Towards Visual Simulation in Multimodal Language Models

Catherine Finegan-Dollak
Department of Computer Science
University of Richmond
Richmond, Virginia, USA
cfinegan@richmond.edu

Abstract

Inspired by extensive evidence that humans use *visual simulation* to understand parts of language, this position paper explores analogous behavior in multimodal language models. Although grounding models in vision has been an active area of research in the AI community, visual simulation remains underexplored. We address this gap by formally defining visual simulation, identifying the architectural and training components to enable it, and proposing a multi-pronged approach to determine whether a model engages in visual simulation.

1 Introduction

When humans hear or read a sentence, our brains simulate the objects and events from that sentence: our visual cortex activates in patterns similar to those that would occur if we actually saw the event, auditory cortex activates as though we heard it, motor cortex activates as though we were taking the actions described, and so on [Bergen, 2015]. Cognitive scientists theorize that such activations—sometimes called *embodied simulation* or *grounded cognition* [Barsalou, 2008, Reilly et al., 2025]—are part of how human brains represent the meaning conveyed by language.

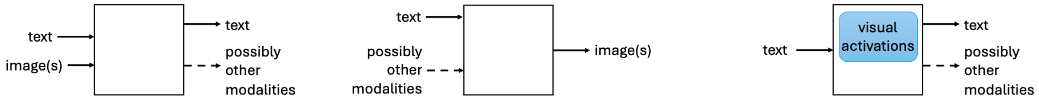
Grounding language models has been an extremely active area of research in recent years [e.g., Suhr et al., 2019, Bisk et al., 2020, Thrush et al., 2022, Li et al., 2022, among others]. To date, however, embodied simulation remains an underexplored facet of grounding language models.[1] This position paper is a first step towards addressing this gap. While a full solution would involve many forms of simulation, we focus on *visual simulation*.

Our contributions are

- A definition of visual simulation for multimodal models (Section 2)
- Identification of architectural and training elements that enable visual simulation with brief analysis of how current models do/do not incorporate these elements (Section 3)
- Proposals for how to determine if a model engages in visual simulation (Section 4)

[1] We use "grounding" to refer broadly to connecting linguistic representations with perceptual or action-based ones. This is related to, but distinct from, grounded *understanding*, which would require demonstrating that language behavior depends on perceptual mechanisms. We do not contend that visual simulation constitutes grounded understanding, but propose it as a useful lens for examining how multimodal models might approach grounding.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: First Workshop on CogInterp: Interpreting Cognition in Deep Learning Models.


text
image(s)
text
possibly other modalities
text
possibly other modalities
image(s)
text
visual activations
text
possibly other modalities

(a) Typical VLM (b) Image generation model (c) Visual simulation

Figure 1: Architectural comparison. (a) Typical VLM jointly processes text and images. (b) Image generation models map text into images. (c) In visual simulation, text input without images produces visual activations as part of model state at inference time.

2 Background: Defining Visual Simulation

To define visual simulation for multimodal language models (MLLMs), we should first understand what it means in humans. Brain imaging studies show that human brain regions involved in visual perception are activated in response to visual concepts expressed in words [Kiefer and Pulvermüller, 2012]; for example, Simmons et al. [2007] showed activity in the left fusiform gyrus for both color perception and color knowledge retrieval based only on words. Behavioral studies provide additional evidence of this phenomenon and point to numerous visual features that are subject to simulation, such as orientation, shape, location, direction of motion, and perspective [Bergen, 2012, 2015].[2]

By analogy, then, we define visual simulation in an MLLM in three parts: (1) activation of portions of the model involved in visual perception (2) to match the activations that would occur if the described object/event were seen, but (3) caused by linguistic, rather than visual, input. We use the term "visual simulation" rather than "visualization" to avoid implying any subjective experience of "seeing."

3 Implications for Language Models

This simple definition implies several desiderata for architecture and training in order for a model to engage in visual simulation.

3.1 Architecture

Visual simulation places specific demands on model architecture. Part 1 of the definition tells us that to engage in visual simulation, the model must have some component capable of visual perception. However, part 3 specifies that at the time a model is engaging in visual simulation, it is not perceiving visual input; the simulation must be triggered by the linguistic input, not by visual input, or it is visual *perception*, not visual *simulation*. That is, at test time, the model should receive text, *not* images. As most vision and language models (VLMs) are designed to take images as part of their input (Figure 1a), they are not engaging in visual simulation. Naturally, this does not rule out the use of images during training. Moreover, a model where image input is optional, such as an any-to-any model [Tang et al., 2023, Wu et al., 2024, Chameleon, 2025], might incorporate a visual simulation capability; however, it is essential to remember that when the image is provided, the model is not engaging in visual simulation.

Visual simulation also does *not* require image generation (Figure 1b). While visual simulation may, in humans, cause a subjective experience of an image, no image is actually created, only visual cortex activations. While generation, unlike input images, is permissible, it is neither necessary nor sufficient for visual simulation.

Having said what the architecture is *not*, what *does* qualify as visual simulation? As shown in Figure 1c, text input causes some component within the model to recreate visual activations. There is, however, room for debate over exactly how to define recreating visual activations.

If we understand recreating visual activations as broadly as possible, we can say that models that pull the text representation towards the vision embedding space are capable of some degree of visual

[2] This is distinct from Battaglia et al. [2013]'s concept of mental simulation as a kind of physics engine used to make predictions about a scene. Their work proposes a mechanism to explain known functions; in contrast, embodied simulation describes observed phenomena in brains, the function of which remains an open question [*e.g.,* Ghio and Tettamanti, 2016, Bechtold et al., 2023].

Table 1: Architectural desiderata for visual simulation across major MLLMs.

Model	Text-only input option	Aligns text representation to vision	Deep alignment with vision encoder	Handles Video
CLIP[1]	Yes	Yes	No	No
BLIP[2]	Yes	Yes	No	No
BLIP-2[3]	No	No (frozen)	No	No
Flamingo[4]	No	No (frozen)	No	Yes
Kosmos-1[5]	Yes	Yes	No	No
LLaVA[6]	No	No (projects vision to text space)	No	No
NExT-GPT[7]	Yes	No (frozen)	No	Yes

[1]Radford et al. [2021]; [2]Li et al. [2022]; [3]Li et al. [2023]; [4]Alayrac et al. [2022];

[5]Huang et al. [2023]; [6]Liu et al. [2023]; [7]Wu et al. [2024];

simulation. For example, CLIP [Radford et al., 2021] trains its text and vision encoders to maximize the similarity between the embeddings for matched text/image pairs and minimize that between unmatched pairs; this simultaneously pulls the text representation towards the vision space and the vision representation towards the text space. Empirically, Jones et al. [2024] demonstrated with behavioral measures developed to study embodied simulation in humans that CLIP's text encoder evidences some visual simulation, though far less than in humans. In its unimodal encoder form, BLIP Li et al. [2022] similarly pulls text representations and image representations closer together.

Stronger forms of visual simulation may require deeper alignment than the single embedding alignment in CLIP. Vision models have many layers, so multiple layers of the text encoder mirroring multiple layers of a vision encoder should be explored. One approach would be an architecture similar to deeply-supervised knowledge distillation Luo et al. [2023].

Finally, visual simulation in models may benefit from extending beyond static representations. Simulating a static image may support certain types of spatial reasoning or implicit world knowledge (e.g., understanding why a polar bear might cover its nose while hunting in snow [Bergen, 2012]). However, human visual simulation also involves dynamic components, such as eye movements that track described events or the mental rotation of objects. Incorporating such dynamics in model architectures could enable richer forms of reasoning, particularly for tasks that require understanding temporal sequences or transformations in space.

In summary, at a minimum, basic visual simulation requires a model that (1) takes text input without an image input, and (2) learns to represent its text input in a way that aligns with a visual encoder. More sophisticated visual simulation should include a deeper alignment of text representations to visual, rather than just one embedding for each modality. Dynamic vision, as opposed to static, offers additional advantages. Table 1 provides a brief overview of how existing models align with these architectural desiderata for visual simulation. CLIP and any-to-any models show promise, but it is clear from this table that existing models do not prioritize visual simulation.

3.2 Training

Having considered structural requirements, we now turn to training signals that might enable simulation. Existing training objectives and datasets were not designed for visual simulation. To achieve visual simulation, training must cause the model to mimic the activations that a vision component would have if it saw the object or event described by the text input. This has implications for three aspects of training: model freezing, data, and tasks.

A frozen LM cannot adapt its representations to recreate visual activations. Models like BLIP-2 [Li et al., 2023] and Flamingo [Alayrac et al., 2022] deliberately freeze the language model, preserving its pretrained text space rather than pulling it toward vision. Under our definition, such models cannot learn visual simulation because the LM is never moved toward a visual space. In contrast, freezing the vision encoder is permissible: the text encoder can still be trained to approximate its representations. While this may reduce training costs, it is not required, nor does it align with the human case, where the visual cortex is influenced by language and remains plastic over time [Lupyan et al., 2020, Castaldi et al., 2020, Rosa et al., 2013].

On the data side, if we want a model to learn to recreate the activations a visual component would have if it saw the same thing the text describes, we should train on aligned images and descriptive text. This means images and their captions are not ideal, nor are interleaved text and images: captions complement the images or explain how they relate to a larger text, rather than describing the images. Quality alt text is better, when available. As noted in our discussion of architecture, dynamic visual simulation has advantages over static, so pairs of videos and their descriptions are highly desirable.

Tasks should, at a minimum, draw the encoding of the text into a visual encoding space; as described in the discussion of architecture, alignment at multiple levels is more desirable than alignment of a single embedding. Many tasks on which MLLMs are most commonly evaluated—VQA, captioning, and so on—are not suitable training tasks, as they lead to models that expect visual input, rather than emulating it.

Image generation models such as latent diffusion models [Rombach et al., 2022] could conceivably pull text representations into an image-related space, provided the language model is not frozen. Predicting masked portions of images [e.g., Chen et al., 2020] might work similarly. However, in both cases, the desired updates to the vision encoder would be at most a side effect; we contend that tasks more targeted to visual simulation would be more effective for training.

CLIP's contrastive learning has demonstrated value; it might be expanded to draw intermediate states of multiple layers of the text encoder closer to corresponding layers of the image encoder. A task that has not, to our knowledge, been explored is to directly predict the activations of a vision encoder given the text; again, this suggests something that looks like deep model distillation.

Overall, these directions suggest that successful training for visual simulation will require datasets with genuinely descriptive text and tasks that align text encodings with vision activations across multiple layers, moving beyond objectives designed for multimodal perception.

An additional consideration, which we leave for future work, is the choice of the vision encoder itself. A vision model trained solely for object recognition on static images may not provide the most useful representational space for visual simulation. Identifying which pretraining tasks and datasets yield visual features most conducive to simulation remains an open question.

4 Detecting Visual Simulation

In humans, the evidence for embodied simulation can be grouped into mechanistic and behavioral. Mechanistic evidence includes things like fMRI evidence of activation of the visual cortex; behavioral evidence includes things like eye movement over a blank screen tracking the movement described in a story. We propose an analogous two pronged approach for models.

On the behavioral side, there are two types of experiment we might pursue. The first, pioneered by Jones et al. [2024], adapts the same behavioral experiments used to find evidence of embodied simulation in humans to a machine context.

The second, which is more familiar to the machine learning community, is to create benchmarks for tasks requiring visual simulation. However, such benchmarks are not sufficient by themselves, as there may be alternative explanations for models' success. Consider the following example from the spatial reasoning component of EWoK [Ivanova et al., 2024]: given the input "The nail is east of the wheel," the model must determine which is more likely, "The wheel is west of the nail," or "The wheel is east of the nail." Visualization is one way to solve this problem, but a model with a symbolic relation "east-of" that can be inverted using a rule could make a correct prediction without visualization. Text-only models can learn such a relation—or even a more sophisticated chained relation (e.g., "The wheel is west of the nail, and the cup is east of the nail; is the cup east of the wheel?")—from a sufficiently large corpus. Indeed, Patel and Pavlick [2021] demonstrated that a large, language-only model could learn to answer simple questions about spatial relations in a grid world depicted in text. Thus, benchmarks should be used in conjunction with other types of evidence, not by themselves, to detect visual simulation.

On the mechanistic side, the goal is to measure whether text-induced activations resemble vision activations. If the architecture explicitly predicts the visual encoder's hidden states, this is straightforward, as we can compare the text encoder's predictions with the actual hidden states induced by the corresponding visual stimulus, using similarity measures such as KL divergence. Additional analyses of how much attention subsequent layers pay to these components could also be informative. Mechanistic measures are more complex for other architectures; however, the probing and interpretability literature for LLMs and MLLMs provides a rapidly expanding selection of relevant tools. Causal mediation analysis (CMA) such as ROME [Zhou et al., 2023] and NOTICE [Golovanevsky et al., 2025] can be used to identify causal pathways ("circuits") involved in completing visualization benchmarks. These circuits could likewise be compared against a visual encoder; circuits that align with vision features such as edges would provide strong evidence of visual simulation. Identifying circuits that support performance across multiple visual-simulation-related tasks would also provide mechanistic evidence of visual simulation. Additional methods of understanding such circuits remains an open challenge worth further study.

Thus, combining behavioral measures of both forms with mechanistic probing can build converging evidence of visual simulation. This distinction echoes recent thinking in the long-running "imagery debate" in cognitive science, where behavioral results alone can often be explained in purely symbolic terms, but mechanistic evidence provides stronger support for the existence of genuinely visual representations Pearson and Kosslyn [2015].

5 Conclusion

Visual simulation remains an underexplored but crucial dimension of grounding in multimodal LLMs. By outlining how it might be defined, enabled, and detected, this paper aims not to settle the question but to open it: to invite systematic study of visual simulation as a phenomenon and to lay groundwork for models and methods that can capture it.

Acknowledgments and Disclosure of Funding

We thank Arryn Robbins for serving as a sounding board during early development of this paper and the anonymous reviewers for their insightful feedback. We gratefully acknowledge support from the University of Richmond Arts & Sciences Faculty Research Committee Summer Fellowship, which provided funding that made this work possible.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millicah, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 23716–23736, Red Hook, NY, USA, November 2022. Curran Associates Inc. ISBN 978-1-71387-108-8.

Lawrence W. Barsalou. Grounded Cognition. *Annual Review of Psychology*, 59(Volume 59, 2008): 617–645, January 2008. ISSN 0066-4308, 1545-2085. doi: 10.1146/annurev.psych.59.103006.093639. URL `https://www.annualreviews.org/content/journals/10.1146/annurev.psych.59.103006.093639`. Publisher: Annual Reviews.

Peter W. Battaglia, Jessica B. Hamrick, and Joshua B. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, November 2013. doi: 10.1073/pnas.1306572110. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1306572110`. Publisher: Proceedings of the National Academy of Sciences.

Laura Bechtold, Samuel H. Cosper, Anastasia Malyshevskaya, Maria Montefinese, Piermatteo Morucci, Valentina Niccolai, Claudia Repetto, Ana Zappa, and Yury Shtyrov. Brain Signatures of Embodied Semantics and Language: A Consensus Paper. *Journal of Cognition*, 6(1):61, 2023. ISSN 2514-4820. doi: 10.5334/joc.237. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10573703/`.

Benjamin Bergen. Embodiment, simulation and meaning. In *The Routledge Handbook of Semantics*. Routledge, 2015. ISBN 978-1-315-68553-3. Num Pages: 16.

Benjamin K. Bergen. *Louder Than Words: The New Science of How the Mind Makes Meaning*. Basic Books, October 2012. ISBN 978-0-465-02829-0. Google-Books-ID: BrEszJ1fYGQC.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience Grounds Language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.703. URL https://aclanthology.org/2020.emnlp-main.703.

Elisa Castaldi, Claudia Lunghi, and Maria Concetta Morrone. Neuroplasticity in adult human visual cortex. *Neuroscience & Biobehavioral Reviews*, 112:542–552, May 2020. ISSN 0149-7634. doi: 10.1016/j.neubiorev.2020.02.028. URL https://www.sciencedirect.com/science/article/pii/S0149763419303288.

Team Chameleon. Chameleon: Mixed-Modal Early-Fusion Foundation Models, March 2025. URL http://arxiv.org/abs/2405.09818. arXiv:2405.09818 [cs].

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: UNiversal Image-TExt Representation Learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 104–120, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58577-8. doi: 10.1007/978-3-030-58577-8_7.

Marta Ghio and Marco Tettamanti. Chapter 52 - Grounding Sentence Processing in the Sensory-Motor System. In Gregory Hickok and Steven L. Small, editors, *Neurobiology of Language*, pages 647–657. Academic Press, San Diego, January 2016. ISBN 978-0-12-407794-2. doi: 10.1016/B978-0-12-407794-2.00052-3. URL https://www.sciencedirect.com/science/article/pii/B9780124077942000523.

Michal Golovanevsky, William Rudman, Vedant Palit, Carsten Eickhoff, and Ritambhara Singh. What Do VLMs NOTICE? A Mechanistic Interpretability Pipeline for Gaussian-Noise-free Text-Image Corruption and Evaluation. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11462–11482, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. URL https://aclanthology.org/2025.naacl-long.571/.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Nils Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language Is Not All You Need: Aligning Perception with Language Models. *Advances in Neural Information Processing Systems*, 36:72096–72109, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e425b75bac5742a008d643826428787c-Abstract-Conference.html.

Anna A. Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H. Clark, Carina Kauf, Jennifer Hu, R. T. Pramod, Gabriel Grand, Vivian Paulun, Maria Ryskina, Ekin Akyürek, Ethan Wilcox, Nafisa Rashid, Leshem Choshen, Roger Levy, Evelina Fedorenko, Joshua Tenenbaum, and Jacob Andreas. Elements of World Knowledge (EWOK): A cognition-inspired framework for evaluating basic world knowledge in language models, May 2024. URL http://arxiv.org/abs/2405.09605. arXiv:2405.09605 [cs].

Cameron R. Jones, Benjamin Bergen, and Sean Trott. Do Multimodal Large Language Models and Humans Ground Language Similarly? *Computational Linguistics*, 50(4):1415–1440, December 2024. ISSN 0891-2017. doi: 10.1162/coli_a_00531.

Markus Kiefer and Friedemann Pulvermüller. Conceptual representations in mind and brain: Theoretical developments, current evidence and future directions. *Cortex*, 48(7):805–825, July 2012. ISSN 0010-9452. doi: 10.1016/j.cortex.2011.04.006. URL https://www.sciencedirect.com/science/article/pii/S0010945211001018.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12888–12900. PMLR, June 2022. URL https://proceedings.mlr.press/v162/li22n.html. ISSN: 2640-3498.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742. PMLR, July 2023. URL https://proceedings.mlr.press/v202/li23q.html. ISSN: 2640-3498.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, December 2023.

Shiya Luo, Defang Chen, and Can Wang. Knowledge Distillation with Deep Supervision. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, June 2023. doi: 10.1109/IJCNN54540.2023.10191309. URL https://ieeexplore.ieee.org/abstract/document/10191309. ISSN: 2161-4407.

Gary Lupyan, Rasha Abdel Rahman, Lera Boroditsky, and Andy Clark. Effects of Language on Visual Perception. *Trends in Cognitive Sciences*, 24(11):930–944, November 2020. ISSN 1879-307X. doi: 10.1016/j.tics.2020.08.005.

Roma Patel and Ellie Pavlick. Mapping Language Models to Grounded Conceptual Spaces. In *The Tenth International Conference on Learning Representations, ICLR*, October 2021. URL https://openreview.net/forum.

Joel Pearson and Stephen M. Kosslyn. The heterogeneity of mental representation: Ending the imagery debate. *Proceedings of the National Academy of Sciences*, 112(33):10089–10092, August 2015. doi: 10.1073/pnas.1504933112. URL https://www.pnas.org/doi/full/10.1073/pnas.1504933112. Publisher: Proceedings of the National Academy of Sciences.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, July 2021. URL https://proceedings.mlr.press/v139/radford21a.html. ISSN: 2640-3498.

Jamie Reilly, Cory Shain, Valentina Borghesani, Philipp Kuhnke, Gabriella Vigliocco, Jonathan E. Peelle, Bradford Z. Mahon, Laurel J. Buxbaum, Asifa Majid, Marc Brysbaert, Anna M. Borghi, Simon De Deyne, Guy Dove, Liuba Papeo, Penny M. Pexman, David Poeppel, Gary Lupyan, Paulo Boggio, Gregory Hickok, Laura Gwilliams, Leonardo Fernandino, Daniel Mirman, Evangelia G. Chrysikou, Chaleece W. Sandberg, Sebastian J. Crutch, Liina Pylkkänen, Eiling Yee, Rebecca L. Jackson, Jennifer M. Rodd, Marina Bedny, Louise Connell, Markus Kiefer, David Kemmerer, Greig de Zubicaray, Elizabeth Jefferies, Dermot Lynott, Cynthia S.Q. Siew, Rutvik H. Desai, Ken McRae, Michele T. Diaz, Marianna Bolognesi, Evelina Fedorenko, Swathi Kiran, Maria Montefinese, Jeffrey R. Binder, Melvin J. Yap, Gesa Hartwigsen, Jessica Cantlon, Yanchao Bi, Paul Hoffman, Frank E. Garcea, and David Vinson. What we mean when we say semantic: Toward a multidisciplinary semantic glossary. *Psychonomic Bulletin & Review*, 32(1):243–280, February 2025. ISSN 1531-5320. doi: 10.3758/s13423-024-02556-7. URL https://doi.org/10.3758/s13423-024-02556-7.

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. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.

Andreia Martins Rosa, Maria Fátima Silva, Sónia Ferreira, Joaquim Murta, and Miguel Castelo-Branco. Plasticity in the Human Visual Cortex: An Ophthalmology-Based Perspective. *BioMed Research International*, 2013:568354, 2013. ISSN 2314-6133. doi: 10.1155/2013/568354. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3800612/.

W. Kyle Simmons, Vimal Ramjee, Michael S. Beauchamp, Ken McRae, Alex Martin, and Lawrence W. Barsalou. A common neural substrate for perceiving and knowing about color. *Neuropsychologia*, 45(12):2802–2810, September 2007. ISSN 0028-3932. doi: 10.1016/j.neuropsychologia.2007.05.002. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3596878/.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A Corpus for Reasoning about Natural Language Grounded in Photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1644. URL https://www.aclweb.org/anthology/P19-1644.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-Any Generation via Composable Diffusion. In *Advances in Neural Information Processing Systems*, volume 36, pages 16083–16099, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/33edf072fe44f19079d66713a1831550-Abstract-Conference.html.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5228–5238, New Orleans, LA, USA, June 2022. IEEE. ISBN 978-1-66546-946-3. doi: 10.1109/CVPR52688.2022.00517. URL https://ieeexplore.ieee.org/document/9878945/.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NExT-GPT: any-to-any multimodal LLM. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 53366–53397, Vienna, Austria, July 2024. JMLR.org.

Kankan Zhou, Eason Lai, Wei Bin Au Yeong, Kyriakos Mouratidis, and Jing Jiang. ROME: Evaluating Pre-trained Vision-Language Models on Reasoning beyond Visual Common Sense. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10185–10197, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.683. URL https://aclanthology.org/2023.findings-emnlp.683.