

Investigating Mechanisms for In-Context Vision Language Binding

Darshana Saravanan

Makarand Tapaswi

Vineet Gandhi

CVIT, IIIT Hyderabad, India

Abstract

To understand a prompt, Vision-Language models (VLMs) must perceive the image, comprehend the text, and build associations within and across both modalities. For instance, given an ‘image of a red toy car’, the model should associate this image to phrases like ‘car’, ‘red toy’, ‘red object’, etc. Feng and Steinhardt [4] propose the Binding ID mechanism in LLMs, suggesting that the entity and its corresponding attribute tokens share a Binding ID in the model activations. We investigate this for image-text binding in VLMs using a synthetic dataset and task that requires models to associate 3D objects in an image with their descriptions in the text. Our experiments demonstrate that VLMs assign a distinct Binding ID to an object’s image tokens and its textual references, enabling in-context association.

1. Introduction

As Vision-Language models (VLMs) like Gemini [15] and GPT-4o [6] become ubiquitous, it is crucial to understand how they function to determine why they respond the way they do, especially in safety-critical applications. A fundamental ability of VLMs is to associate information across an image and text to reason about a query. For example, given an *image of a furniture store that has a chair with a yellow tag* and the caption *All furniture with a yellow tag have a 30% discount*, a VLM should be able to infer that the chair has a discounted selling price. Our goal is to study this ability to *bind* objects in an image to information in text. To this end, we propose the *Shapes* task, a controlled synthetic task that requires models to associate 3D objects in an image with their references in the text. In Fig. 1, the image contains two 3D objects: a *green sphere* and a *red cube*. The green sphere is referred to as the *green object* in the context. So, to answer the question ‘*What does the sphere contain?*’, the model needs to internally learn that the sphere corresponds to the phrase *green object*: *is*(‘*green sphere patches*’, ‘*green object*’), and that this *green object* contains item P: *contains*(‘*green object*’, ‘*item P*’).

The binding ID mechanism proposed in [4] suggests that LLMs’ internal activations represent binding information

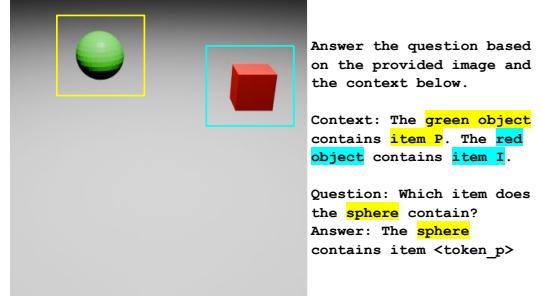


Figure 1. **Shapes Task.** Given an image with two 3D objects and a text description (context), the model needs to comprehend the question and identify the correct item (`<token_p>`) contained in the queried object. Image and text tokens highlighted with the same color are expected to contain the same binding IDs, allowing the model to predict the correct answer.

by attaching binding ID vectors to the corresponding entities and attributes. We investigate whether VLMs use a similar mechanism to represent associations between image tokens and text tokens. We study the most commonly used VLM architecture that consists of a visual encoder, a multimodal projector and a language model.

VLMs and LLMs have some key differences that necessitate careful experimentation. (i) Text tokens have fixed embeddings, while concepts in an image (objects, colors, textures, etc.) do not have fixed embeddings; they are represented in the patch tokens obtained from the vision encoder. (ii) Recent powerful VLMs like LLaVA-OneVision [10], Molmbo [3], and Qwen2-VL [17] utilize an image encoder that converts the input image into a set of multiscale, multi-crop images and independently maps each of these images into a set of vision tokens. This leads to multiple sets of tokens for the same visual concept. We adapt the causal mediation based experiments from [4] to account for these differences and make the following observations: (i) Image tokens corresponding to the location of the visual concept represent information related to that concept. This is applicable even when there are multiple tokens corresponding to multiple crops from the same image. (ii) VLMs implement the binding ID mechanism. There are binding ID vectors that associate the image tokens corresponding to a visual object and its references in the text tokens.

2. Task Definition and Notations

Shapes task. This task consists of images with two 3D objects (O_0, O_1) with distinct shapes and colors. The context refers to both objects using their color (C_0, C_1) and assigns a unique item (I_0, I_1). We use the notation $c = \text{ctxt}(O_0 \leftrightarrow C_0 \leftrightarrow I_0, O_1 \leftrightarrow C_1 \leftrightarrow I_1)$ to denote a context where object O_0 of the color C_0 contains item I_0 and object O_1 of the color C_1 contains item I_1 . In Fig. 1, O_0 and O_1 correspond to the *green sphere* and *red cube* patches in the image, C_0 and C_1 correspond to *green* and *red* in the text and I_0 and I_1 correspond to *item P* and *item I* in text respectively. The question refers to one of the objects using its shape and queries the item assigned to it. Note that ‘*item P/I*’ are randomly chosen uppercase English letters with no inherent meaning.

We generate the images using Blender [2]. We consider four choices for the shape (cone, cube, cylinder and sphere) and six choices for the color (red, blue, green, yellow, cyan and purple). The objects occupy a fixed number of patches and are located in fixed positions.

Notation. Let $\Phi_v(\cdot)$ denote the vision encoder and $g(\cdot)$ denote the multi-modal projector. For an image X_v , the patch embeddings are $t_v = g(\Phi_v(X_v))$. Now, let t_c denote the prompt tokens, comprising image tokens t_v and text tokens up to the context’s end, just before the question. Let the LLM have L transformer layers and D -dimensional activation space. For every token position p , $Z_p \in R^{L \times D}$ is the stacked set of residual stream activations. The activations at the object, color, and item positions are denoted as Z_{O_k} , Z_{C_k} and Z_{I_k} respectively where $k \in \{0, 1\}$.

3. Do Binding IDs Occur in VLMs?

Binding Id Mechanism. Feng and Steinhardt [4] suggest that LLMs associate concepts through binding ID vectors in their activations. Specifically, the activations of an LLM can be decomposed into vectors that encode the concept and those that encode the binding information. Each binding ID consists of similar vector pairs in a subspace, with associated concepts sharing one vector from the same ID. Extending this, we describe our hypothesis for the existence of binding IDs in VLMs using the Shapes task below:

- Consider 3D objects as visual entities and their colors and items mentioned in the text as their attributes. For the k^{th} visual entity-attributes tuple (I_k, C_k, O_k) , the model represents binding vectors in its activations in an abstract form, independent of any particular object, color, or item.
- For object patch tokens, the activations Z_{O_k} can be decomposed as $Z_{O_k} = f_O(O_k) + b_O(k)$. Similarly, $Z_{C_k} = f_C(C_k) + b_C(k)$ and $Z_{I_k} = f_I(I_k) + b_I(k)$. Here $f_O(O_k)$, $f_C(C_k)$, $f_I(I_k)$ are the content vectors and the set of binding vectors $(b_O(k), b_C(k), b_I(k))$ form the binding

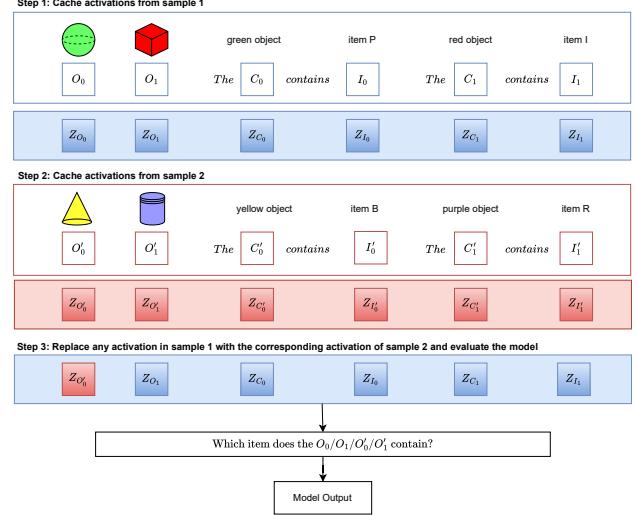


Figure 2. Causal intervention. In steps 1 and 2, activations from the first and second samples are saved. In step 3, object/color/item activations in the first sample are replaced with those from the second. This new activation stack is frozen, and the model is queried with all four objects to observe the change in predictions.

ID for the k^{th} tuple.

- To answer the question about an object, the model selects the item that shares the same binding ID.

Note that, since binding IDs are independent of the entity/attribute, we can manipulate the associations built by the model by exchanging the binding IDs in the activations as $\hat{Z}_{O_k} := Z_{O_k} - b_O(k) + b_O(k')$ where $k \neq k'$. In the following sections, we assert the existence of the binding ID mechanism by establishing two of its properties: Factorizability (Sec. 3.1) and Position independence (Sec. 3.2). Then, we exchange the associations built by the model using Mean interventions (Sec. 3.3).

3.1. Factorizability

Fig. 2 shows two samples from the Shapes task with the contexts $c = \text{ctxt}(O_0 \leftrightarrow C_0 \leftrightarrow I_0, O_1 \leftrightarrow C_1 \leftrightarrow I_1)$ and $c' = \text{ctxt}(O'_0 \leftrightarrow C'_0 \leftrightarrow I'_0, O'_1 \leftrightarrow C'_1 \leftrightarrow I'_1)$.

The Binding ID mechanism assumes that the information linking a concept to its attributes is stored locally within the activations at its token positions and is independent of the specific concept itself. This implies that the activations of the *sphere* (O_0) in the first sample and the *cone* (O'_0) in the second sample should contain the same binding vector $b_O(0)$ as they both correspond to the 0th visual entity-attributes tuple in their respective samples. Replacing Z_{O_0} with $Z_{O'_0}$ should now bind the *cone* with the text tokens *green object* and *item P*. We demonstrate this using causal interventions [16] on the activations as described below.

- Cache all activations Z_c from the model run on c .
- Cache activations $Z_{O'_0}$ and $Z_{O'_1}$ from the model run on c' .
- Construct a new stack of activations Z_c^* by modifying Z_c

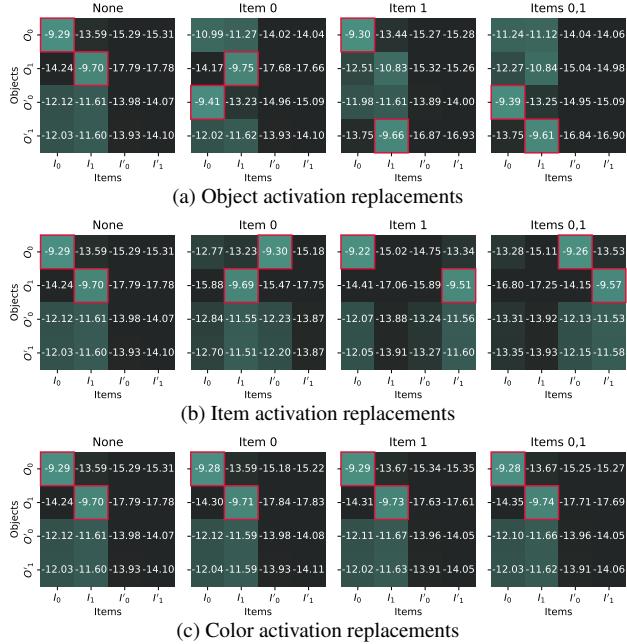


Figure 3. Factorizability results. Each row shows the model’s mean log probabilities of an item contained in an object. The first grid in each case shows results with unaltered activations. Squares highlighted in red denote the expected predictions based on our hypothesis. Model outputs match hypothesis suggesting a multi-modal binding ID mechanism.

- such that Z_{O_k} is replaced with $Z_{O'_k}$ for any $k \in \{0, 1\}$.
- Re-evaluate the model by probing what item each shape (O_0, O_1, O'_0, O'_1) contains by freezing the activation cache as Z_c^* . We expect the model to now associate O'_k with I_k since both Z_{O_k} and $Z_{O'_k}$ contain the same binding ID vector $b_O(k)$.

Results. Fig. 3 shows the mean log probability of choosing an item before and after interventions. We show the factorizability results for object patch tokens, color tokens and item tokens. In Fig. 3a, the **first** grid shows the results when the *activations are unaltered*. As expected, for objects O_0 and O_1 , items I_0 and I_1 are chosen at a higher rate, respectively and for objects O'_0 and O'_1 , items I_0 and I_1 are chosen at a roughly equal rate since these objects do not exist in the image. In the **second** grid, we replace Z_{O_0} with $Z_{O'_0}$. Now, when the model is queried for the item contained by O'_0 , the model picks item I_0 over I_1 . The **third** grid follows the same pattern, Z_{O_1} is replaced by $Z_{O'_1}$ resulting in O'_1 containing I_1 . Finally, both object activations are replaced in the **fourth** grid and we observe that the model chooses $I_{0/1}$ for $O'_{0/1}$ respectively. Note that when the object patches are replaced, the color of the new object no longer matches the color description in the text. Nevertheless, the new object is still associated with the same item as the original object, as they both contain the same binding vector.

We observe a similar behavior for replacing items in

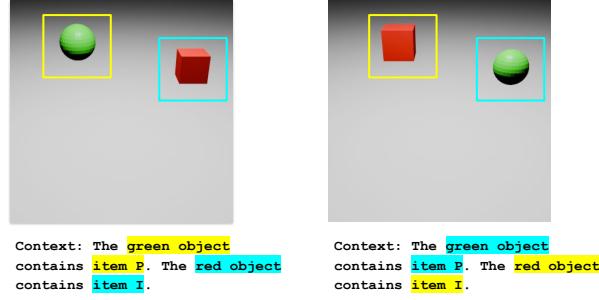


Figure 4. Mean intervention samples.

Fig. 3b. When Z_{I_k} is replaced by $Z_{I'_k}$, the model prefers item I'_k for object O_k . However, when we intervene on the color activations Z_{C_k} , the results are similar to when there are no interventions (Fig. 3c). This is expected since both Z_{C_k} and $Z_{C'_k}$ contain the same binding ID vectors.

3.2. Position Independence

Next, we hypothesize that the associations formed by the model are invariant to the activation positions of the object, color, or item, as they rely solely on the binding IDs. This implies that swapping the positions of Z_{O_0} and Z_{O_1} should not change items associated with the objects. To validate this, we first obtain the activations of the context tokens Z_c (Sec. 3.1). Then, we compute a new stack of activations Z_c^* wherein the positions of Z_{O_0} and Z_{O_1} are altered, following the procedure described in [4], adapted for models that use Rotary Position Embedding (RoPE) [14]. Unlike absolute position embeddings, RoPE incorporate positional information only through the attention score computations, without injecting it directly into the residual stream activations.

Results. Fig. 5 shows the mean log probabilities when the positions of Z_{O_0} and Z_{O_1} are progressively adjusted to get closer and ultimately swapped. We observe that the model answers with the correct item regardless of positions.

3.3. Mean Interventions

The factorizability and position independence results show that binding vectors are contained within the activations corresponding to the object, color, and item tokens and cause the model to form associations across image and text. If binding vectors were directly accessible, we could interchange them to observe if the model changes its answer. While this is not feasible, we can approximate the difference in binding vectors from the difference in activations. To estimate $\Delta_O = b_O(1) - b_O(0)$, we consider two instances of the Shapes task as shown in Fig. 4. Let O_0, O_1 denote the objects in the first instance and O'_0, O'_1 denote the objects in the second instance. Notice that both O_0 and O'_1 are the same object, a *green sphere*. However, we expect their activations to contain different binding IDs. We can now estimate Δ_O as the difference $Z_{O'_1} - Z_{O_0}$. Concretely,

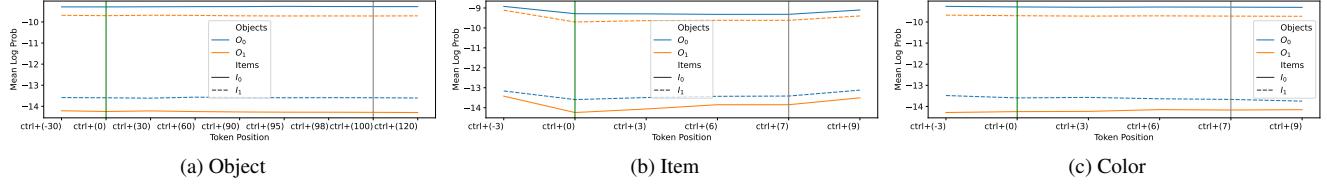


Figure 5. Position independence results. The integers in the x-axis show how much the position of the first and second objects/items/colors are incremented and decremented respectively. The green line corresponds to no change in positions and the gray line corresponds to swapped positions. In all cases $O_k \leftrightarrow I_k$ (blue solid O_0, I_0 and orange dashed O_1, I_1) have a higher probability than $O_k \leftrightarrow I_k'$.

Condition	Mean vectors		Random vectors	
	$O_0 \leftrightarrow I_0$	$O_1 \leftrightarrow I_1$	$O_0 \leftrightarrow I_0$	$O_1 \leftrightarrow I_1$
None	1.00	1.00	-	-
O	0.00	0.05	1.00	1.00
I	0.05	0.00	1.00	1.00
C	1.00	1.00	1.00	1.00
O, I	1.00	0.95	1.00	1.00
O, I, C	1.00	0.95	1.00	1.00

Table 1. Mean ablation accuracies: Object (O), Item (I), Color (C).

we compute Δ_O as the mean of the difference of activations over multiple pairs of instances ($\Delta_O \approx \text{mean}_{O_0, O'_1} [Z_{O'_1} - Z_{O_0}]$). Similarly, we compute $\Delta_C = b_C(1) - b_C(0)$ and $\Delta_I = b_I(1) - b_I(0)$ from the color and item activations.

Using these mean vectors ($\Delta_O, \Delta_C, \Delta_I$), we can now edit the binding vectors in the activations to alter the model response. For any new instance with the context $c^* = \text{ctxt}(O_0^* \leftrightarrow C_0^* \leftrightarrow I_0^*, O_1^* \leftrightarrow C_1^* \leftrightarrow I_1^*)$, we can alter the binding vector of the objects as $Z_{O_0^*} := Z_{O_0^*} + \Delta_O$ and $Z_{O_1^*} := Z_{O_1^*} - \Delta_O$. This should result in a swap of object-item binding with O_0^* and O_1^* being bound to I_1^* and I_0^* respectively. Similarly, altering the binding vector of the items as $Z_{I_0^*} := Z_{I_0^*} + \Delta_I$ and $Z_{I_1^*} := Z_{I_1^*} - \Delta_I$ should also exchange the model response. Altering the binding vectors in color token activations will make the model now associate $O_k, C_{k'}$ and I_k where $k \neq k'$. However, O_k is still bound to I_k , and we expect no change in response.

Results. Tab. 1 shows the accuracy, measured as the fraction of samples where the correct item has the highest log probability among the possible items in the context. As expected, both object and item interventions individually change the model’s response, while color interventions do not. Further, simultaneously performing object and item interventions restores the model’s original response since they now have the same binding IDs. We also repeat these experiments with random vectors that have the same magnitude but different directions. These vectors do not alter the model response, indicating that the specific directions of the mean vectors causally affect the binding.

3.4. Experimental Details

Throughout the paper, we report results with LLaVA-OneVision-7B [10], which uses the SigLIP [18] vision en-

coder and encodes multiple crops from a single image. The Shapes task images are of size 384×384 , with each object appearing in two crops and occupying 5×5 patch tokens. Empirically, we found that when intervening on object token activations, a 3-token padding on all sides in both crops yields optimal results. To estimate the difference of binding vectors, we use a separate set that contains different shapes (frustum, pyramid, prism and toroid), colors (lime, pink, gold, brown, orange and azure) and items (lowercase English alphabet). All colors and items span two text tokens.

4. Related Work

The Binding ID mechanism explains how LLMs associate concepts in context, leading to the identification of a binding subspace where bound tokens have a higher similarity than unbound ones [5]. Concurrently, researchers uncovered circuits for entity tracking in LLMs, allowing inference of entity properties from context [12]. The Shapes task is inspired by the text-based entity tracking task [9], which requires predicting an entity’s state based on its initial description and applied operations.

Prior works have analyzed attention heads in VLMs to understand visual processing [8], shown that object information is localized to corresponding image token positions [11], and developed methods to manipulate image token representations to mitigate hallucinations [7]. Our work complements these efforts by examining the association between image and text representations.

Benchmarks like VTQA [1] and MuMuQA [13] pose multi-hop questions that require synthesis of visual and textual information, going beyond traditional VQA where answers rely primarily on visual inputs. They present an opportunity to explore how mechanisms such as Binding IDs could enhance reasoning in complex, realistic scenarios.

5. Conclusion

In this work, we explore how in-context associations occur in VLMs. We formulate the Shapes task, a simple and controlled QA task which requires the model to associate 3D objects in an image with their references in the text. Through experiments, we demonstrate that VLMs utilize binding ID vectors to bind concepts across image and text.

References

- [1] Kang Chen and Xiangqian Wu. VTQA: Visual Text Question Answering via Entity Alignment and Cross-Media Reasoning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 4
- [2] Blender Online Community. *Blender - A 3D Modelling and Rendering Package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 2
- [3] Matt Deitke et al. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. *arXiv preprint arXiv:2409.17146*, 2024. 1
- [4] Jiahai Feng and Jacob Steinhardt. How do Language Models Bind Entities in Context? In *International Conference on Learning Representations (ICLR)*, 2024. 1, 2, 3
- [5] Jiahai Feng, Stuart Russell, and Jacob Steinhardt. Monitoring Latent World States in Language Models with Propositional Probes. In *International Conference on Learning Representations (ICLR)*, 2025. 4
- [6] Aaron Hurst et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1
- [7] Nicholas Jiang, Anish Kachinthaya, Suzanne Petryk, and Yossi Gandelsman. Interpreting and Editing Vision-Language Representations to Mitigate Hallucinations. In *International Conference on Learning Representations (ICLR)*, 2025. 4
- [8] Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the Image? A Deep-Dive into the Vision of Vision Language Models. *arXiv preprint arXiv:2411.17491*, 2024. 4
- [9] Najoung Kim and Sebastian Schuster. Entity Tracking in Language Models. In *Association of Computational Linguistics (ACL)*, 2023. 4
- [10] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 4
- [11] Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards Interpreting Visual Information Processing in Vision-Language Models. In *International Conference on Learning Representations (ICLR)*, 2025. 4
- [12] Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-Tuning Enhances Existing Mechanisms: A Case Study on Entity Tracking. In *International Conference on Learning Representations (ICLR)*, 2024. 4
- [13] Revant Gangi Reddy et al. MuMuQA: Multimedia Multi-Hop News Question Answering via Cross-Media Knowledge Extraction and Grounding. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2022. 4
- [14] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. *arXiv preprint arXiv:2104.09864*, 2021. 3
- [15] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1
- [16] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2
- [17] Peng Wang et al. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1
- [18] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *International Conference on Computer Vision (ICCV)*, 2023. 4