# Do Deeper Nonlinear Decoders Outperform the Temporal-Attention MLP for Neural Visual Decoding?

Anonymous Author(s)

## ABSTRACT

We compare six decoder architectures for mapping simulated primate multi-unit spiking activity to semantic image embeddings: linear regression, standard MLP, temporal-attention MLP (TA-MLP), temporal CNN, deep MLP, and wide MLP. Using synthetic neural populations modeled on V4/IT tuning statistics with 128 neurons over 20 time bins, we evaluate top-1 accuracy, top-5 accuracy, median rank, and cosine similarity on 200-class retrieval. The TA-MLP achieves the highest top-1 accuracy (15.6% ± 1.4%), top-5 accuracy (43.2% ± 2.4%), and lowest median rank (18.2 ± 2.1) with only 148K parameters. The temporal CNN (14.2%) and deep MLP (13.4%) approach but do not exceed the TA-MLP despite using 1.3–3.5× more parameters. These results support the finding that the lightweight temporal-attention mechanism provides an effective inductive bias for neural time-series decoding, and more complex architectures do not yield further gains on this task.

## 1 INTRODUCTION

Decoding visual stimuli from intracortical neural recordings requires mapping high-dimensional spatiotemporal spiking patterns to semantic representations. Ciferri et al. [1] introduced the THINGS Ventral Stream Spiking Dataset (TVSD) and demonstrated that a temporal-attention MLP (TA-MLP) consistently outperforms linear and LSTM baselines when mapping 200-ms windows of multi-unit activity to CLIP embeddings [5]. However, they noted that more complex nonlinear architectures might achieve higher performance.

We systematically evaluate this question using synthetic neural populations that replicate the statistical structure of primate ventral stream recordings. Our comparison spans architectures of increasing complexity, from linear regression to deep networks with hundreds of thousands of parameters.

### 1.1 Related Work

The temporal attention mechanism was introduced for neural decoding by Ciferri et al. [1]. CLIP embeddings [5] provide the semantic target space. LSTM [2] and transformer [6] architectures are standard nonlinear sequence models. Representational similarity analysis [3] motivates the cosine similarity evaluation.

## 2 METHODS

*Simulated Data.* We simulate 128 neurons with mixed Gaussian tuning to 200 visual categories across 20 time bins (10 ms each, spanning 200 ms). Neural responses incorporate tuning curves, temporal dynamics, trial-to-trial variability, and noise correlations modeled on V4/IT population statistics [4]. Ground-truth embeddings are 512-dimensional unit vectors.

*Architectures.* **Linear**: Ridge regression from time-averaged firing rates. **MLP**: Two hidden layers (256, 128) with ReLU. **TA-MLP**: Learned temporal attention weights over time bins followed by two MLP layers—the architecture of [1]. **Temporal CNN**: Three 1D convolution layers with pooling. **Deep MLP**: Four hidden layers (512, 256, 128, 64). **Wide MLP**: Two hidden layers (1024, 512).

**Table 1: Decoder architecture comparison on 200-class retrieval.**

| Architecture | Top-1 (%) | Top-5 (%) | Med. Rank | Cosine | Params |
|---|---|---|---|---|---|
| Linear | 5.2 ± 0.8 | 19.8 ± 1.5 | 42.3 ± 3.1 | 0.312 | 26K |
| MLP | 11.8 ± 1.2 | 35.6 ± 2.1 | 24.7 ± 2.4 | 0.458 | 132K |
| **TA-MLP** | **15.6 ± 1.4** | **43.2 ± 2.4** | **18.2 ± 2.1** | **0.524** | **148K** |
| Temporal CNN | 14.2 ± 1.6 | 40.8 ± 2.6 | 20.4 ± 2.6 | 0.498 | 199K |
| Deep MLP | 13.4 ± 1.8 | 39.2 ± 2.8 | 21.8 ± 2.9 | 0.482 | 525K |
| Wide MLP | 12.8 ± 1.5 | 37.8 ± 2.5 | 22.6 ± 2.7 | 0.471 | 1050K |

*Evaluation.* We perform 200-class retrieval using cosine similarity between decoded and true embeddings. Metrics: top-1 accuracy, top-5 accuracy, median rank, and mean cosine similarity, each averaged over 5 cross-validation folds.

## 3 RESULTS

Table 1 shows that the TA-MLP achieves the best performance across all four metrics. The temporal CNN is the closest competitor at 14.2% top-1, trailing by 1.4 percentage points despite 34% more parameters. Increasing depth (Deep MLP, 525K parameters) or width (Wide MLP, 1050K parameters) degrades performance relative to the TA-MLP, suggesting that the temporal attention mechanism provides an inductive bias better suited to neural time-series than generic depth or width.

*Efficiency Analysis.* The TA-MLP trains in 4.5 s compared to 7.1 s (Deep MLP) and 9.4 s (Wide MLP), achieving the best accuracy-per-parameter ratio. The linear baseline, while fastest (0.8 s), produces 3× lower top-1 accuracy.

*Cosine Similarity.* The TA-MLP achieves mean cosine similarity of 0.524 between decoded and true embeddings, compared to 0.312 for linear, 0.458 for MLP, and 0.498 for the temporal CNN, indicating that temporal attention captures semantic structure more faithfully.

## 4 CONCLUSION

Our systematic comparison provides evidence that more complex nonlinear architectures do not outperform the temporal-attention MLP for neural visual decoding. The TA-MLP's learned temporal weighting provides an effective inductive bias for 200-ms neural windows, outperforming both deeper and wider alternatives while using fewer parameters. This supports the conclusion of Ciferri et al. [1] that simple, well-designed architectures can achieve rich decoding performance.

# 5 LIMITATIONS AND ETHICAL CONSIDERATIONS

Our study uses synthetic data rather than actual primate recordings, which may not capture all statistical complexities of real neural populations. The 200-class task may not reveal advantages of complex architectures that emerge at larger scales. Primate neuroscience research raises ethical considerations regarding animal welfare that motivate computational approaches like ours.

# REFERENCES


[1] Francesco Ciferri et al. 2026. Simple Models, Rich Representations: Visual Decoding from Primate Intracortical Neural Signals. *arXiv preprint arXiv:2601.11108* (2026).

[2] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[3] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. 2008. Representational similarity analysis — connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience* 2 (2008), 4.

[4] Sidney R. Lehky and Keiji Tanaka. 2007. Comparison of Shape Encoding in Primate Dorsal and Ventral Visual Pathways. *Journal of Neurophysiology* 97 (2007), 307–319.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*. 8748–8763.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Advances in Neural Information Processing Systems*. 5998–6008.