LANGUAGE CONDITIONING ELICITS SYMBOL GROUNDING IN VISUAL PROCESSING

Anonymous authors
Paper under double-blind review

ABSTRACT

Grounding is a core ability for visual reasoning in embodied intelligence. Within multimodal models, prior work has identified mechanisms for grounding, such as binding and retrieval, in language models. We posit that language conditioning can also elicit grounding in visual processing. Specifically, we inject linguistic context into a frozen pretrained vision transformer (ViT) via cross-attention, shaping the representations of objects and their features to serve as potential grounders for linguistic entities. We evaluate this approach on visual question answering tasks across vision backbones and analyze the representations they encode. We examine the representations at global and local levels. First, geometric analysis shows that ViTs encode object representations with attribute-related geometric structure, with language conditioning reshaping this hierarchy and substructures according to the queried attribute. Further analysis on geometry across layers reveals two mechanisms for grounding: binding segregates scenes containing the referent in the middle layers; organization repartitions scenes into groups aligned with the queried attribute values for retrieval in the latter layers. Second, controlled experiments on embeddings demonstrate that linguistic context enhances attribute alignment of the referent and its queried attribute while suppressing those of non-referents and unqueried attributes; referent-selective modulation also holds for natural images. This study opens a new direction for studying symbol/language grounding, the one that treats perception itself as an active component of grounding and investigates how linguistic signals can shape its representation.

1 INTRODUCTION

Grounding is a core ability for visual reasoning in embodied intelligence (Pavlick, 2023). For example, given a scene (*e.g.*, Figure 1, left) and a question "*What is the color of the shiny cube?*", we can compose features *shiny* and *cube* into the linguistic entity *shiny cube*, bind that entity to a particular object in the scene, and retrieve its color *red* from the object to produce the answer.

Symbol grounding concerns how linguistic symbols acquire meaning through their connection to perception and the external world (Harnad, 1990). Much of the existing work on vision-language grounding, however, studies this problem primarily from the language side: visual inputs are represented as fixed features extracted by a pretrained frozen vision encoder, and learning focuses on how linguistic representations are aligned with these visual features. This formulation implicitly treats the visual representation as a fixed grounding target (referred to as *grounder*) and assumes that grounding occurs largely through adaptation on the language side. However, this captures only part of the grounding process. The perceptual signals arising from the environment that provide the grounding for linguistic symbols can themselves be shaped by language, task context, and interaction. A more complete account of symbol grounding therefore requires studying the bidirectional relationship between language and perception: not only how language is grounded in visual representations, but also how linguistic signals influence what is perceived, represented, and attended to in the visual world.

We therefore ask: *can language conditioning elicit grounding in visual processing?* Motivated by classical cognitive theories of visual processing and perceptual representation (Treisman & Gelade, 1980; Shepard & Metzler, 1971; Marr, 1982; Biederman, 1987; Barsalou, 1999), we hypothesize that language conditioning organizes visual representations around the referent and the semantic features

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

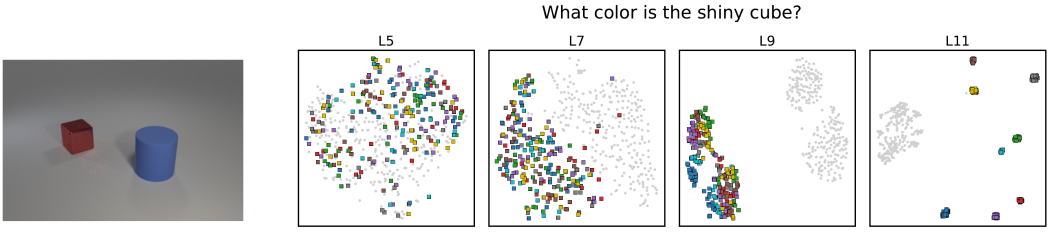


Figure 1: An example scene and t-SNE visualizations of scene representations in DINOv2. Each point represents a scene, obtained by mean-pooling visual patch tokens. Marker size, shape, and color denote the designated object's actual attribute values, respectively: ○ sphere, △ cylinder, and □ cube; darker color shiny, and lighter color matte. Each panel is embedded independently; positions are not comparable across panels. For the question "What color is the shiny cube?", the geometry separates the scenes containing a described object (shiny cube) by L7; by L11, the matching scenes form subgroups by the object's color. Point colors indicate description satisfaction; markers indicate the matching object's color.

specified by the linguistic context. To validate this hypothesis, we extend a typical vision-language architecture by injecting linguistic context into a frozen pretrained vision transformer (ViT) through learnable cross-attention. We examine how language conditioning steers visual representations via cross-attention to serve as potential grounders for linguistic entities.

More specifically, we use CLEVR (Johnson et al., 2017) as a controlled testbed to examine geometric relations of grounding-related visual representations at global and local levels across vision backbones. First, language reorganizes geometry around the referent and its attributes. Across layers, the model progressively separates scenes containing the object satisfying the described attributes, and then forms subgroups according to the attribute being queried, revealing a progression from binding to organization for retrieval. Second, this reorganization is reflected locally at object representations: language enhances the queried attribute of the referent while suppressing corresponding information for non-referents and unqueried attributes. Third, the same referent-selective modulation recurs after training on GQA, showing that the phenomenon extends beyond synthetic scenes to natural images.

In summary, our findings show that language conditioning can elicit grounding within vision encoders, shaping visual representations in ways that reflect the semantic content expressed in language. This perspective opens a new direction for studying symbol/language grounding, the one that treats perception itself as an active component of grounding and investigates how linguistic signals shape its representation.

2 Related work

Symbol processing and grounding The symbol grounding problem concerns how symbols acquire meaning through connections to the world (Harnad, 1990). Prior work has studied how symbol-like mechanisms emerge in neural networks across architectures and modalities. Early studies showed that symbol processing can emerge from architectural inductive biases (Webb et al., 2021; Lepori et al., 2023). More recent mechanistic analyses have localized specific operations, including entity binding (Feng & Steinhardt, 2024; Li et al., 2025) and abstract reasoning (Yang et al., 2025). Within VLMs, related work has identified binding and retrieval (Assouel et al., 2026; Saravanan et al., 2025; Song et al., 2026), and grounding (Wu et al., 2026; Kang et al., 2025). Song et al. (2026) examine how linguistic context recodes visual representation in vision-language models. Although recent work finds object binding in pre-trained vision transformers (Li et al., 2025), a long-held belief in the multimodal-model community is that mechanisms for grounding lie within the language model (Feng & Steinhardt, 2024; Yang et al., 2025; Assouel et al., 2026; Wu et al., 2026; Kang et al., 2025; Song et al., 2026). Our work asks whether such mechanisms also arise when language conditions the visual representation, shifting the analysis from language components to vision ones.

Cross-Modal Integration in VLMs Vision-language models differ in where they fuse modalities. Late-fusion architectures encode images into query-independent visual tokens and route most cross-modal computation through the language model (Liu et al., 2024; Wang et al., 2024). Early-fusion architectures allow one modality to condition the other during representation formation. FiLM applies language-conditioned affine transformations to visual modules (Perez et al., 2018). Flamingo injects visual information into a language model for few-shot generation (Alayrac et al., 2022); Q-former architectures use learned queries to extract visual features for language-image alignment (Li et al., 2022; 2023); SteerViT injects language into vision transformers to steer visual representations (Ruthardt et al., 2026). Cross-attention also supports language-to-vision control in image generation (Chefer et al., 2023) and grounded pre-training (Zhang et al., 2022). We adopt the language-to-vision early-fusion setting, but shift the emphasis from capability to understanding: rather than asking whether cross-attention improves performance, we ask whether and how linguistic context shapes visual representations to support grounding.

3 Symbol Grounding in Visual Processing

Grounding is a broad and often overloaded term, used in a variety of contexts ranging from grounded reasoning to communication and common ground. In this paper, we use the term in the specific sense of symbol grounding in the visual world: establishing correspondences between linguistic entities and entities in visual perception.

We therefore adopt visual question answering tasks as a controlled setting in our study. In visual question answering tasks, given a visual input v and a question q, a model $\mathcal{M}$ maps (v, q) to an answer y: $\mathcal{M} : (v, q) \mapsto y$. Given training data $\mathcal{D} = \{(v, q, a^*)\}$, the model f_θ is trained to maximize the likelihood of the correct answer. Answering visual questions involves connecting linguistic expressions to visual information relevant to the answer. For example, querying the material of the red cube requires identifying the described red cube and retrieving its queried attribute value from the scene. Our key hypothesis is that linguistic context can be used to shape visual representations to support grounding. To validate this hypothesis, we introduce learnable cross-attention into a frozen pretrained vision transformer (ViT).

3.1 Architecture

A vision transformer (ViT) maps an image to patch tokens $z^{(0)} \in \mathbb{R}^{N \times d}$, obtained by linearly projecting image patches, and updates them through L transformer blocks. Each block applies self-attention (SA) and feed-forward multi-layer perceptron (MLP) with residual connections:

$$z^{(\ell)} = z^{(\ell-1)} + \mathrm{SA}^{(\ell)}(z^{(\ell-1)}) + \mathrm{MLP}^{(\ell)}(z^{(\ell-1)}).$$

Unrolling this recursion yields the residual-stream view (Elhage et al., 2021; Ferrando et al., 2024), each layer writes to a shared additive stream, and the representation at layer ℓ accumulates the contributions of preceding layers. This view motivates an analysis of how language conditioning steers the feature geometry. To enable language conditioning, we inject cross-attention (CA) modules with a $\tanh$ gate before self-attention in selected transformer blocks. Each CA module conditions the visual tokens on text tokens z_t and writes the result into the visual residual stream:

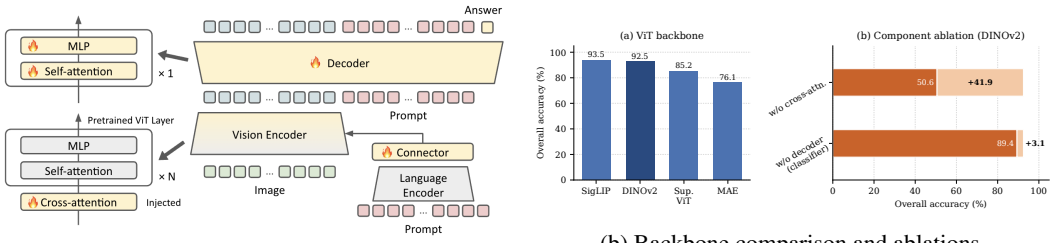
$$\tilde{z}^{(\ell)} = z^{(\ell-1)} + \tanh(\alpha^{(\ell)}) \cdot \mathrm{CA}^{(\ell)}(z^{(\ell-1)}, z_t),$$
$$z^{(\ell)} = \tilde{z}^{(\ell)} + \mathrm{SA}^{(\ell)}(\tilde{z}^{(\ell)}) + \mathrm{MLP}^{(\ell)}(\tilde{z}^{(\ell)}).$$

Here z_t denotes the text-token representations, and $\alpha^{(\ell)}$ is a learned gate initialized near zero.

3.2 Setup

The proposed model comprises a ViT backbone with gated cross-attention, a text encoder with a connector, and a decoder to predict output as illustrated in Figure 2a.

ViT backbone. We build on a pretrained ViT backbone (Oquab et al., 2024) with gated cross-attention appearing at alternating transformer blocks (e.g., layers 1, 3, 5, 7, 9, and 11 for a 12-layer ViT). The choice of gates and CA placement follows prior early-fusion architectures for stability (Alayrac et al., 2022; Ruthardt et al., 2026).



(a) Architecture.

(b) Backbone comparison and ablations.

Figure 2: Architecture and CLEVR validation accuracy. (a) the vision and language encoders are frozen (gray); cross-attention, the connector, and the decoder are trainable (yellow). (b) accuracy (%) averaged over 3 seeds. DINOv2 achieves 92.5%, comparable to SigLIP and higher than supervised ViT and MAE. For DINOv2 ablations, orange denotes the ablated variant, and light orange denotes the gain from restoring the component. Cross-attention adds 41.9 percentage points; replacing the classifier with a decoder adds 3.1 points.

Text encoder. Text queries are encoded by a frozen RoBERTa-Large model (Liu et al., 2019) and projected to the cross-attention via a learned linear connector. The module produces text embeddings z_t fed into CA.

Readout head. We select decoder and classifier heads for evaluation. The decoder head is a one-layer transformer decoder (SA+MLP) where inputs are concatenated visual patches and text tokens. The classifier head is a one-layer MLP which contains 28 CLEVR answer categories.

Training. We adopt CLEVR (Johnson et al., 2017) as a controlled testbed, which allows us to control the generation of scenes and objects through a program. We train on approximately 700K questions over 70K scenes and evaluate on 150K validation questions over 15K held-out scenes across 28 answer classes. The CA modules with gates, connector, and decoder are randomly initialized, while the visual and text backbones remain frozen.

3.3 Validating the Architecture

Before analyzing the model's internals, we verify that the architecture is effective and identify which components drive its performance. We first compare ViT backbones under the same language-conditioned architecture, then ablate its key components.

Backbone comparison. We evaluate four pretrained backbones: DINOv2, SigLIP, MAE, and an ImageNet-21K supervised ViT. As shown in Figure 2b(a), the choice of backbone matters: DINOv2 and SigLIP reach 92.5 and 93.5 overall accuracy, respectively, while the supervised ViT and MAE perform much worse (85.2 and 76.1). Throughout the paper, we use DINOv2 as the default backbone, given its wide adoption in prior work (Jiang et al., 2023; Kim et al., 2024; Krojer et al., 2026) and its strong performance in our setting.

Component ablations. Using DINOv2, we ablate two components: we remove the gated cross-attention so the visual stream receives no language conditioning, and we replace the decoder with a linear classifier. As shown in Figure 2b(b), language conditioning is critical in this architecture: adding cross-attention to the question improves accuracy by 41.9 points. In contrast, the decoder adds only 3.1 points over a linear classifier. Together, these results indicate that the language-conditioned visual representations already encode most of the information required for accurate prediction, with relatively limited additional gains from the output head. Appendix A provides the full evaluation results.

4 LANGUAGE-CONDITIONED VISUAL GEOMETRY

We now ask how language reorganizes the object information already available in the vision encoders. Specifically, we conduct controlled experiments to study the geometric structure of the embedding space that is shaped by linguistic context.

Setup. We generate 324 matched pairs of single-object and two-object scenes using CLEVR. For each pair, the single-object scene contains object A, with attribute combinations uniformly represented across three random positions (3 shapes $\times$ 2 materials $\times$ 2 sizes $\times$ 9 colors $\times$ 3 positions); the two-object scene additionally contains object B with distinct attributes. We first compare the paired scenes processed by the original backbone without cross-attention to examine the encoding of a single object and how adding another object changes the geometry. Next, we hold each two-object image and vary the question supplied to the language-conditioned encoder described in Section 3.1. A generic question specifies an attribute without selecting an object (*e.g.*, "What color is the object?"). We then specify object A as referent (*e.g.*, "What color is the {description of A}?") and change the queried attribute while keeping the description fixed (*e.g.*, "What shape is the {description of A}?"). These comparisons examine how the geometry changes when linguistic context selects an object or specifies a different attribute. We use t-distributed stochastic neighbor embedding (t-SNE), which visualizes local neighborhood geometric structure in low dimensions. Each point represents a scene, obtained by mean-pooling block-11 patch tokens.

Results. As shown in Figure 4, representations of single-object scenes form attribute-related geometric substructures, while they are less apparent when a second object is added. In the third plot, a generic color question shifts the representations according to color (*e.g.*, scenes with yellow objects shift to the left), but the scenes remain mixed, indicating that querying an attribute without specifying the object does not yield separable geometry. However, in the fourth and fifth plots, when the question specifies the referent, the scenes cluster into nine groups by color and three groups by shape, respectively, depending on the queried attribute. These patterns suggest that language conditioning reshapes the geometry according to the referent and the queried attribute.

5 GROUNDING-RELATED ORGANIZATION ACROSS LAYERS

To further investigate the effect of language conditioning, we look into the geometry progression in the visual stream. Section 5.1 provides a case study that visualizes geometry across layers using t-SNE; Section 5.2 verifies that geometry shifts according to described and queried attributes with representational similarity analysis.

5.1 CASE STUDY: T-SNE LAYER-WISE VISUALIZATION

We select a question "What color is the shiny cube?" and 500 distinct images to build the geometry. Each scene is processed under the same question, and its patch tokens are mean-pooled to obtain a scene embedding; each point in the t-SNE visualization represents one scene.

Result. As shown in Figure 1, scenes satisfying the object description first become separable around L7. By L11, these scenes are repartitioned into subgroups aligned with the answer of the queried attribute, making the answer separable to the decoder. Notably, scenes that do not satisfy the description are not involved in the repartition, indicating that queried attribute repartitioning is conditional on description matching separation. Appendix B reports attention-head interventions that test how disrupting description processing affects subsequent attribute organization.

5.2 REPRESENTATIONAL SIMILARITY ANALYSIS

We further use RSA to quantify whether the geometry separates according to object description and groups by queried attribute values across layers.

Measurement. We use representational dissimilarity matrices (RDMs) to compare the geometry of scene representations with the semantic groupings expected for binding and retrieval-related

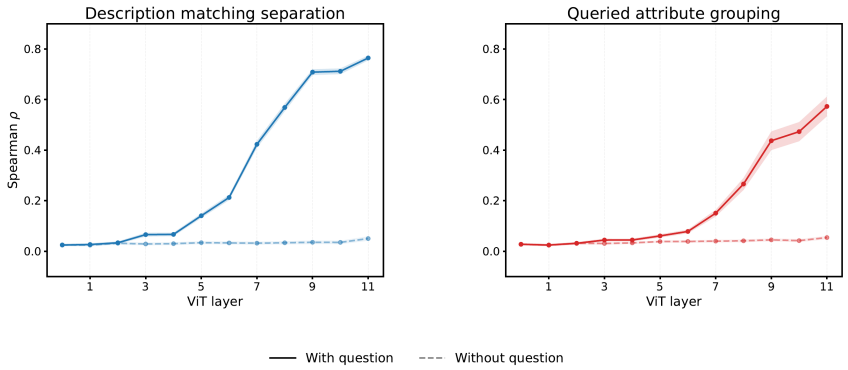


Figure 3: RSA results across layers in DINOv2. Curves show mean Spearman correlations across 72 questions with 500 scenes; queried attribute grouping is evaluated conditioned on description matching scenes. Solid and dashed lines compare inference with and without the question in the same model.

organization. Each RDM entry describes the dissimilarity between two scenes. For binding, the semantic RDM assigns 0 when both scenes contain an object matching the description or neither does, and 1 when only one does. For organization, we restrict the comparison to description-matching scenes and assign 0 when the matching objects share the queried attribute value and 1 otherwise. At each layer, we construct a neural RDM using correlation distances between scene representations and compute its Spearman correlation with each semantic RDM using the off-diagonal entries from one triangle of the matrices. Higher correlations indicate that the visual geometry more closely follows the corresponding semantic grouping: separation by description satisfaction for binding, and grouping by queried attribute value within matching scenes for retrieval.

Setup. We sample 72 questions and 500 distinct images from CLEVR. The scenes contain 3–5 objects, excluding those in which the described object in the question is not unique. Scenes with no matching object are retained for the binding comparison. We extract the mean-pooling representation at each layer with the question supplied through cross-attention and with cross-attention disabled in the same model.

Results. As shown in Figure 3, RSA correlation for description satisfaction rises from 0.14 at L5 to 0.42 at L7 and 0.71 at L9. Within description-matching scenes, correlation for the queried attribute value rises from 0.15 at L7 to 0.43 at L9 and 0.57 at L11. Without language conditioning, both correlations remain below 0.1. These results show that language conditioning progressively organizes scene representations: scenes first separate by description satisfaction, and matching scenes then group by the queried attribute value. We characterize these patterns as binding, which connects linguistic entities to objects in the scene, and organization for retrieval, which makes the queried attribute values separable.

6 Modulation in Object Representation

Next, we investigate how linguistic context modulates representations of objects and their attributes. For binding, we test how the described object is distinguished as the referent. For retrieval, we test how its queried attribute is emphasized to support the answer. We hypothesize that language conditioning adds a shared referent vector to the described object's patch representations and enhances alignment with its queried-attribute prototype, while suppressing alignment with the corresponding prototypes for non-referents and with unqueried-attribute prototypes for the referent.

6.1 Modulation for binding described objects

Setup. We generate 127 two-object scenes across four vision backbones. The two objects have different shapes. For each image, we ask "What color is the {referent}?", changing the shape description to select each object in turn while keeping the image and queried attribute fixed. At each

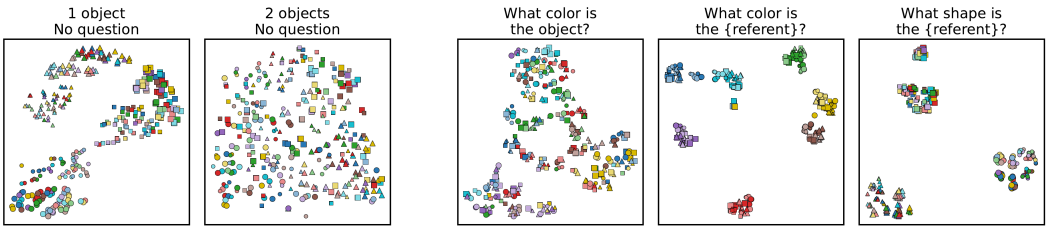


Figure 4: t-SNE visualizations of different scenes and questions in DINOv2. Each point represents a scene, obtained by mean-pooling visual patch tokens. Marker size, shape, and color denote the designated object's actual attribute values respectively: ○ sphere, △ cylinder, and □ cube; darker color shiny, and lighter color matte. Each panel is embedded independently; positions are not comparable across panels. Attribute structure becomes less apparent when a second object is added, while language conditioning reorganizes scenes by the referent's color or shape, depending on the queried attribute.

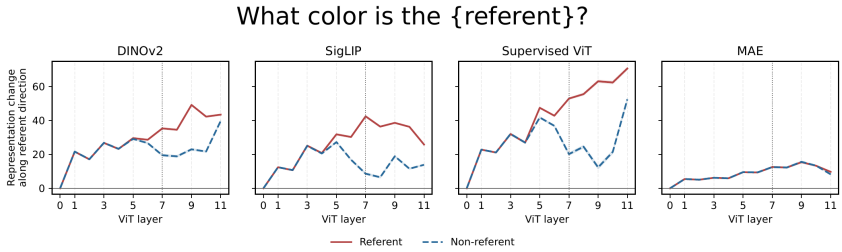


Figure 5: Referent modulation under "What color is the {referent}?". The direction is estimated only from referent-minus-no-question displacements in training images. Held-out referent (red) and non-referent (blue) displacements are projected onto this direction; zero is the no-question baseline.

layer, we average patch tokens within each object's mask to obtain its representation. We estimate a referent direction from referent-minus-no-question displacements in the training images. We then project held-out referent and non-referent displacements, each measured relative to the corresponding no-question representation, onto this direction. This design tests whether referent displacements have larger projections onto the estimated direction than non-referent displacements.

Results. As shown in Figure 5, referent and non-referent representations initially change similarly along the estimated referent direction. By L7, referents show larger changes than non-referents in DINOv2, SigLIP, and supervised-ViT; MAE shows a smaller difference and emerges later. These results support our hypothesis that language conditioning adds a shared referent component to the selected object's representation.

6.2 Modulation for Queried Attribute Retrieval

Setup. We test whether language conditioning selectively enhances the queried attribute of the referent through two comparisons. At each layer, we average patch tokens within each object's mask and measure cosine similarity to its own attribute-value prototype, estimated from single-object images. We subtract the corresponding no-question similarity to quantify enhancement or suppression. First, we compare referents and non-referents under color questions. Keeping each two-object image fixed, we change the referring description to select each object in turn and measure alignment with each object's own color prototype. This comparison tests whether the queried attribute is enhanced for the selected object and suppressed for the other object. Second, we compare queried and unqueried attributes across 140 two-object scenes. Under "What color is the {referent}?", we measure the referent's alignment with its own color and material prototypes. Each object is queried separately, and measurements are averaged within each image. This comparison tests whether conditioning enhances the queried color while suppressing the unqueried material.

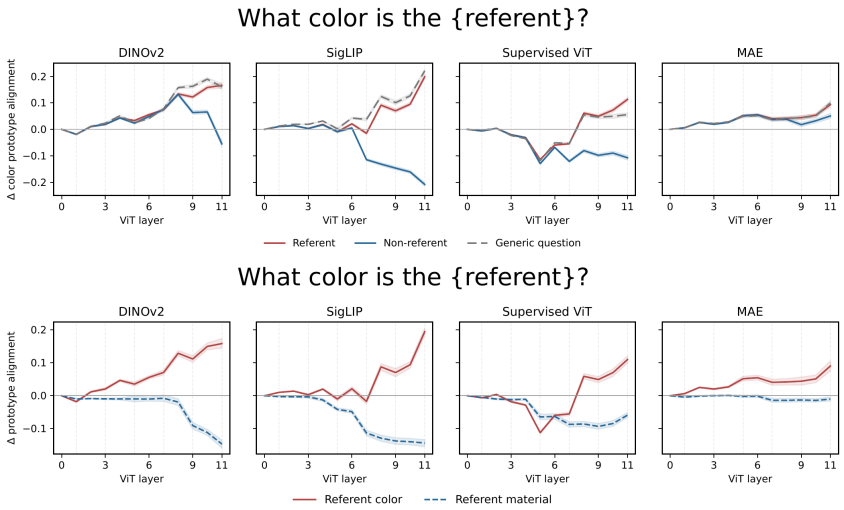


Figure 6: Attribute modulation under the question "What color is the {referent}?" Red and blue curves show changes in the referent's cosine similarity to its own color and material prototypes, respectively, relative to the same object's no-question baseline.

Results. Figure 6 shows higher color-prototype alignment for referents than non-referents across all four backbones. Relative to no question, referent alignment increases, whereas non-referent alignment decreases in DINOv2, SigLIP, and supervised ViT; non-referent alignment remains above baseline in MAE. Figure 6 further shows that under color questions, referent alignment with the queried color prototype increases in later layers, while alignment with the unqueried material prototype decreases across all four backbones. These results show that language conditioning selectively modulates attribute alignment according to which object is referenced and which attribute is queried.

7 Grounding in Natural Images

We extend the attribute modulation analysis in Section 6.2 to natural images using GQA (Hudson & Manning, 2019). We train language-conditioned SigLIP-B/16 and DINOv2-L/14 with one-layer decoders on GQA. For paired validation questions, we hold the image and queried attribute fixed while changing the referred object. We measure each object's cosine similarity to its own queried-attribute value prototype, relative to no question. Prototypes are estimated from a separate pool of 1,661 annotated objects; the analysis includes 345 object-role observations across 144 images.

Figure 7 shows enhanced alignment for referents and suppressed alignment for non-referents. At the final layer, the respective changes are $+0.027$ and -0.046 for SigLIP, and $+0.022$ and -0.040 for DINOv2. These results extend the referent-selective attribute modulation observed in CLEVR to natural images. The analysis primarily covers color questions.

8 Discussion

Binding and organization for retrieval. Our findings connect scene-level organization with modulation of individual object representations. First, scene representations separate according to whether the described object is present, and matching scenes subsequently group by its queried attribute value. Second, at object locations, referents acquire a larger component along the estimated referent direction than non-referents, while attribute alignment changes selectively with the question. These observations support two complementary functions: distinguishing the described object and emphasizing its queried attribute for retrieval. The referent component distinguishes an object's role in the question, whereas prototype alignment measures how strongly its representation emphasizes a particular attribute value. The recurrence of referent enhancement and non-referent suppression after GQA training extends the attribute-modulation finding to natural images, although this analysis primarily covers color questions and does not test the full layer-wise progression.

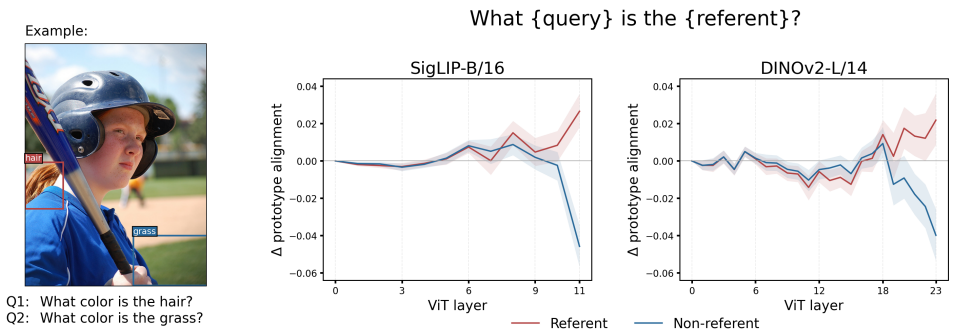


Figure 7: Grounding-related representations after GQA training. Left: an actual paired-question example with annotated objects. Center: raw reference-direction projection change between the two questions. Right: cosine similarity to each object's own attribute-value direction, relative to no question.

Indexing and recoding in symbol processing. Assouel et al. (2026) identify spatial position IDs that language-model attention heads retrieve, select, and use to access an object's features. Song et al. (2026) identify visual recoding within the language model: linguistic context marks goal-relevant objects and modulates their attributes in image-token representations. In these autoregressive models, visual information must become available at the position predicting the next token; indexing directs access to the appropriate object, while recoding shapes the information available for prediction. In our model, recoding occurs within the encoder's visual stream, where linguistic context distinguishes the referent and emphasizes its queried attribute before the representations reach the output head. Together, these findings suggest complementary roles for indexing and recoding in grounding: indexing selects where to retrieve visual information, while recoding changes how that information is represented for the current linguistic context.

Revisiting symbol grounding. Harnad (1990) argues that symbols must be grounded bottom-up in perception, rather than acquiring meaning solely through connections to other symbols. This account raises a further question: how are perceptual representations organized for a particular linguistic context? In prevailing vision-language architectures, the language model performs reasoning over information supplied by the vision encoder. Our findings suggest an alternative: the visual stream itself supports binding and organizes visual information for retrieval, as language conditioning organizes visual representations of objects and their attributes—the grounders that connect linguistic entities to the visual world—to distinguish the referent and emphasize the queried attribute. This perspective is consistent with perceptual symbol systems (Barsalou, 1999), in which perceptual representations support conceptual processing and inference. Within a language-to-vision fusion architecture, perception takes an active computational role in connecting linguistic entities to the visual world. We therefore view grounding as an interaction between perception and linguistic context, rather than as adaptation on the language side alone.

9 Conclusion

Language conditioning enables the visual stream to distinguish described objects and organize their attributes for answering questions. Across scene geometry and object representations, our analyses show how linguistic context shapes the perceptual information that serves as grounders for linguistic entities. Referent-selective attribute modulation also occurs after training on natural images, further supporting an active role for the visual stream in this grounding process.

These findings characterize grounding-related organization that can emerge under language-to-vision conditioning, without implying that such organization is unique to this fusion strategy. Future work should investigate how modulation of individual objects contributes to the organization of scene representations through both representational and mechanistic analyses. Comparisons across fusion strategies can test how architecture shapes this organization, while relational and compositional questions in natural images can extend the analysis beyond direct attribute queries.

AI USE STATEMENT

We used generative AI tools to assist with manuscript writing, language polishing, reference checking, manuscript consistency checks, and implementing experimental and analysis code. All AI-assisted text and code were reviewed and revised by the authors to ensure accuracy and consistency with the reported methods and results. The authors take full responsibility for the final content of this paper.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.

Rim Assouel, Declan Campbell, Yoshua Bengio, and Taylor Webb. Visual symbolic mechanisms: Emergent symbol processing in vision language models. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/pdf/a5cbf39d7924cd3639a8a7a69c8a762bd4c9bc97.pdf.

Lawrence W Barsalou. Perceptual symbol systems. *Behavioral and brain sciences*, 22(4):577–660, 1999.

Irving Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94(2):115–147, 1987.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics*, 42(4):1–10, 2023. doi: 10.1145/3592116.

Arthur Conmy, Augustine N Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *International Conference on Learning Representations (ICLR)*, 2024.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. A primer on the inner workings of transformer-based language models. *arXiv preprint arXiv:2405.00208*, 2024. URL https://arxiv.org/abs/2405.00208.

Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. Interpreting CLIP's image representation via text-based decomposition. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5Ca9sSzuDp.

Michal Golovanevsky, William Schoenegger, Atishay Sharma, Ritambhara Singh, Ellie Pavlick, and Chen Sun. What do VLMs NOTICE? a mechanistic interpretability pipeline for gaussian-noise-free text-image corruption and evaluation. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 2025. URL https://aclanthology.org/2025.naacl-long.571/.

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1–3):335–346, 1990. doi: 10.1016/0167-2789(90)90087-6.

Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6700–6709, 2019.

Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin'e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*, 2023.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2901–2910, 2017.

Sonia Joseph, Praneet Suresh, Lorenz Hufe, Edward Stevinson, Robert Graham, Yash Vadi, Danilo Bzdok, Sebastian Lapuschkin, Lee Sharkey, and Blake Aaron Richards. Prisma: An open source toolkit for mechanistic interpretability in vision and video. In *CVPR 2025 Workshop on Mechanistic Interpretability in Vision (MIV)*, 2025. URL https://arxiv.org/abs/2504.19475. Oral Presentation, arXiv:2504.19475.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Kang_Your_Large_Vision-Language_Model_Only_Needs_A_Few_Attention_Heads_CVPR_2025_paper.pdf.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Benno Krojer, Shravan Nayak, Oscar Mañas, Vaibhav Adlakha, Desmond Elliott, Siva Reddy, and Marius Mosbach. Latentlens: Revealing highly interpretable visual tokens in llms. *arXiv preprint arXiv:2602.00462*, 2026.

Michael A Lepori, Thomas Serre, and Ellie Pavlick. Break it down: Evidence for structural compositionality in neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, 2022.

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 *International Conference on Machine Learning (ICML)*, 2023.

Yihao Li, Saeed Salehi, Lyle H. Ungar, and Konrad Kording. Does object binding naturally emerge in large pretrained vision transformers? In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0119. URL https://papers.neurips.cc/paper_files/paper/2025/hash/050b8ff31bee2dfea65b731e71baccd5-Abstract-Conference.html.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, 2024.

Yiming Liu, Yuhui Zhang, and Serena Yeung-Levy. Mechanistic interpretability meets vision language models: Insights and limitations. ICLR 2025 Blogpost Track, 2025. URL https://openreview.net/forum?id=pZqvfsUpeh.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman, 1982.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Jungwon Park, Jungmin Ko, Dongnam Byun, Jangwon Suh, and Wonjong Rhee. Cross-attention head position patterns can align with human visual concepts in text-to-image generative models. In *International Conference on Learning Representations (ICLR)*, 2025.

Ellie Pavlick. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A*, 381(2251), 2023.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Jona Ruthardt, Manu Gaur, Deva Ramanan, Makarand Tapaswi, and Yuki M. Asano. Steerable visual representations. In *Computer Vision – ECCV 2026*, volume 17071 of *Lecture Notes in Computer Science*, pp. 357–374. Springer, 2026. URL https://jonaruthardt.github.io/project/SteerViT/.

Darshana Saravanan, Makarand Tapaswi, and Vineet Gandhi. Investigating mechanisms for in-context vision language binding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025W/MIV/papers/Saravanan_Investigating_Mechanisms_for_In-Context_Vision_Language_Binding_CVPRW_2025_paper.pdf.

Roger N. Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703, 1971.

Brian Song, Michael A. Lepori, and Ellie Pavlick. Linguistic context recodes visual representations in vision-language models, 2026. URL https://arxiv.org/abs/2608.00035.

Samuel Stevens, Wei-Lun Chao, Tanya Berger-Wolf, and Yu Su. Interpretable and testable vision features via sparse autoencoders. *arXiv preprint arXiv:2502.06755*, 2025. URL https://arxiv.org/abs/2502.06755.

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

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*, 2023.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Qidong Wang, Junjie Hu, and Ming Jiang. V-SEAM: Visual semantic editing and attention modulating for causal interpretability of vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025. URL https://aclanthology.org/2025.emnlp-main.880/.

Taylor W Webb, Ishita Sinha, and Jonathan D Cohen. Emergent symbols through binding in external memory. In *International Conference on Learning Representations (ICLR)*, 2021.

Shuyu Wu, Ziqiao Ma, Xiaoxi Luo, Yidong Huang, Josue Torres-Fonseca, Freda Shi, and Joyce Chai. The mechanistic emergence of symbol grounding in language models. In *International Conference on Machine Learning (ICML)*, 2026. URL https://sled.eecs.umich.edu/publication/wu-2026-emergence/.

Yukang Yang, Declan Iain Campbell, Kaixuan Huang, Mengdi Wang, Jonathan D. Cohen, and Taylor Whittington Webb. Emergent symbolic mechanisms support abstract reasoning in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70515–70549, 2025. URL https://proceedings.mlr.press/v267/yang25c.html.

Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/ea370419760b421ce12e3082eb2ae1a8-Abstract-Conference.html.

A Full Evaluation Results

We report CLEVR validation accuracy for attribute queries (QryAttr), attribute comparisons (EqAttr), existence (Exist), counting (Count), and integer comparisons (CmpInt). The backbone comparison and the primary DINOv2 ablations average seeds 42, 43, and 44.

Additional ablations. The following variants use a decoder receiving visual tokens only; the question conditions the visual stream through cross-attention.

Table 1: CLEVR validation accuracy (%), averaged over three seeds, with language-to-vision cross-attention and a decoder receiving visual and question tokens.

Model / variant	Overall	QryAttr	EqAttr	Exist	Count	CmpInt
DINOv2	92.5	99.1	92.5	96.1	86.1	77.7
SigLIP	93.5	99.2	95.1	96.8	87.0	79.9
Supervised-ViT	85.2	94.3	82.4	89.6	75.5	73.5
MAE	76.1	91.2	61.7	80.6	63.3	71.7

Table 2: Classifier-readout accuracy (%) for seed 42. These single-seed results are separate from the three-seed DINOv2 ablation below.

Model / variant	Overall	QryAttr	EqAttr	Exist	Count	CmpInt
DINOv2	90.1	98.1	88.7	94.2	82.4	75.7
SigLIP	84.8	92.8	82.6	89.8	76.2	72.2
Supervised-ViT	86.6	92.5	87.6	91.8	79.3	73.0
MAE	77.0	86.4	74.4	81.5	65.2	68.9

Table 3: DINOv2 component ablations on CLEVR validation (%). The decoder baseline receives visual and question tokens; removing CA retains the question input to the decoder.

Model / variant	Overall	QryAttr	EqAttr	Exist	Count	CmpInt
DINOv2 + decoder	92.5	99.1	92.5	96.1	86.1	77.7
- Decoder + classifier	89.4	97.0	87.8	93.9	81.8	75.9
- CA (no language conditioning)	50.6	52.0	53.2	59.0	41.7	50.5

Table 4: Additional DINOv2 ablations with visual tokens only supplied to the decoder.

Model / variant	Overall	QryAttr	EqAttr	Exist	Count	CmpInt
Visual tokens only to decoder	91.0	98.7	89.7	94.9	84.0	74.9
ViT backbone from scratch	52.6	48.6	51.6	66.5	45.5	67.8
- RoBERTa + learned text encoder	80.4	85.1	82.5	90.4	73.2	61.7

B MECHANISTIC INTERPRETABILITY ANALYSES

Scope and motivation. To better understand how language conditioning supports grounding, we additionally examine the model's internal components through activation patching and interventions on representational geometry. Component-level analyses can inform targeted model editing by identifying where an intervention changes predictions (Meng et al., 2022). Here, however, the identified heads belong to a model trained for our controlled setting, limiting their direct applicability to editing other models. We therefore present these model-specific analyses in the appendix, as a complement to the representational findings in the main text.

Mechanistic Interpretability in VLMs Mechanistic interpretability has developed mature tools for language models, including activation patching, causal tracing, and circuit discovery (Wang et al., 2023; Conmy et al., 2023; Meng et al., 2022; Olsson et al., 2022). Extending these tools to vision and vision-language models remains difficult because visual representations encode dense, continuous, and often ambiguous scene information (Liu et al., 2025). Recent work has analyzed visual concepts in image encoders (Joseph et al., 2025; Stevens et al., 2025; Gandelsman et al., 2024) and semantic structure in cross-attention (Park et al., 2025). NOTICE uses semantically controlled image-text corruptions with activation patching to support causal interpretation (Golovanevsky et al., 2025). V-SEAM combines visual semantic editing with head-level modulation to measure causal effects across semantic categories (Wang et al., 2025). We use mechanistic analyses to characterize grounding-related visual geometry.

B.1 EXPERIMENTAL OVERVIEW

We next ask which information sources and model components support the binding and organization mechanisms identified in Section 5. We first inspect layer-wise cross-attention patterns to test whether different parts of the question are accessed at different stages of processing. We then use head-wise activation patching to identify components that causally influence description-related and queried-attribute representations.

B.2 ACTIVATION PATCHING

We further investigate which components account for the mechanism through head-wise activation patching (Meng et al., 2022; Wang et al., 2023). We construct corrupted inputs by swapping keywords in the question: *attribute value* (e.g., "red" $\rightarrow$ "blue") and *attribute query* (e.g., "what shape" $\rightarrow$ "what material") across four attributes. Patching restores the clean activation of a single head and measures recovery of the clean prediction.

To identify the causal contribution of individual model components to text-conditioned visual reasoning, we employ activation patching in both denoising and noising directions. Given a query question q and its text-corrupted counterpart $\tilde{q}$ obtained by swapping attribute values, we perform two complementary interventions. In the denoising direction, a forward pass is performed with the corrupted input $\tilde{q}$ while restoring the activations of a target component from a clean forward pass with q. In the noising direction, we run a clean forward pass with q but replace the target head activation with the corresponding activation recorded from the corrupted-question forward pass $\tilde{q}$.

Specifically, for a CA head at layer ℓ with head index h, we record the pre-projection activations $\mathbf{z}^{(\ell,h)}$ at the output of the cross-attention module and substitute the corresponding slice during the counterfactual forward pass on the projection layer. We report the change in the ground-truth answer logit: $\Delta = \mathrm{logit}_{\mathrm{patched}}(a^*) - \mathrm{logit}_{\mathrm{baseline}}(a^*)$, averaged over corruption pairs sampled per attribute type.

As shown in Figure 8, for described-attribute corruption, the largest cross-attention effects occur at L5H0 for color ($+1.06$), L7H9 for material ($+1.44$), L7H11 for size ($+0.56$), and L7H3 for shape ($+0.67$). As shown in Figure 9, for queried-attribute corruption, mean recovery across self-attention heads is largest at L11 for all four attribute types. These results identify components whose activations influence predictions under changes to the description or queried attribute; they motivate the following interventions on representational organization.

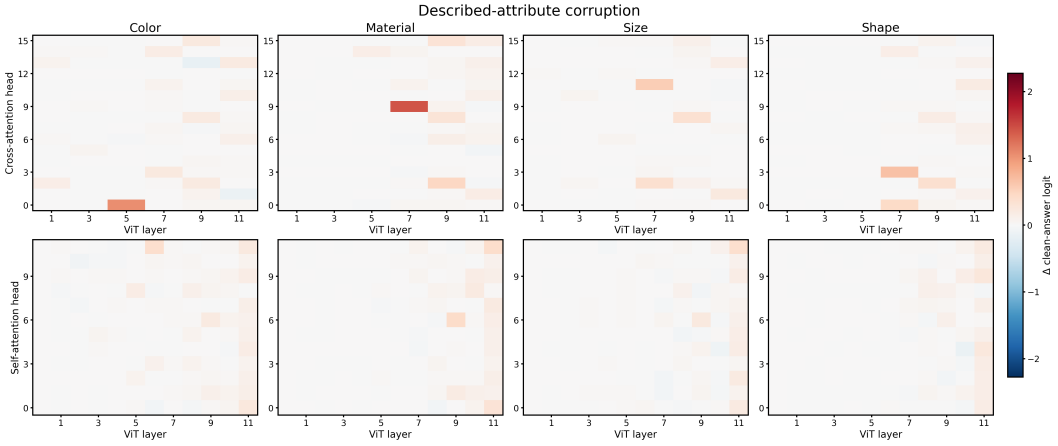


Figure 8: Head-wise patching after described-attribute corruption. Each cell shows the mean change in the clean answer's logit over 50 corruption pairs when one clean head activation replaces its corrupted counterpart. Top: cross-attention heads; bottom: self-attention heads. Red denotes increased and blue decreased clean-answer logits; layer and head indices are zero-based.

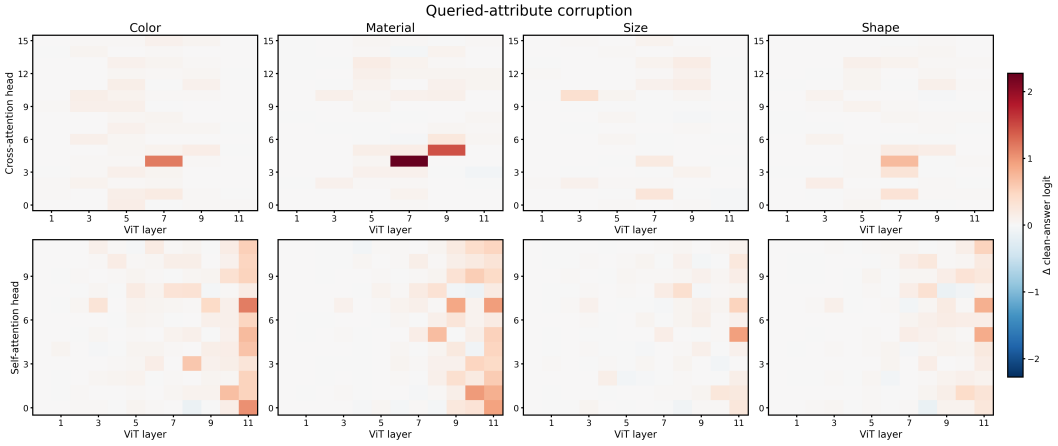


Figure 9: Head-wise patching after queried-attribute corruption, using the same measurement and color scale as Figure 8. Columns identify the queried attribute in the clean question. Each cell averages 50 corruption pairs.

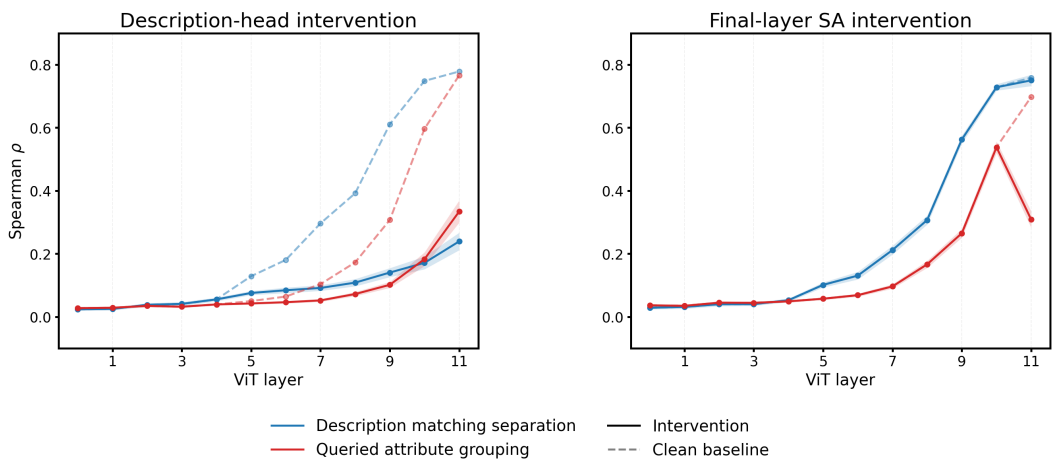


Figure 10: Representational similarity after attention interventions. Left: noising four selected cross-attention heads. Right: noising final-block self-attention. Solid curves show intervention results and faint dashed curves the original reference curves. Blue denotes description matching separation and red denotes queried attribute grouping (See Section 5.2).

B.3 Causal Intervention

After identifying the description-matching and queried-attribute components, we use causal interventions with RSA to probe their functional roles. As shown in Figure 10, noising the four description matching separation heads (L5H0, L7H9, L7H11, L7H3) reduces description matching separation. Notably, at L11, the queried attribute grouping RSA still rises despite the disrupted upstream geometry. The model still partitions matching scenes by queried attribute value even when the upstream description matching separation is disrupted, producing the mixed clusters. In contrast, SA11 late-stage intervention selectively disrupts queried attribute grouping while preserving description matching separation. This dissociation supports distinct causal roles for the identified components in binding and organization for retrieval.

B.4 Interpretation and Limitations

These operations are analogous to mechanisms previously localized in language-model components (Assouel et al., 2026). In the visual stream, however, symbol-like processing appears as description- and attribute-aligned geometry rather than position-based indexing over discrete tokens. Our results provide preliminary evidence linking model function to intermediate visual geometry, but they do not yet constitute a complete circuit-level account. Future work should use path patching and related causal interventions to identify the pathways connecting the components, and to test more directly whether description- and attribute-aligned geometry mediates the model's final predictions.