Unmasking the functionality of early layers in VLMs

Max Hartman[1*], Vidhata Jayaraman[1*], Moulik Choraria[1], Akhil Bhimaraju[1], Lav R. Varshney[2]

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Introduction

- Recent work has shown that key vision processing occurs in the middle-to-late layers of VLMs.
- We compare the performance of DeepInsert [1] and masking to reveal a difference in early layer processing (See Figure 1)
- We show that vision copying occurs in the early layers of VLMs.

Early Layers copy vision tokens

- The clear performance gap (Tables 1 & 2) between skipping and masking shows a functionality difference between the two methods.
- Inspired by [2], we define copying heads as being:

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*

- The mathematical formulation of masking and skipping is:

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**: $Y^{skip} = \mathrm{VLM}^{\ell-k}\left(\begin{pmatrix}\mathrm{VLM}^k(X_{text}) \\ X_{vis}\end{pmatrix}\right)$, 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).

- Where VLM^n is the VLM model with n layers, X_{text} and X_{vision} are the text and vision tokens in the prompt respectively

1: The University of Illinois Urbana-Champaign, Department of Electrical and Computer Engineering and Coordinated Science Lab
2: Stony Brook University, AI Innovation Institute
* Denotes Equal Contribution

Figures & Tables:

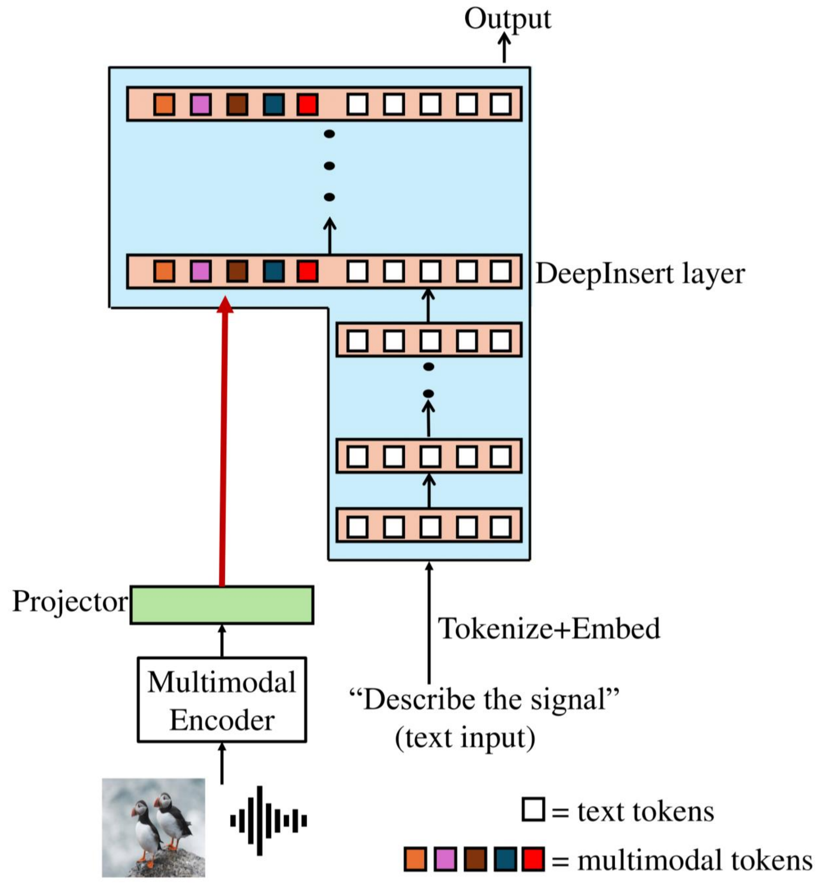


Figure 1. DeepInsert Architecture

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 DeepInsert vs Masking on LLaVA 1.5 7B.

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 DeepInsert vs Masking on LLaVA 1.5 13B.

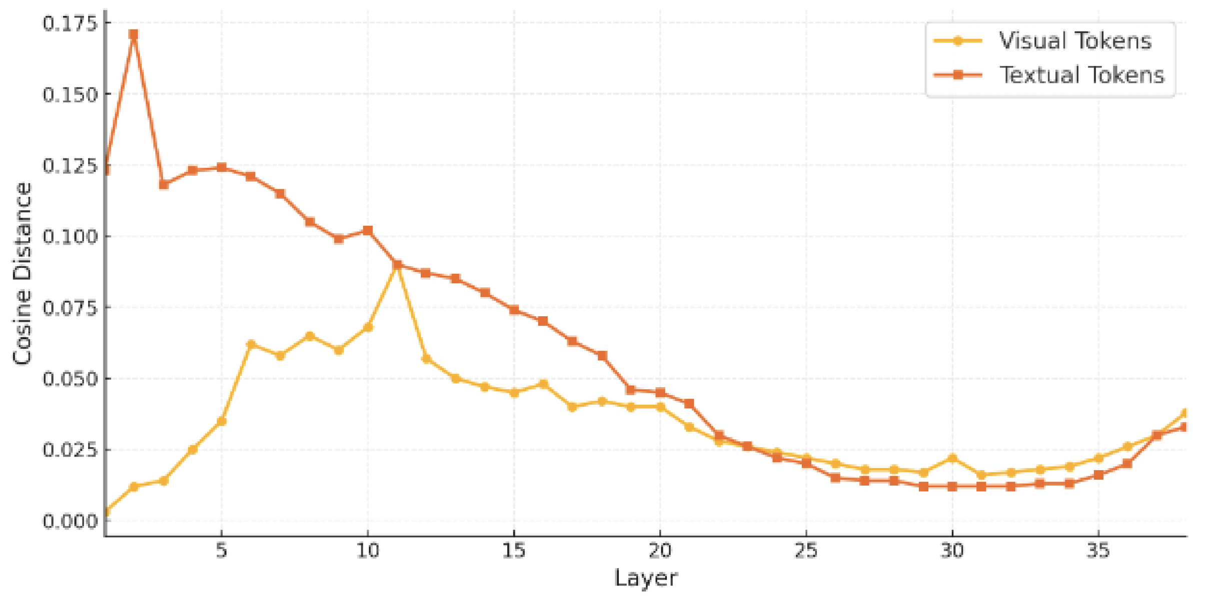
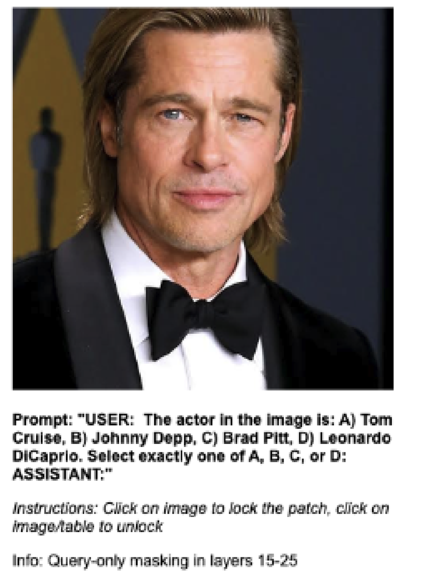


Figure 2. Average cosine distance between adjacent layers in LLaVA 1.5 13B



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

Figure 3. "Logit Lens" outputs for the early layer vision tokens

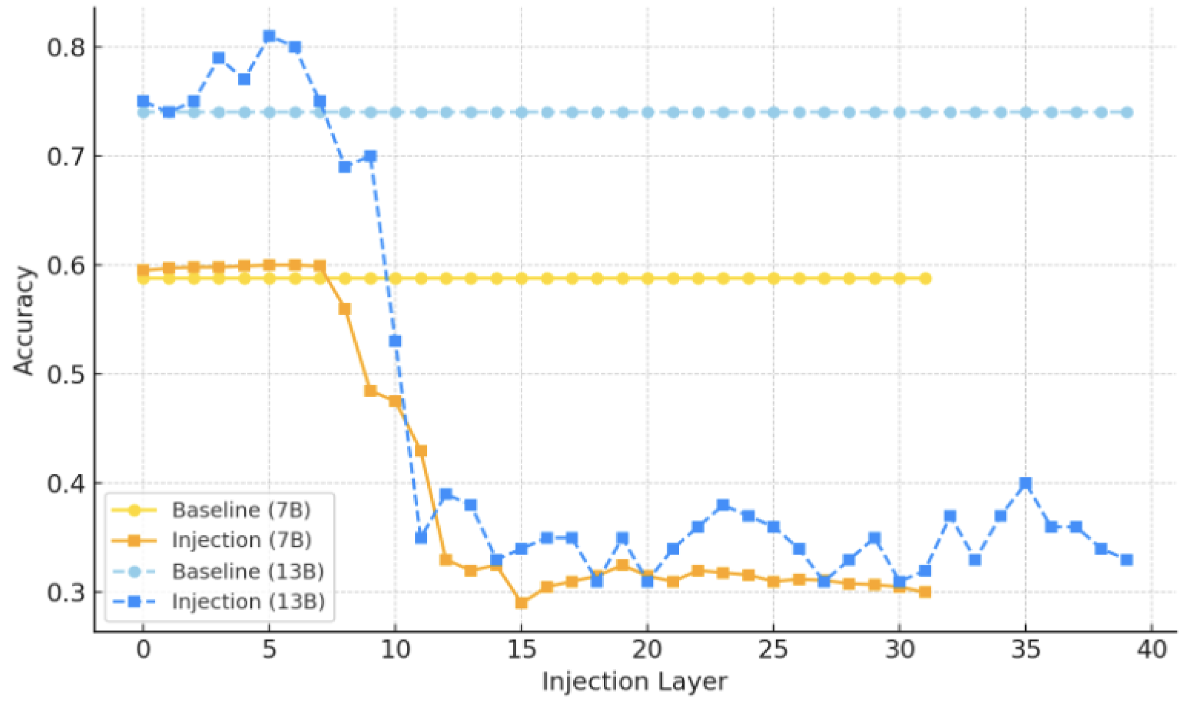


Figure 4. Accuracy of LLaVA 1.5 13B on random image with correct forward pass injected for all layers ≥ n (n = 0, 1, … , 39)

Discussion

- In our subsequent work [3], we theoretically ground when copying occurs.
- We define four notions of redundancy:
  - Geometric: Small average cosine distance
  - Proximal: Small cosine distance with high probability
  - Functional: Optimal mean square estimators are similar
  - Informational: Low conditional entropy

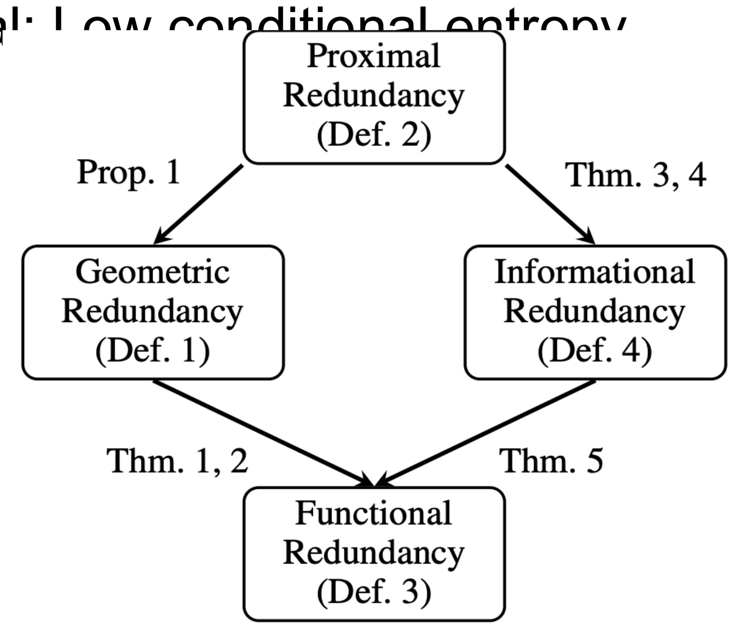


Figure 5. Implications of redundancy notions. See further details in "Skip-It? Theoretical Conditions for Layer Skipping in Vision–Language Models (Posted on Arxiv)

- We validate our framework using DeepInsert (extended to both late entry and early exit) on additional models and datasets.

Conclusion

- We see a performance difference between DeepInsert and masking.
- We show that early layer act as copying heads for vision tokens

References

[1] DeepInsert: Early layer bypass for efficient and performant multimodal understanding, 2025. Choraria et al.

[2] A mathematical framework for transformer circuits, 2021. Elhage et al.

[3] Skip-It? Theoretical Conditions for Layer Skipping in Vision-Language Models, 2025. Hartman et al.

Acknowledgements

We thank the Office of Undergraduate Research (OUR) at University of Illinois Urbana-Champaign for their research grant