Unmasking the functionality of early layers in VLMs

Max Hartman*
University of Illinois Urbana-Champaign
maxh3@illinois.edu

Vidhata Arjun Jayaraman*
University of Illinois Urbana-Champaign
vidhata2@illinois.edu

Moulik Choraria
University of Illinois Urbana-Champaign
moulikc2@illinois.edu

Akhil Bhimaraju
University of Illinois Urbana-Champaign
akhilb3@illinois.edu

Lav R Varshney
Stony Brook University
lav.varshney@stonybrook.edu

Abstract

Recent work on analyzing the functionality of vision language models has observed that key multimodal processing occurs in the middle to late layers. This raises a natural question about the role of early layers, which is the central focus of our investigation. To explore this, we employ layer skipping as a more suitable alternative to the commonly used attention masking, motivated by our finding that masking vision tokens in early layers substantially degrades performance, whereas skipping the same layers does not. Using layer skipping, we find that vision tokens largely pass unaltered through the early layers of the language model, supporting the view that these layers primarily function as copying heads for visual tokens. This insight highlights opportunities to improve model efficiency by reducing redundant computation in the early stages.

1. Introduction

With the rapid advancement and widespread adoption of large language models (LLMs), interpretability has become a central research focus [19]. Interpretability research has pushed our understanding on how models internally represent and process information further, with studies showing that a combination of attention heads and MLP layers, operating as circuits, can specialize to perform functional objectives such as copying, induction, and factual recall [5, 17, 22]. Specifically, Elhage et al. [5] demonstrated that induction heads in GPT-2 copy tokens from earlier in the context, while Olsson et al. [17] showed these induction heads are also the primary mechanism driving in-context learning. More recently, circuit tracing has been used to trace computational pathways that implement a specific behavior, such as multi-step reasoning and chain-of-thought faithfulness [1, 12]. Meanwhile, sparse autoencoders have been used to discover and isolate interpretable features within model activations [6, 21].

Since vision language models (VLMs) are often built with an LLM as their backbone [2, 11, 13, 23], many of these techniques can be directly extended to them. For instance, activation probing is a commonly used technique that maps hidden states into interpretable formats—one common example being the logit lens [16], that was recently used in VLMs [15] to show that visual token representations in VLMs become increasingly human-interpretable across layers. Another line of work relates to determining the functionality of certain layers. Akin to studying layer functionality in LLMs, [14, 24], Jiang et al. [10], Skean et al. [20] show that the middle-to-late layers act as the main site for multimodal processing (and therefore, represent the best sites for extracting) information. When it comes to early layers, recent empirical evidence [4] suggests that for VLMs, early layers are often redundant for multimodal processing, and that multimodal tokens can be trained to bypass these layers for improved efficiency [4]. However, the reasons for this redundancy and the way multimodal representations evolve in the early layers remain unclear.

In this work, we investigate why this redundancy exists. We first argue for the use of layer skipping as a tool for exploring counterfactuals—i.e., examining VLM behavior with or without multimodal tokens—over the more commonly used attention masking. We then leverage this approach to motivate our central hypothesis that the early layers of VLMs function similarly to *copying heads*, largely preserving internal multimodal representations. We substantiate this with analytical experiments using established tools from the interpretability literature. Our results provide new insight into the early-layer redundancy observed in prior work and open up pressing questions on probing the mechanisms that give rise to this phenomenon.

2. Layer Skipping and Masking in VLMs

We seek to understand how early layers of VLMs operate, with special regard to multimodal tokens. Previous works have used masking as an perturbation to see how inputs are used. In this section, we demonstrate that its applicability is limited via a counter-example, showing that masking degrades model performance quickly and may lead to incorrect conclusions from such studies. To avoid this, we use layer skipping.

2.1. Mathematical Formulation

For the following experiments, we consider variations of perturbed forward passes beyond the standard forward pass. For clarity, we state them explicitly. Let $X = \begin{pmatrix} X_{text} \\ X_{vis} \end{pmatrix} \in \mathbb{R}^{(n_{text}+n_{vis})\times d}$ correspond to the text and vision input and let $\mathrm{VLM}^{\ell}(X)$ be the output of the vision-language model with ℓ layers when the input is X. Then:

1. **Baseline**: $Y^{base} = VLM^{\ell}(X)$ represents a standard forward pass of the model.
2. **Masked input**: $Y^{base} = \mathrm{VLM}^{\ell}(M^{\text{vis},k}_{i:j}(X))$ where $M^{\text{vis}}_{i:j}(X)$ masks the attention to/from the vision tokens upto layer k, with $1 \leq i \leq j \leq n$ for some visual-text prompt of size n.
3. **Layer skip**: $Y^{skip} = \mathrm{VLM}^{\ell-k}(\begin{pmatrix} \mathrm{VLM}^{k}(X_{text}) \\ X_{vis} \end{pmatrix})$, which runs the model on just the text input for the first k layers and then adds the vision tokens back into the prompt after k^{th} layer (see Fig. 1).

2.2. Experimental Setup

In this experiment, we consider a direct comparison of the performance degradation of layer masking versus skipping. We use Llava 1.5 7B and 13B [13] and the Visual7W dataset [25] with user queries formatted to first hold the vision tokens, then the question as a multiple choice answer to be responded with A, B, C, or D. Because of the breadth of questions in the dataset, asking multiple choice questions still represents complex generation tasks with simple accuracy validation while still allowing simple analysis.

2.3. Experimental Results

From Table 1, skipping visual tokens in the first 4 layers yields considerably higher performance than masking them, while Table 2 shows a similar advantage when in the first 8 layers. Furthermore, from Figure 2, the cosine similarity between the representations of the masking study and the baseline is much lower than that of the layer skipping study and the baseline. This shows that there is a fundamental difference between the two methods, with masking leading to irrecoverable loss of information from visual representations. On the other hand, while layer skipping may interfere with positional encodings, it is much less likely to introduce these unexpected perturbations or noise which obscures the counterfactual results.

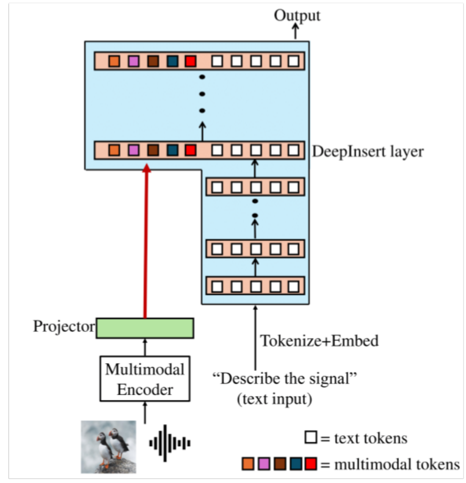


Figure 1. Visual representation of early layer skipping [4]. Specifically, the visual tokens are not passed into the first few layers but instead, directly inserted along with the prompt to the chosen layer for insertion.

Condition	Masking	Skipping
Baseline	0.579	0.579
0–3	0.370	**0.549**
0–7	0.316	**0.357**
0–11	**0.269**	0.264
0–15	0.2760	**0.284**

Table 1. Accuracy of correct answer prediction when masking vs. layer skipping layers $0 - n$ in Llava 1.5 7B and 13B.

Condition	Masking	Skipping
Baseline	0.778	0.778
0–3	0.378	**0.788**
0–7	0.351	**0.695**
0–11	0.264	**0.433**
0–15	0.268	**0.400**

Table 2. Accuracy of correct answer prediction when masking vs. layer skipping layers $0 - n$ in Llava 1.5 13B.

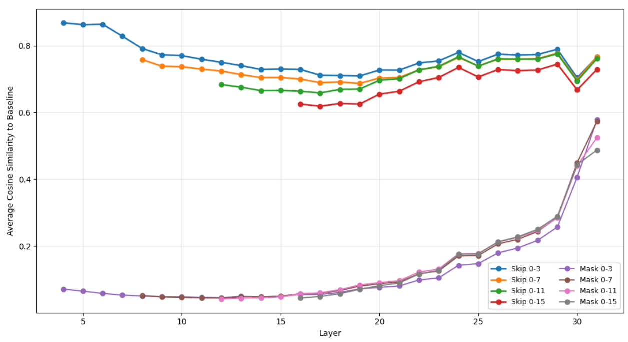


Figure 2. Layer-by-layer cosine similarity between baseline hidden states and those obtained with masking or layer skipping using Llava 1.5 7B. The hidden states with the layer skipping method showed much higher cosine similarity to the baseline hidden states than the hidden states with masking.

3. Early layers are vision copying heads

One may expect masking and skipping studies to yield similar results, since both aim to achieve the same goal of preventing the model from using visual tokens in order to probe their specific role. Yet, we observe a marked performance gap between the two, prompting the question of why this discrepancy arises. To address this, we start with the key observation from the previous section that directly inserting vision information into the later layers of the model produces little change in VLM performance. This suggests that the trained model expects vision representations, after passing through all layers, to reside in a subspace closely aligned with that of skipped (and thus unaltered) tokens. Similar conclusions have been drawn in prior work, though with model retraining [4]. This motivates our hypothesis that the early layers, consisting of both attention and MLP components, function as *copying heads* for visual tokens. In the remainder of this section, we test this hypothesis through an analysis of hidden states, unembedded visual tokens, and an additional skipping-based perturbation study.

3.1. Definition

Induction heads were defined in [5] as a mechanism consisting of two parts: a copying head which copies the previous token and an inductive head which completes the memorized pattern learned from the training data. As an example, if an induction head learns from training data that $[a]$ precedes $[b]$, then when the induction head sees an $[a]$ in practice, it copies $[b]$ to be the prediction.

Because induction heads have also been hypothesized to exist in larger language models and serve as an underlying mechanism for in-context learning [17], we hypothesize that the early layers of VLMs also serve as copying heads. More rigorously, we give the following definition.

Definition (Copying Head): *A layer ℓ functions as a copying head if $d(x_{\ell-1}, x_\ell) < \epsilon$ for some metric $d : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ and small $\epsilon > 0$ where $x_i \in \mathbb{R}^d$ is the representation of a token at layer i*

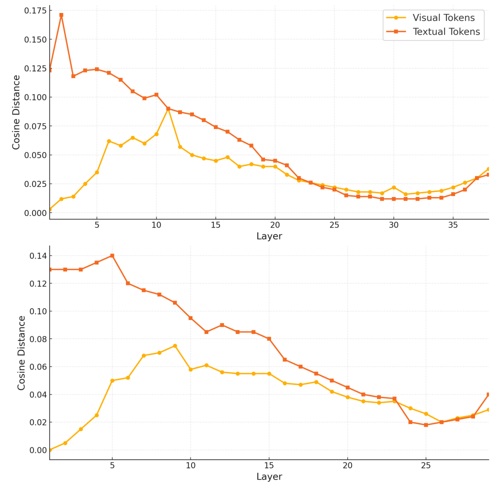


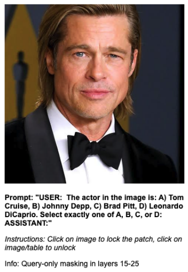
Figure 3. Cosine distance of visual and textual hidden states of adjacent layers in Llava 1.5 7B (bottom) and 13B (top). In the early layers the distance between hidden states of adjacent layers is quite small.

3.2. Empirical Evidence

In these experiments use Llava 1.5 7B and 13B [13] and the Visual7W dataset [25] with user queries formatted to first hold the vision tokens. The answers are left open-ended for the experiments which don't compute an accuracy and use multiple choice (similar to 2.2) if an accuracy is required. In Figure 3, we run a forward pass of the model and computed the average cosine distance between the hidden states of tokens in adjacent layers across examples. We can see that in the early layers the hidden states of the visual tokens experience minimal change, relative to the textual tokens. This indicates that the model seems to copy the hidden states of visual tokens between early layers.

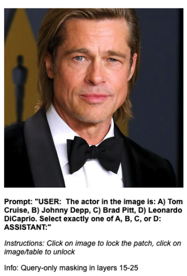
To present further evidence of these copying heads, we use the *logit lens*, as in [15], to analyze the token predictions in the early layers for visual tokens versus textual tokens. From Figure 4, the image token predictions for the early layers stay the same, providing evidence for the claim that these layers act as copying heads. However, from Figure 5, the text token predictions for the early layers vary layer-by-layer, indicating that these layers do not act as copying heads for text tokens.

We further quantify these results by finding the probability of any layer predicting the same token as the initial layer. In Figure 6, we see that initial token prediction of the first layer has more than an 80% chance of being the predicted token of layers 2-4. This suggests that the first 4 layers have a high probability of repeating the initial token



Token/Layer	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5	Layer 6
<IMG154>	AXI	AXI	eye	eye	AXI	eye
<IMG155>	eye	eye	eye	eye	eye	eye
<IMG156>	schau	schau	schau	schau	schau	AXI
<IMG157>	AXI	AXI	AXI	AXI	AXI	AXI
<IMG158>	schau	schau	schau	schau	schau	schau
<IMG159>	schau	schau	schau	schau	schau	schau
<IMG160>	schau	schau	schau	schau	AXI	AXI
<IMG161>	automat	automat	automat	automat	AXI	automat
<IMG162>	AXI	AXI	AXI	AXI	AXI	AXI
<IMG163>	dust	dust	dust	olas	olas	olas
<IMG164>	dust	dust	dust	dust	dust	oni
<IMG165>	IGN	conf	IGN	conf	conf	conf
<IMG166>	dust	dust	dust	olas	olas	olas
<IMG167>	dust	dust	dust	dust	dust	oni
<IMG168>	dust	dust	dust	dust	dust	oni
<IMG169>	IGN	IGN	IGN	IGN	previous	Einzeln
<IMG170>	lassen	lassen	lassen	onas	onas	ulas

Figure 4. Example logit lens output for a set of image tokens for the early to mid layers using [15]. In these early layers, the image token predictions largely stay the same, indicating a possible copying functionality.



Token/Layer	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5	Layer 6
C	HP	archivi	archivi	ali	otte	oco
)	necess	пута	ublic	Ward	ющей	ющей
Brad	ley	ley	ley	ley	ley	ley
Pitt	ainer	ord	hook	ession	cli	ass
,	eth	пута	씨	Außer	kwiet	kwiet
D	dra	è	ゼ	ゼ	ientes	spole
)	necess	пута	ublic	ole	prospect	etwork
Leon	‸	‸	‸	anten	ardo	ardo
ardo	unte	wohl	—	ogy	ogy	ession
Di	Mot	vert	vert	aries	flat	aries
Cap	acity	acity	acity	�	ita	ital
rio	atin	valid	valid	рез	ft	zon
.	ˆ	пута	пута	guez	esterni	esterni
Select	ive	ive	ive	ively	ive	ive
exactly	pher	Roth	że	Einzeln	Einzeln	exactly
one	hundred	hundred	hundred	�	‸	iv

Figure 5. Example logit lens output for a set of text tokens for early to mid layers using [15]. The text token predictions seem to change layer-by-layer not showing any copying functionality.

and thus functioning as copying heads.

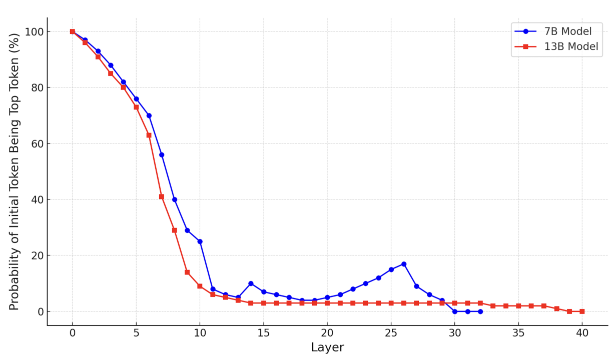


Figure 6. The probability of the initial unembedded visual token being generated in future layers. The probability of predicting the initial token in the first 4 layers stays above 80% in both models.

Finally, to analyze the actual impact of image token representations in early layers on model outputs, we run an experiment similar to the *causal tracing* described in [3].

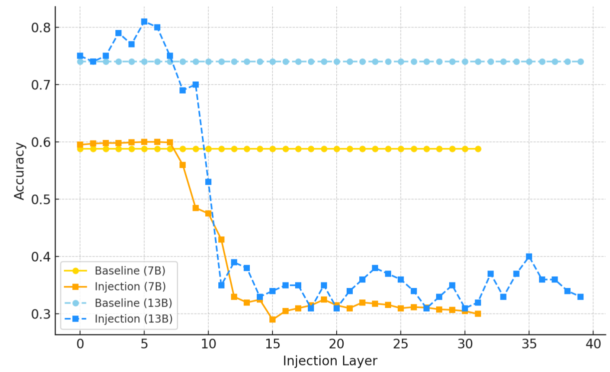


Figure 7. Model accuracy of using an unrelated image for the first n layers before injecting the correct visual representation of a clean forward pass at layer n. The accuracy of the model does not drastically change when an unrelated image is used for the first 7 layers.

We first run a clean forward pass with the correct image and prompt. We then run a modified forward pass with a random image and the original prompt, where at specified layers we inject the visual token representations from the clean forward pass into the layer. Figure 7 shows the model accuracy for different injection layers. We see that the drop in performance only begins to occur around Layer 7, further supporting the claim that the early layers are visual copying heads.

4. Conclusion

In this work, we find that skipping visual tokens in the early layers (i.e. layer skipping) considerably outperforms masking attention to the vision tokens, leading to the hypothesis that the early layers act as copying heads for visual tokens. We validate this by analyzing cosine distances of token representations, analyzing the outputs of the VLM-extension of the *logit lens*, and by inspecting the performance drops when inputting a random image and injecting clean visual token representations at certain layers.

We provide initial experiments for audio-language models in A

However, the most pressing question that demands attention for future work is *why* the model exhibits this behavior and whether it is an inherent limitation of our pretraining methodologies/alignment, or whether this is a deliberate mechanism to demarcate context (i.e. image) vs. the query (the textual prompt) to facilitate better predictions.

In future work, we provide a set of theoretical conditions to determine when layer skipping, such as the early vision tokens, can be used with minimal performance degradation [9].

References

[1] Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. 1

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 1

[3] Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding information storage and transfer in multi-modal large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 4

[4] Moulik Choraria, Xinbo Wu, Akhil Bhimaraju, Nitesh Sekhar, Yue Wu, Xu Zhang, Prateek Singhal, and Lav R. Varshney. Platonic grounding for efficient multimodal language models, 2025. 1, 2, 3

[5] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html. 1, 3

[6] Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *ICLR*, 2025. 1

[7] Yuan Gong, Jin Yu, and James Glass. Vocalsound: A dataset for improving human vocal sounds recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 151–155, 2022. 6

[8] Yuan Gong, Alexander H Liu, Hongyin Luo, Leonid Karlinsky, and James Glass. Joint audio and speech understanding. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023. 6

[9] Max Hartman, Vidhata Jayaraman, Moulik Choraria, Akhil Bhimaraju, and Lav R. Varshney. Skip-it? theoretical conditions for layer skipping in vision-language models, 2025. 4

[10] Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. *arXiv preprint arXiv:2411.16724*, 2024. 1

[11] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 1

[12] Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025. 1

[13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, pages 34892–34916. Curran Associates, Inc., 2023. 1, 2, 3

[14] Zhu Liu, Cunliang Kong, Ying Liu, and Maosong Sun. Fantastic semantics and where to find them: Investigating which layers of generative llms reflect lexical semantics, 2024. 1

[15] Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 3, 4

[16] nostalgebraist. Interpreting gpt: the logit lens. `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`, 2020. 1

[17] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html. 1, 3

[18] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press, 2015. 6

[19] Andrew D. Selbst and Solon Barocas. The intuitive appeal of explainable machines. *Fordham Law Review*, 2018. 1

[20] Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *Forty-second International Conference on Machine Learning*, 2025. 1

[21] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Ex-

tracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. 1

[22] Yifei Wang, Yuheng Chen, Wanting Wen, Yu Sheng, Linjing Li, and Daniel Dajun Zeng. Unveiling factual recall behaviors of large language models through knowledge neurons. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7402. Association for Computational Linguistics, 2024. 1

[23] Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1662, Vienna, Austria, 2025. Association for Computational Linguistics. 1

[24] Yang Zhang, Yanfei Dong, and Kenji Kawaguchi. Investigating layer importance in large language models. In *The 7th BlackboxNLP Workshop*, 2024. 1

[25] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7W: Grounded Question Answering in Images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2, 3

A. Appendix: Extensions to other modalities

As for extensions to other modalities, our initial experiments with audio-language models (LTU-7B [8]) in Table 3 show that skipping early layers also does not significantly impact performance either, suggesting that copying heads may represent a general underlying mechanism for multimodal language models.

Skip Layers	**ESC50**	**VS**
0	81.30	56.39
0–3	81.70	56.36
0–5	80.40	54.86
0–7	74.50	41.21
0–11	66.40	31.05
0–15	32.35	32.36

Table 3. Performance across layers on ESC50 [18] and VS [7] audio benchmarks. Skipping early layers seems to not drastically impact performance.