A Computational Model for Binding by Enhanced Firing Rate: Implementing Smooth Power-law enhancement in Object-Centric Representations

Ishanvir S. Choongh
Cognitive Systems
University of British Columbia
Vancouver, BC Canada V6T 1Z3
ishanvir@student.ubc.ca

Manu S. Madhav
School of Biomedical Engineering
Djawad Mowafaghian Centre for Brain Health
University of British Columbia
Vancouver, BC Canada V6T 1Z3
manu.madhav@ubc.ca

Abstract

The binding problem remains a core challenge in perception. A proposed neural mechanism is binding by firing-rate enhancements (BBRE), where attention amplifies neuronal activity for same-object features without the need for synchronized oscillations. We interpret BBRE as a contrast mechanism and test its functional utility computationally via the Minimal Viable Binding Architecture (MVBA), a proof-of-concept model extending Slot Attention with a smooth power-law enhancement ($f(x) = \mathrm{sign}(x) \cdot |x|^{\alpha}$) to amplify strong signals and suppress weak ones in isolated spatial ("WHERE") and feature ("WHAT") binding modules, inspired by dual-stream theories of the visual system. Seven ablated variants were trained unsupervised on a synthetic dataset using MSE reconstruction loss. Results showed 71.4% MSE improvement overall (*baseline* 1.174 to *full* 0.335), with 18.4% from power-law enhancement alone. Visualizations validated reduced feature bleeding. MVBA empirically supports BBRE as a contrast mechanism, bridging neuroscience, psychology, and machine learning for biologically plausible object-centric learning.

1 Introduction

The binding problem is a fundamental challenge in visual perception: How does the brain integrate distributed sensory features into coherent object representations within its parallel processing architecture? Feature Integration Theory (FIT) [1] provides a key framework, proposing that early vision decomposes scenes into basic features across separate neural maps, but focused attention is required to "bind" them into wholes, preventing miscombinations like illusory conjunctions [1, 2]. Historical debates have contrasted synchrony-based theories, where temporal correlations and oscillatory phases link features [3], with attention-modulated firing rate alternatives [4]. Roelfsema critiques synchrony models for lacking empirical support and proposes binding by enhanced firing rates (BBRE), where attention boosts neuronal activity for features of the same object, creating perceptual separation without oscillations [4].

Computational modeling bridges psychology and neuroscience by enabling testable simulations of binding mechanisms, serving as a tool to evaluate BBRE's functional role as a contrast enhancement mechanism that amplifies relevant signals while suppressing noise. In machine learning, object-centric paradigms decompose scenes into interpretable components, inspired by perceptual grouping [5]. In the Slot Attention paradigm [6], slots are vectors that act as abstract containers for scene components, and they compete iteratively to parse images into objects without supervision, mirroring parallel feature extraction and serial binding.

Despite advances, these models often fail to fully address binding, exhibiting features "leaking" across objects, especially in unsupervised training on complex images. The cause is attributed to features being insufficiently segregated or enhanced relative to the background noise [7]. This gap highlights the need for stricter enhancement/suppression mechanisms beyond standard attention, prompting computational tests of biologically inspired solutions like BBRE as a contrast mechanism.

To address this, we introduce the Minimal Viable Binding architecture (MVBA) as a proof-of-concept model that interprets BBRE as a contrast mechanism and tests whether it improves binding, assessed through reconstruction quality. We implement this as a smooth power-law enhancement, which amplifies strong features while dampening weak ones through a dynamic top-down transformation, without modeling actual firing-rate dynamics in the neural network. Our central hypothesis is that this power-law enhancement, as a functional analog of BBRE, significantly improves reconstruction quality measured via model ablations on multi-object scenes, providing empirical validation for rate-based binding as a contrast mechanism.

2 Methods

2.1 Model Architecture and Design Rationale

MVBA is a modular neural network for object-centric representations with explicit binding mechanisms, comprising six modules that process visual input hierarchically (Fig. 1a). Feature extraction uses a three-layer CNN to yield 64-dimensional visual features at full 64×64 image resolution, concatenated channel-wise with 8-dimensional learnable positional encoding for a $[72 \times 64 \times 64]$ tensor. This map feeds Slot Attention, inspired by [6], where four slots (128-dimensional) compete via multi-headed attention (4 heads, 32 dims/head) over three iterations, with symmetry breaking through Gaussian initialization and dimensional biasing to prevent identical convergence. The slot states are refined through gated recurrent unit (GRU) updates and residual MLPs. After three iterations of Slot attention, each slot produces a dynamic top-down $\alpha \in [1, 3]$ parameter via dual 64-hidden MLPs, yielding separate spatial (α_s) and feature (α_f) parameters for adaptive sharpening.

The pipeline then splits into Spatial Binding and Feature Binding. Spatial Binding computes "WHERE" spatial maps via 32-dimensional query-key matching (slot queries: $128 \rightarrow 32$; feature keys: 1×1 conv $72 \rightarrow 32$), with power-law enhancement on logits before softmax to amplify strong matches and suppress weak ones, embodying biological attention by simultaneously enhancing relevant signals and suppressing noise [2]. Multi-scale refinement (3×3, 5×5, 7×7 kernels, 16 channels) enforces coherent regions. Feature Binding extracts "WHAT" properties by weighting features with spatial maps, applying power-law enhancement, and using slot-gated pooling projected to 128 dimensions. Reconstruction attempts to recreate the original image by decoding slots via upsampling ($128 \rightarrow 256 \rightarrow 64 \times 8 \times 8$, then transposed conv to $3 \times 64 \times 64$) with per-slot RGB/masks combined via normalized softmax, enabling end-to-end differentiability. This design isolates binding for ablation, with a power-law mimicking contrast akin to BBRE via post-matching sharpening.

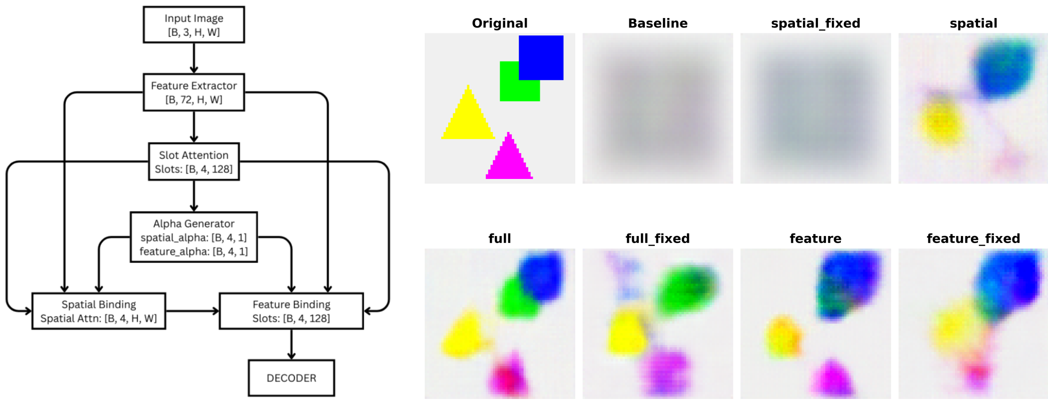
The core innovation provided by MVBA is smooth power-law enhancement, mimicking neural contrast: $f(x) = \text{sign}(x) \cdot |x|^{\alpha}$, where $\alpha > 1$ is the learned parameter that amplifies strong signals and suppresses weak ones, inspired by gain modulation [8]. As a BBRE analog, it simulates firing rate increase for attended features, creating perceptual separation against the background.

2.2 Model Variants and Ablation Setup

Seven variants were constructed to isolate effects: *baseline* (Slot Attention only), *spatial_fixed* ($\alpha_s = 1$, no Feature Binding), *feature_fixed* ($\alpha_f = 1$, no Spatial Binding), *full_fixed* ($\alpha_s = \alpha_f = 1$), *spatial* ($\alpha_s \in [1, 3]$, no Feature Binding), *feature* ($\alpha_f \in [1, 3]$, no Spatial binding), *full* ($\alpha_s, \alpha_f \in [1, 3]$). Variants were trained with identical hyperparameters end-to-end.

2.3 Dataset, Training, and Evaluation Metrics

We generated the *SimpleObjects* dataset (8,000 training, 2,000 test) using OpenAI o3 LLM [9], featuring 2 to 4 geometric shapes with randomized colours/sizes/positions, including overlaps for binding challenges. For example, Figure 1b. "Original" image is sample 0961 from the *SimpleObjects* test dataset. Training used unsupervised Mean-Squared Error (MSE) reconstruction loss on a



(a) MVBA architecture

(b) Reconstruction quality comparison

Figure 1: **(a) MVBA architecture**. Visual input is processed through six components: feature extraction, slot attention, dynamic alpha generation, spatial binding, feature binding, and image reconstruction (decoder). Tensor shapes denote batch size B, channels, height H, and width W; see Method 2.1 for module details. **(b) Reconstruction quality comparison**. Original image for test sample 0961 and reconstructed images from the 7 MVBA variants. The *_fixed* suffix indicates α being restricted to 1 (no enhancement). The *full* model demonstrates reduced feature bleeding and more coherent object boundaries.

single 16GB GPU (optimizer: Adam, batch size: 16, epochs: 50). We computed the following metrics for evaluation: Mean-Squared Error (MSE), mean absolute error (L1), Peak Signal-to-Noise Ratio (PSNR $= 20\log_{10}(1/\sqrt{\mathrm{MSE}+\epsilon})$), and Structural Similarity Index Measure [10] (SSIM $= \frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)}$, where μ_x, μ_y are means, σ_x^2, σ_y^2 are variances, σ_{xy} is covariance, $c_1 = 0.0001, c_2 = 0.0009$). We used per-slot MSE for decomposition. For statistical analysis, we conducted paired t-tests (21 comparisons) to assess metric differences, with Cohen's d and 95% bootstrap confidence intervals (CIs) over 1000 resamples, averaged over the 2000 test images. The codebase was implemented in PyTorch, see Appendix B.

3 Results

Across the seven variants, metrics showed substantial MSE reductions (Table 1), with a drop from *baseline* MSE = 1.174 to *full* MSE = 0.335 (71.4% improvement). These improvements were driven by architecture gains (*baseline* to *full_fixed*: 65.0%) and BBRE enhancements (*full_fixed* to *full*: 18.4%). Secondary metrics showed the same pattern for *full* vs. *baseline*: PSNR$= 5.36$ vs. -0.32, SSIM$= 0.714$ vs. 0.086, L1$= 0.190$ vs. 0.566, with reduced standard deviations indicating stability. Effect sizes were meaningful (e.g. *baseline* to *full*: $d = 2.37$, *full_fixed* to *full*: $d = 0.49$), all $p < 0.001$, and 95% bootstrap CIs confirmed reliable differences. See Appendix A for the complete statistical analysis (Tables 2, 3, 4, 5).

Table 1: Mean performance metrics across MVBA variants

Variant	MSE (std) (lower is better)	PSNR (std) (higher is better)	SSIM (std) (higher is better)	L1 (std) (lower is better)
baseline	$1.1736(0.4620)$	$-0.32(1.89)$	$0.0864(0.0799)$	$0.5664(0.1996)$
spatial_fixed	$1.1758(0.4624)$	$-0.33(1.89)$	$0.0685(0.0791)$	$0.5357(0.1912)$
spatial	$0.6003(0.3052)$	$2.81(2.36)$	$0.4844(0.3600)$	$0.3118(0.1147)$
feature_fixed	$0.5438(0.2994)$	$3.31(2.49)$	$0.5883(0.3019)$	$0.2662(0.1096)$
feature	$0.4424(0.2472)$	$4.22(2.51)$	$0.6582(0.3012)$	$0.2219(0.0905)$
full_fixed	$0.4108(0.2421)$	$4.61(2.63)$	$0.6210(0.3489)$	$0.2131(0.0913)$
full	$\mathbf{0.3353}(\mathbf{0.1791})$	$\mathbf{5.36}(\mathbf{2.36})$	$\mathbf{0.7141}(\mathbf{0.2923})$	$\mathbf{0.1902}(\mathbf{0.0703})$

Pairwise t-tests (21 comparisons) highlighted architecture significance (e.g., *baseline* to *full_fixed*: $d = 2.32$, 65.0% improvement, *baseline* to *feature_fixed*: $d = 2.08$, 53.7% improvement) and power-law boosts, especially in spatial binding (*spatial_fixed* to *spatial*: $d = 2.00$, 49% improvement). Synergistic effects emerge, as *full* model gains (71.4%) were less than the summed components, showing non-additive benefits. Shapiro-Wilk tests noted normality violations, but large $n = 2000$ ensured robustness. Visualizations of reconstructions across variants demonstrated improved binding: *baseline* showed blurred masses, while the *full* model showed coherent bound objects (Fig. 1b).

4 Conclusion

The results robustly support the hypothesis that smooth power-law enhancement, as a BBRE analog, improves reconstruction quality in object-centric models ($p < 0.001$ across comparisons). Isolation using restricted variants *full_fixed* ($\alpha = 1$) to *full* (dynamic $\alpha \in [1, 3]$) yielded an MSE difference of 0.07552 ($t = 22.07$, $p = 8.15 \times 10^{-97}$, $d = 0.494$, 18.38% MSE reduction, 95% CI $[17.0\%, 19.8\%]$), with spatial binding gaining most (*spatial_fixed* to *spatial*: 49% MSE reduction, $d = 2.00$). Overall, the improvement in MSE from *baseline* to the *full* model was MSE$=$ 1.174 to 0.335 (71.4%). Synergistic effects were evident: the 71.4% improvement $<$ sum of components (*feature* $62.3\%+$ *spatial* 48.9%), mimicking integration due to attention [2]. This affirms power-law's imitation of an attention-based contrast enhancement mechanism. Limitations of this study include SimpleObject's toy nature, which limits generalizability to complex real-world scenes, and reconstruction quality serving as an indirect proxy for binding quality. Future work should focus on extending the model to complex scenes and assessing its performance using direct metrics, and having dynamic, ranked slots for robust perception. Overall, MVBA bridges neuroscience, psychology, and machine learning to empirically validate a testable analog for firing-rate contrast enhancements in attention.

5 Related Work

Computational modeling is a useful tool for evaluating binding mechanisms in visual perception. Recent approaches have integrated spiking neural networks (SNNs) and artificial neural networks (ANNs) to address binding via covariance and spike timing, relying on synchrony [11]. However, such oscillation-based theories are critiqued for lacking sufficient empirical support, advocating instead for binding by enhanced firing rates (BBRE) [4], yet computational tests of this mechanism remain scarce beyond conceptual critiques. In machine learning, object-centric paradigms extend these ideas, with Slot Attention [6] enabling unsupervised scene decomposition into slots, but prone to feature bleeding. Recent variants prioritize architectural refinements: GLASS [12] uses guided latent diffusion for semantic slots; ROLA [13] optimizes attention for real-world data; Adaptive Slot Attention [14] dynamically adjusts slot numbers; and probabilistic priors to Slot Attention improve object identifiability [15]. These represent architectural tweaks rather than biologically plausible enhancements. Together, these works highlight a gap in models empirically testing neuroscientific/psychological theories like BBRE as a functional enhancement mechanism. MVBA addresses this gap by implementing smooth power-law amplification in dual-stream binding, validating the contrastive role of a rate-based enhancement through ablations.

6 Acknowledgements

We would like to acknowledge Doug Altshuler and Khanh Dao Duc for their valuable mentorship to Ishanvir S. Choongh, comments and feedback on this manuscript. This work was originally conceived as a project for UBC course COGS 402, instructed by Christopher Mole. This work was supported by the Tier 2 Canada Research Chair in Neural Circuits of Cognition and Control (CRC-2020-00226), and a New Frontiers in Research Fund Exploration grant (NFRFE-2022-00524).

References

[1] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[2] Anne Treisman. How the deployment of attention determines what we see. In *Progress in Psychological Science around the World. Volume 1 Neural, Cognitive and Developmental Issues.*, pages 245–277. Psychology Press, 2013.

[3] Wolf Singer. Neuronal synchrony: a versatile code for the definition of relations? *Neuron*, 24(1):49–65, 1999.

[4] Pieter R Roelfsema. Solving the binding problem: Assemblies form when neurons enhance their firing rate—they don't need to oscillate or synchronize. *Neuron*, 111(7):1003–1019, 2023.

[5] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *International conference on machine learning*, pages 2424–2433. PMLR, 2019.

[6] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020.

[7] Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, et al. Bridging the gap to real-world object-centric learning. *arXiv preprint arXiv:2209.14860*, 2022.

[8] Emilio Salinas and Peter Thier. Gain modulation: a major computational principle of the central nervous system. *Neuron*, 27(1):15–21, 2000.

[9] OpenAI. o3 model documentation, 2025. Accessed: 2025-08-22.

[10] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[11] Hao Zheng, Hui Lin, Rui Zhao, and Luping Shi. Dance of snn and ann: solving binding problem by combining spike timing and reconstructive attention. *Advances in Neural Information Processing Systems*, 35:31430–31443, 2022.

[12] Krishnakant Singh, Simone Schaub-Meyer, and Stefan Roth. Guided latent slot diffusion for object-centric learning. *arXiv preprint arXiv:2407.17929*, 2024.

[13] Qian Tang, Hao Wang, Xiaofeng Zhu, Zhen Lei, and Zheng Zhang. Rola: real-world object-centric learning with attention optimization. *Science China Information Sciences*, 68(9):192105, 2025.

[14] Ke Fan, Zechen Bai, Tianjun Xiao, Tong He, Max Horn, Yanwei Fu, Francesco Locatello, and Zheng Zhang. Adaptive slot attention: Object discovery with dynamic slot number. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23062–23071, 2024.

[15] Avinash Kori, Francesco Locatello, Ainkaran Santhirasekaram, Francesca Toni, Ben Glocker, and Fabio De Sousa Ribeiro. Identifiable object-centric representation learning via probabilistic slot attention. *Advances in Neural Information Processing Systems*, 37:93300–93335, 2024.

7 Appendix

A Statistical Comparisons

A.1 MSE (Mean Squared Error) Comparisons

Table 2: Pairwise MSE statistical comparisons between MVBA variants

Comparison	Mean Diff	t stat	p value	Cohen's d	Effect Size	Improv %	95% CI
Architecture-Only Comparisons (Fixed $\alpha = 1$**)**							
baseline to *full_fixed*	0.7628	103.87	$< 10^{-100}$	2.323	Huge	65.00	$[64.38, 65.59]$
baseline to *feature_fixed*	0.6298	93.13	$< 10^{-100}$	2.083	Huge	53.66	$[52.89, 54.37]$
baseline to *spatial_fixed*	-0.0022	-6.27	4.49×10^{-10}	-0.14	Negligible	-0.19	$[-0.25, -0.13]$
spatial_fixed to *full_fixed*	0.765	104.1	$< 10^{-100}$	2.328	Huge	65.06	$[64.49, 65.67]$
spatial_fixed to *feature_fixed*	0.632	93.19	$< 10^{-100}$	2.084	Huge	53.75	$[53.05, 54.42]$
feature_fixed to *full_fixed*	0.133	34	$< 10^{-100}$	0.76	Medium	24.46	$[23.28, 25.72]$
BBRE Enhancement Effects							
full_fixed to *full*	0.0755	22.07	8.15×10^{-97}	0.494	Small-Med	18.38	$[17.06, 19.79]$
feature_fixed to *feature*	0.1014	30.94	$< 10^{-100}$	0.692	Medium	18.65	$[17.65, 19.69]$
spatial_fixed to *spatial*	0.5755	89.24	$< 10^{-100}$	1.996	Very Large	48.95	$[48.29, 49.68]$
Enhanced Model Comparisons ($\alpha \in [1, 3]$)							
baseline to *full*	0.8383	105.8	$< 10^{-100}$	2.366	Huge	71.43	$[71.00, 71.87]$
baseline to *feature*	0.7312	99.58	$< 10^{-100}$	2.227	Huge	62.30	$[61.69, 62.90]$
baseline to *spatial*	0.5733	89.19	$< 10^{-100}$	1.995	Very Large	48.85	$[48.15, 49.53]$
spatial to *full*	0.265	60.39	$< 10^{-100}$	1.351	Very Large	44.15	$[43.29, 45.00]$
spatial to *feature*	0.1579	43.66	$< 10^{-100}$	0.976	Large	26.30	$[25.35, 27.27]$
feature to *full*	0.1071	36.27	$< 10^{-100}$	0.811	Large	24.21	$[23.28, 25.16]$
Cross-Type Comparisons							
full_fixed to *feature*	-0.0316	-9.37	1.93×10^{-20}	-0.21	Small	-7.69	$[-9.42, -6.09]$
full_fixed to *spatial*	-0.1895	-46.28	$< 10^{-100}$	-1.035	Large	-46.13	$[-48.41, -44.08]$
feature_fixed to *full*	0.2085	50.84	$< 10^{-100}$	1.137	Large	38.34	$[37.49, 39.24]$
feature_fixed to *spatial*	-0.0565	-15.83	3.01×10^{-53}	-0.354	Small	-10.39	$[-11.68, -9.11]$
spatial_fixed to *full*	0.8405	105.88	$< 10^{-100}$	2.368	Huge	71.48	$[71.08, 71.94]$
spatial_fixed to *feature*	0.7334	99.69	$< 10^{-100}$	2.23	Huge	62.37	$[61.74, 62.99]$

A.2 PSNR (Peak Signal-to-Noise Ratio) Comparisons

Table 3: Pairwise PSNR statistical comparisons between MVBA variants

Comparison	Mean Diff	t stat	p value	Cohen's d	Effect Size	Improv %	95% CI
Architecture-Only Comparisons (Fixed $\alpha = 1$)							
baseline to *full_fixed*	-4.9327	-131.51	$< 10^{-100}$	-2.941	Huge	1546.23	$[1244.4, 2107.1]$
baseline to *feature_fixed*	-3.634	-106.08	$< 10^{-100}$	-2.373	Huge	1139.13	$[886.5, 1540.9]$
baseline to *spatial_fixed*	0.0081	7.08	1.94×10^{-12}	0.158	Negligible	-2.52	$[-3.6, -1.7]$
spatial_fixed to *full_fixed*	-4.9408	-131.79	$< 10^{-100}$	-2.948	Huge	1510.62	$[1215, 2009.7]$
spatial_fixed to *feature_fixed*	-3.6421	-106.06	$< 10^{-100}$	-2.372	Huge	1113.54	$[888.9, 1475.4]$
feature_fixed to *full_fixed*	-1.2987	-42.48	$< 10^{-100}$	-0.95	Large	39.18	$[36.9, 41.5]$
BBRE Enhancement Effects							
full_fixed to *full*	-0.7483	-23.93	$< 10^{-100}$	-0.535	Medium	16.22	$[14.7, 17.8]$
feature_fixed to *feature*	-0.9093	-39.09	$< 10^{-100}$	-0.874	Large	27.43	$[25.7, 29.3]$
spatial_fixed to *spatial*	-3.1375	-101.19	$< 10^{-100}$	-2.263	Huge	959.29	$[760.5, 1302.4]$
Enhanced Model Comparisons ($\alpha \in [1, 3]$)							
baseline to *full*	-5.6811	-170.64	$< 10^{-100}$	-3.817	Huge	1780.81	$[1430.7, 2391.5]$
baseline to *feature*	-4.5433	-127.29	$< 10^{-100}$	-2.847	Huge	1424.17	$[1142, 1915.6]$
baseline to *spatial*	-3.1295	-101.19	$< 10^{-100}$	-2.263	Huge	980.98	$[777, 1342.7]$
spatial to *full*	-2.5516	-87.91	$< 10^{-100}$	-1.966	Very Large	90.79	$[86.4, 95.2]$
spatial to *feature*	-1.4138	-55	$< 10^{-100}$	-1.23	Very Large	50.31	$[47.7, 52.8]$
feature to *full*	-1.1377	-47.66	$< 10^{-100}$	-1.066	Large	26.93	$[25.5, 28.5]$
Cross-Type Comparisons							
full_fixed to *feature*	0.3894	12.67	1.90×10^{-35}	0.283	Small	-8.44	$[-9.7, -7.3]$
full_fixed to *spatial*	1.8032	55.67	$< 10^{-100}$	1.245	Very Large	-39.08	$[-40.3, -37.8]$
feature_fixed to *full*	-2.0471	-76.63	$< 10^{-100}$	-1.714	Very Large	61.75	$[59.1, 64.5]$
feature_fixed to *spatial*	0.5045	20.23	6.36×10^{-83}	0.452	Small	-15.22	$[-16.5, -13.8]$
spatial_fixed to *full*	-5.6891	-170.63	$< 10^{-100}$	-3.816	Huge	1739.42	$[1392.3, 2300.4]$
spatial_fixed to *feature*	-4.5514	-127.35	$< 10^{-100}$	-2.848	Huge	1391.56	$[1111.1, 1892]$

A.3 SSIM (Structural Similarity Index) Comparisons

Table 4: Pairwise SSIM statistical comparisons between MVBA variants

Comparison	Mean Diff	t stat	p value	Cohen's d	Effect Size	Improv %	95% CI
Architecture-Only Comparisons (Fixed $\alpha = 1$)							
baseline to *full_fixed*	-0.5346	-69.32	$< 10^{-100}$	-1.55	Very Large	618.6	$[587.2, 652.1]$
baseline to *feature_fixed*	-0.5019	-77.31	$< 10^{-100}$	-1.729	Very Large	580.69	$[554.2, 609.7]$
baseline to *spatial_fixed*	0.0179	12.3	1.39×10^{-33}	0.275	Small	-20.71	$[-23.7, -17.5]$
spatial_fixed to *full_fixed*	-0.5525	-73.64	$< 10^{-100}$	-1.647	Very Large	806.33	$[765.8, 851.5]$
spatial_fixed to *feature_fixed*	-0.5198	-79.13	$< 10^{-100}$	-1.77	Very Large	758.52	$[718.5, 805]$
feature_fixed to *full_fixed*	-0.0328	-4.02	6.04×10^{-5}	-0.09	Negligible	5.57	$[2.9, 8.3]$
BBRE Enhancement Effects							
full_fixed to *full*	-0.0931	-11.91	1.19×10^{-31}	-0.266	Small	14.99	$[12.4, 17.7]$
feature_fixed to *feature*	-0.0699	-9.82	2.94×10^{-22}	-0.22	Small	11.88	$[9.4, 14.4]$
spatial_fixed to *spatial*	-0.4159	-54.06	$< 10^{-100}$	-1.209	Very Large	606.91	$[574.1, 647.8]$
Enhanced Model Comparisons ($\alpha \in [1, 3]$)							
baseline to *full*	-0.6277	-98.34	$< 10^{-100}$	-2.2	Huge	726.29	$[695.2, 762.7]$
baseline to *feature*	-0.5717	-88.45	$< 10^{-100}$	-1.978	Very Large	661.56	$[634.8, 690.8]$
baseline to *spatial*	-0.398	-51.1	$< 10^{-100}$	-1.143	Large	460.48	$[435.1, 486.1]$
spatial to *full*	-0.2297	-27.33	$< 10^{-100}$	-0.611	Medium	47.42	$[42.8, 52.1]$
spatial to *feature*	-0.1738	-21.33	3.82×10^{-91}	-0.477	Small	35.88	$[31.9, 39.9]$
feature to *full*	-0.0559	-8.24	3.11×10^{-16}	-0.184	Negligible	8.5	$[6.4, 10.7]$
Cross-Type Comparisons							
full_fixed to *feature*	-0.0371	-4.73	2.42×10^{-6}	-0.106	Negligible	5.98	$[3.5, 8.7]$
full_fixed to *spatial*	0.1366	16.16	2.52×10^{-55}	0.362	Small	-22	$[-24.3, -19.6]$
feature_fixed to *full*	-0.1258	-17.86	2.57×10^{-66}	-0.399	Small	21.39	$[18.7, 24.3]$
feature_fixed to *spatial*	0.1039	12.65	2.47×10^{-35}	0.283	Small	-17.66	$[-20.4, -15.1]$
spatial_fixed to *full*	-0.6456	-101.59	$< 10^{-100}$	-2.272	Huge	942.15	$[893.1, 997.5]$
spatial_fixed to *feature*	-0.5896	-90.57	$< 10^{-100}$	-2.026	Huge	860.52	$[816.5, 908.1]$

A.4 L1 (Mean Absolute Error) Comparisons

Table 5: Pairwise L1 statistical comparisons between MVBA variants

Comparison	Mean Diff	t stat	p value	Cohen's d	Effect Size	Improv %	95% CI
Architecture-Only Comparisons (Fixed $\alpha = 1$)							
baseline to *full_fixed*	0.3533	119.5	$< 10^{-100}$	2.673	Huge	62.37	$[62, 62.7]$
baseline to *feature_fixed*	0.3001	109.63	$< 10^{-100}$	2.452	Huge	52.99	$[52.5, 53.4]$
baseline to *spatial_fixed*	0.0307	50.87	$< 10^{-100}$	1.138	Large	5.41	$[5.2, 5.6]$
spatial_fixed to *full_fixed*	0.3226	116.07	$< 10^{-100}$	2.596	Huge	60.22	$[59.8, 60.6]$
spatial_fixed to *feature_fixed*	0.2695	104.29	$< 10^{-100}$	2.333	Huge	50.3	$[49.8, 50.8]$
feature_fixed to *full_fixed*	0.0531	47.5	$< 10^{-100}$	1.062	Large	19.95	$[19.3, 20.6]$
BBRE Enhancement Effects							
full_fixed to *full*	0.0229	22.1	4.84×10^{-97}	0.494	Small	10.75	$[9.9, 11.6]$
feature_fixed to *feature*	0.0444	46.72	$< 10^{-100}$	1.045	Large	16.66	$[16.1, 17.3]$
spatial_fixed to *spatial*	0.2239	90.71	$< 10^{-100}$	2.029	Huge	41.8	$[41.3, 42.3]$
Enhanced Model Comparisons ($\alpha \in [1, 3]$)							
baseline to *full*	0.3762	114.44	$< 10^{-100}$	2.56	Huge	66.42	$[66.1, 66.7]$
baseline to *feature*	0.3445	114.8	$< 10^{-100}$	2.568	Huge	60.83	$[60.4, 61.2]$
baseline to *spatial*	0.2546	97.68	$< 10^{-100}$	2.185	Huge	44.95	$[44.5, 45.4]$
spatial to *full*	0.1216	84.81	$< 10^{-100}$	1.897	Very Large	39	$[38.5, 39.5]$
spatial to *feature*	0.0899	79.22	$< 10^{-100}$	1.772	Very Large	28.84	$[28.3, 29.4]$
feature to *full*	0.0317	37.38	$< 10^{-100}$	0.836	Large	14.28	$[13.7, 14.9]$
Cross-Type Comparisons							
full_fixed to *feature*	-0.0088	-9.3	3.72×10^{-20}	-0.208	Small	-4.11	$[-5, -3.3]$
full_fixed to *spatial*	-0.0987	-78.46	$< 10^{-100}$	-1.755	Very Large	-46.29	$[-47.6, -45]$
feature_fixed to *full*	0.076	60.11	$< 10^{-100}$	1.344	Very Large	28.56	$[28, 29.1]$
feature_fixed to *spatial*	-0.0455	-42.58	$< 10^{-100}$	-0.952	Large	-17.11	$[-18, -16.2]$
spatial_fixed to *full*	0.3455	110.85	$< 10^{-100}$	2.479	Huge	64.5	$[64.1, 64.8]$
spatial_fixed to *feature*	0.3138	110.54	$< 10^{-100}$	2.472	Huge	58.58	$[58.1, 59]$

B Code Availability

https://anonymous.4open.science/r/MVBA-1A7D

1. **MODEL IMPLEMENTATION**: src/models/
2. **DATASET**: data/simple_objects/
3. **TRAINING INFRASTRUCTURE**: src/ & scripts/training/
4. **RAW TRAINING OUTPUT**: train_recon_only/
5. **EXPERIMENT PIPELINE**: experiments_recon_quality_ablation/

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes] .

   Justification: The abstract and introduction claim that MVBA, as a proof-of-concept, tests BBRE as a contrast mechanism via Smooth Power-Law to improve reconstruction quality in ablations on multi-object scenes. This aligns with the methods (architecture, variants, dataset) and results (MSE improvements, statistical tests) presented in the paper. The scope is limited to a toy dataset and indirect metrics, as stated, without overgeneralizing to real-world applications.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes] .

   Justification: The paper explicitly discusses limitations in the "**Conclusion**" Section: (1) the toy nature of the SimpleObjects dataset restricting generalizability and (2) reconstruction metrics as indirect proxy measures for binding.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA] .

Justification: The paper focuses on a computational model and empirical ablations rather than formal theoretical proofs. The Smooth Power-Law is presented as a functional analog with implementation details, but no theorems or mathematical derivations requiring assumptions and proofs are claimed.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes] .

Justification: The methods detail the architecture (components, hyperparameters like (n_slots = 4, $\alpha \in [1, 3]$), variants, dataset (SimpleObjects with 8,000 train / 2,000 test), training (MSE loss, Adam, 50 epochs, batch=16), metrics (MSE, PSNR, SSIM, L1), and analysis (t-tests, Cohen's d, bootstrap), sufficient to replicate the ablation results and claims about improvements.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
    (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
    (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes] .

Justification: See Appendix B.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes] .

Justification: The Methods section details data splits (8,000 train/ 2,000 test), hyperparameters were chosen to be minimally complex (n_slots = 4, $\alpha \in [1, 3]$, iterations = 3), optimizer (Adam), loss (MSE), epochs (50), batch size (16), and evaluation metrics (MSE, L1, PSNR, SSIM) use standard formulas.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes] .

Justification: Results report means with stds (Table 1), of the primary (MSE) and secondary (PSNR, SSIM, L1) metrics averaged over 2000 test samples. $p < 0.001$ for all. Appendix A details a complete statistical analysis.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes] .

   Justification: Training uses a single 16GB GPU, batch size 16, 50 epochs. The 7 model variants share setup, with similar execution times (about 15 mins per variant) in a PyTorch environment on 16 GB GPU hardware.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes] .

   Justification: The research uses a synthetic dataset and ensures transparency through detailed methods and reproducibility via released code. This poses no societal harms or violations of human rights.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No] .

    Justification: Positive impacts (advancing biologically plausible AI for visual perception) are mentioned, but negative impacts are not discussed. The paper focuses on foundational research without exploring harms.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification: The paper uses a synthetic dataset and proof-of-concept model with no high-risk elements for misuse, so safeguards are not applicable.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA] .

Justification: No external assets like code, datasets, or models are used. The SimpleObjects dataset is synthetically generated by an LLM; and our model is inspired by previous work on Slot Attention, which is cited without direct reuse requiring licenses, and dual-stream perceptual theories.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes] .

Justification: The paper introduces the MVBA model and SimpleObjects dataset, documented in 2.Methods, with code publicly available (Appendix B). No external licensing needed as assets are original/synthetic.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA] .

Justification: The paper involves no crowdsourcing or human subjects research, only synthetic data and computational simulations.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification: No human subjects are involved. The research is purely computational on synthetic datasets.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes] .

    Justification: The paper discloses LLM (OpenAI's o3) usage for generating the SimpleObjects dataset in methods, which supports experiments but is not core to the method's originality.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.