, MLP layers reduce cls token information. **(a)** Linear probe accuracy rises after attention blocks and drops after MLPs, indicating improved linear accessibility followed by compression. **(b)** Intrinsic dimensionality shows a matching trend, especially in layers 3–7. Exceptions include MLPs at layers 9 and 11 (ID increase) and the attention block at layer 11 (ID drop), suggesting deeper-layer specialization.

Results Figure 6 compares linear probe accuracy (left) and ID (right) across layers of a ViT-B model. We observe a consistent pattern: attention layers tend to increase both probe accuracy and intrinsic dimensionality, while MLP layers reduce them. This trend is most prominent in the middle layers (3–7), suggesting that attention blocks introduce linearly accessible information, whereas MLPs compress or remove it.

This pattern is not consistent throughout the model. The MLPs at layers 9 and 11 exhibit increases in ID, deviating from the overall compression trend. Conversely, the attention block at layer 11 causes a sharp drop in ID. These exceptions indicate that deeper layers may serve more specialized roles.

B Attention sinks for typographic attention heads

We analyze the attention patterns of head $\mathcal{H}_{5,6}$ in ViT-B, which highest $T_{i,\ell}$ score in the model. More specifically we evaluate its spatial attention norm $\|A^*_{5,6}\|$ on ImageNet-100 and ImageNet-100-Typo.

As shown in Fig. 7, $\|A^*_{5,6}\|$ is systematically higher on ImageNet-100-Typo than on ImageNet-100. While the distribution on typopgrahic images is unimodal, the distribution on the original dataset is bimodal. Manual inspection reveals that the high-norm mode in ImageNet-100 contains incidental text in, such as watermarks or copyright tags.

One possible interpretation of these results is that $\mathcal{H}_{5,6}$ uses the cls-to-cls attention as an attention sink [31], to selectively adjust the impact of this specialized typographic attention head.

Building on these finding we evaluate $\mathcal{H}_{5,6}$'s capabilities to predict if a sample x originates from ImageNet-100 or ImageNet-100-typo. The score $\|A^*_{5,6}\|$ ROC-AUC of 0.887, indicating that this attention signal can be used as a robust classifier. In comparison, a linear classifier trained on the same task reaches an ROC-AUC of 1.0, but overfits to superficial typographic features specific to the ImageNet-100-Typo construction. Many images in the original dataset that contain real-world typography are still correctly classified as non-typographic by this probe - supporting the conclusion that it does not generalize beyond the synthetic intervention.

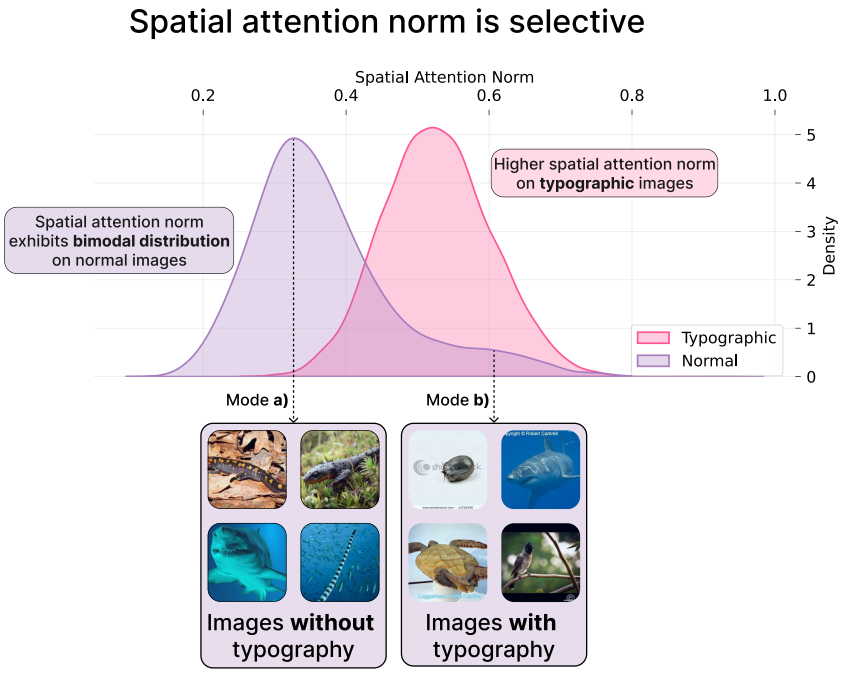

Figure 7: Distribution of the spatial attention norm of ViT-B head $\mathcal{H}_{5,6}$ across ImageNet-100 and ImageNet-100-Typo. The norm is consistently higher for typographic images, while the non-typographic distribution is bimodal. Manual inspection links the higher mode to incidental text, suggesting that the head selectively activates in response to typography, regardless of its origin.

C Datasets

For the zero-shot evaluation we tested a variety of datasets, grouped below by purpose.

RTA-100 [20] consists of 1,000 handcrafted typographic attacks, each written on a Post-it note and overlaid onto natural images.

Disentangling [18] includes 180 typographic attacks, also written on Post-its, designed to probe the separation of visual and textual features in multimodal models.

PAINT-DS [19] comprises 110 Post-it-based typographic attacks and serves to evaluate patch-level vulnerabilities in vision-language models.

Food-101 [27] is a standard image classification dataset containing 101 food categories, each with 1,000 images.

FGVC-Aircraft [26] (referred to as *Aircraft* in our paper) is a fine-grained classification benchmark consisting of 10,000 images across 100 aircraft variants.

ImageNet-100 [24] is a subset of the ImageNet-1k dataset [25], containing 100 object and animal classes with approximately 1,000 images per class.

D Reproducibility

All experiments were conducted on a single RTX Titan GPU with 24 GB of VRAM. For the control dataset, we use a 5% balanced split of the ImageNet-100 training set, where 5% of samples are randomly drawn from each class. This choice approximately matches the size of the validation set and reduces computational overhead during evaluation. We set the accuracy degradation threshold to $\epsilon = 0.01$. Models were accessed through the `OpenCLIP` library [28], and all interventions were implemented using the `Prisma` library [29].

E Relationship between probes and $T_{i,\ell}$ scores

As discussed in Sec. 4, we observe a strong correspondence between layers with elevated $T_{i,\ell}$ scores and those where the typographic probe $P_{\text{typo},\ell}$ exhibits sharp increases in accuracy. In this section, we extend this analysis across all evaluated model sizes to support the trends previously shown for ViT-B. Figures 8 through 12 visualize this relationship for each model. To improve readability, we transpose the probe accuracy plots: the y-axis denotes the layer index, while the x-axis indicates probe accuracy. Adjacent to each plot, we show the full $T_{i,\ell}$ scores across attention heads and layers, with the y-axis again denoting the layer index and the x-axis representing head indices. These figures underpin the alignment between layers that showcase high $T_{i,\ell}$ values and layers where typographic information becomes linearly accessible in the cls token.

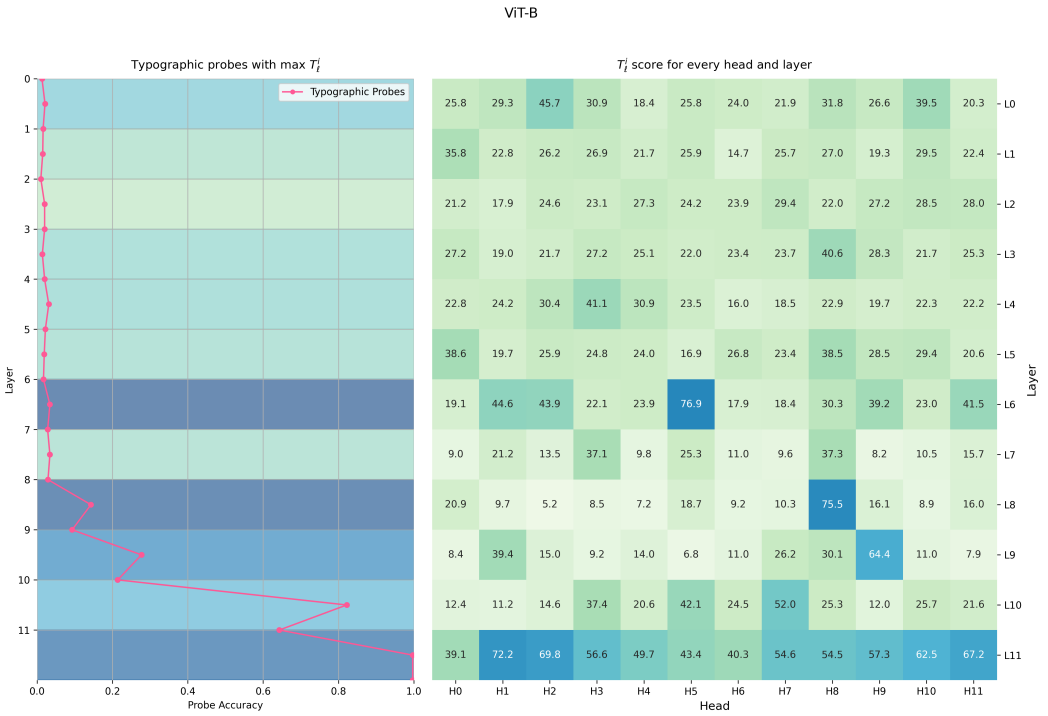

Figure 8: ViT-B: cls probe accuracy (left) and full matrix of $T_{i,\ell}$ scores (right), across all heads and layers. Layers with high typographic attention scores tend to coincide with increased probe accuracy.

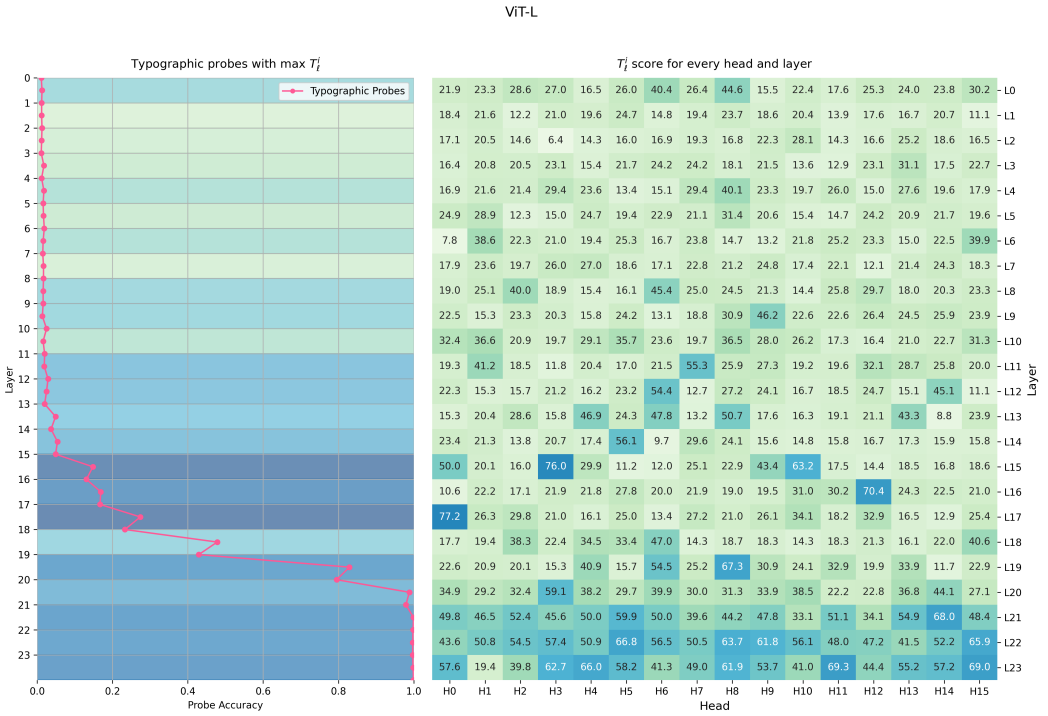

Figure 9: ViT-L: cls probe accuracy (left) and $T_{i,\ell}$ scores (right). A strong layer-wise correlation is visible between typographic attention and linearly decodable signal.

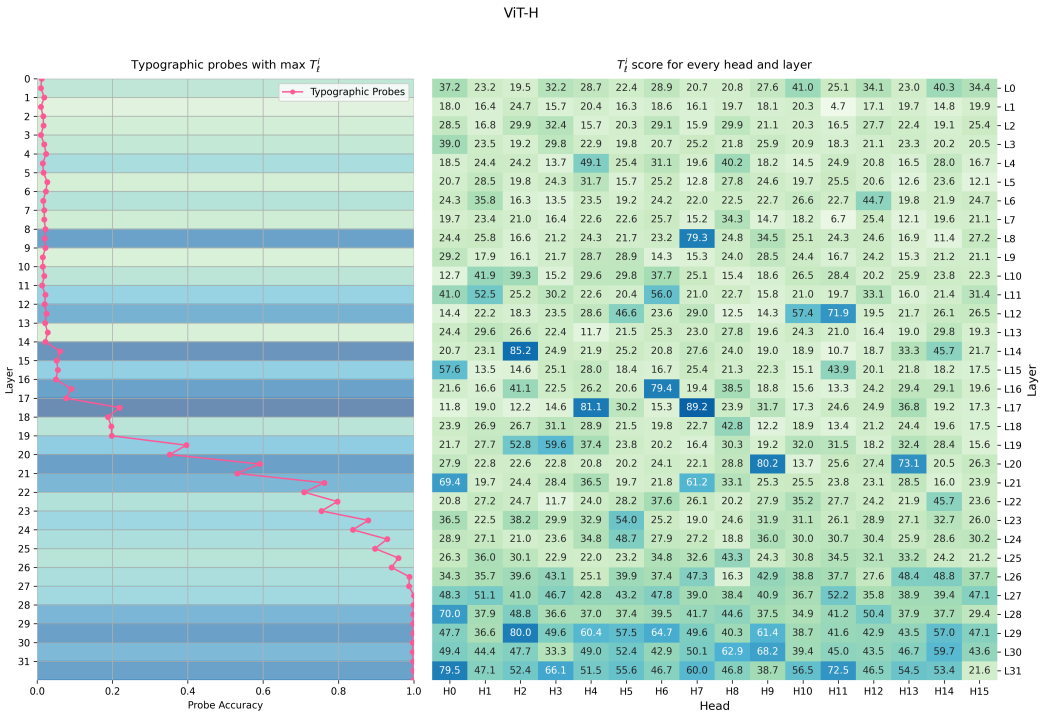

Figure 10: ViT-H: cls probe accuracy (left) and $T_{i,\ell}$ scores (right). The emergence of typographic signal remains aligned with elevated attention scores across heads.

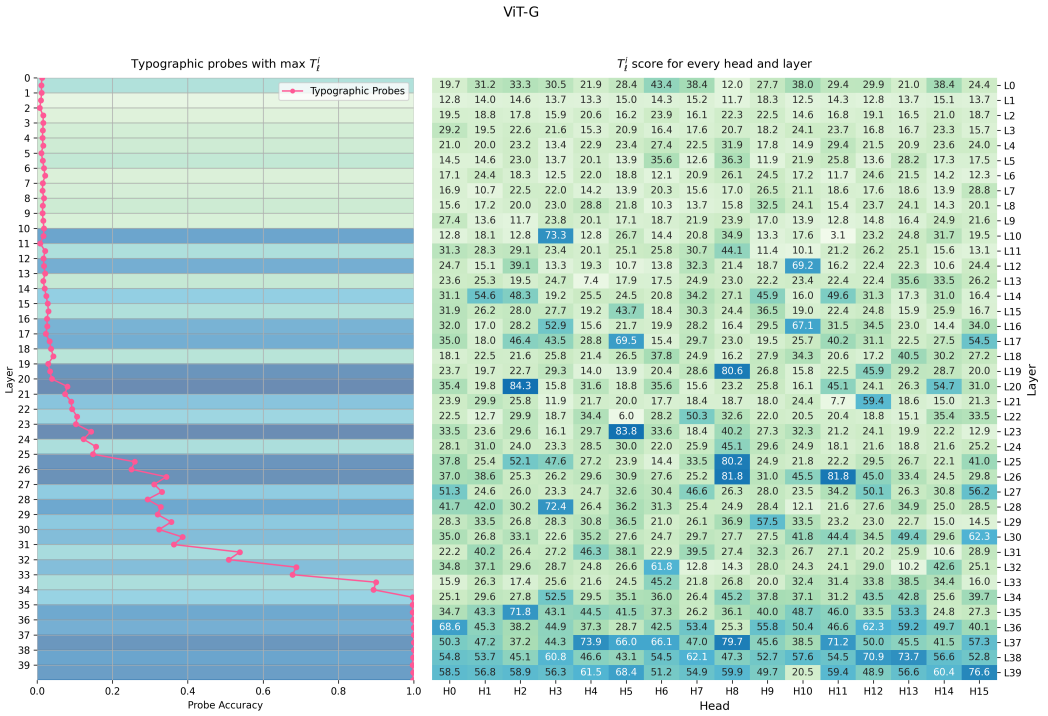

Figure 11: ViT-G: cls probe accuracy (left) and $T_{i,\ell}$ scores (right). Typographic information flow is consistently correlated with high-scoring heads.

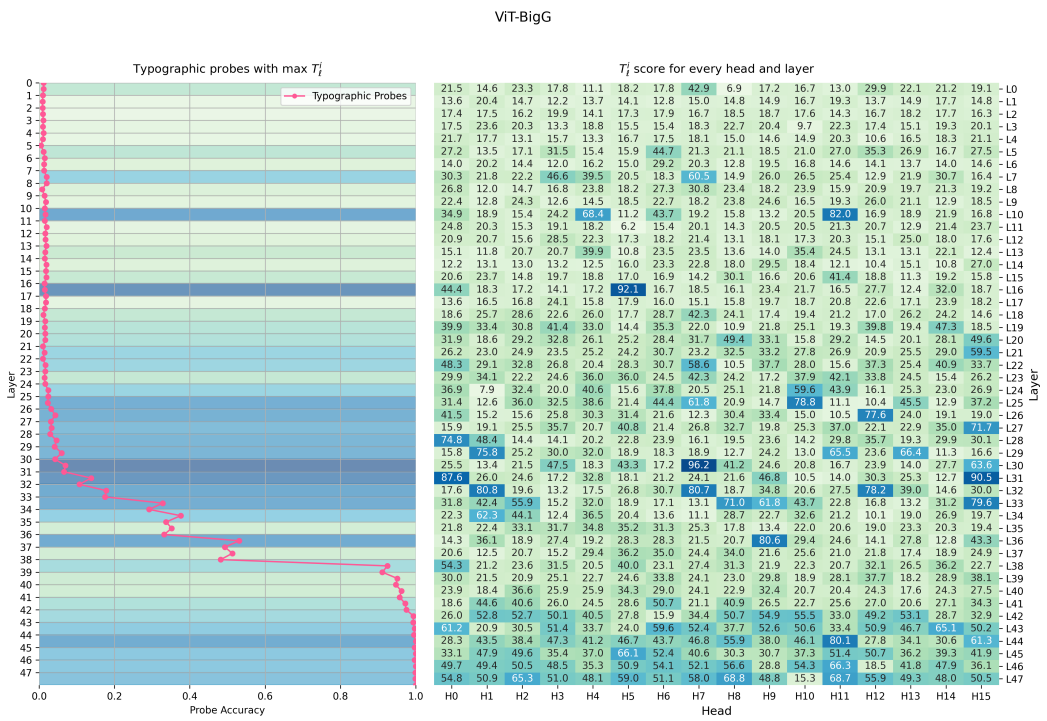

Figure 12: ViT-BigG: cls probe accuracy (left) and $T_{i,\ell}$ scores (right). Even at billion-parameter scale, typographic information remains concentrated in specific layers and heads.