DEFINITELY MAYBE: CAN MULTIMODAL IN-CONTEXT LEARNING INDUCE VISUAL CONCEPTS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) can adapt their predictions to multimodal demonstrations in context, but whether they can induce visual concepts from a few demonstrations during multimodal in-context learning (MICL) remains unclear. Existing evaluations often confound concept induction with recognition of familiar semantic associations. It is therefore difficult to determine whether demonstrations cause MLLMs to acquire new information or only change how existing information is accessed. We study this question in a controlled setting and disentangle these two possibilities. Specifically, we decompose MICL performance into *signal power*, measuring how separable the target concept is inside the low-dimensional subspace the frozen language head can see, and *readability*, measuring how effectively the head aligns with that separation. Across our analysis, we find that MLLMs can induce visual concepts from context. However, demonstrations primarily improve the readability of pre-existing visual signals by steering them toward the frozen readout, rather than substantially changing what the query representation encodes. Additionally, we find that this steering is not selective: demonstrations can also align irrelevant or competing concepts with the readout, creating interference that limits concept induction. Our findings suggest that current MLLMs primarily *learn to read rather than write* during MICL: demonstrations largely teach the model how to access information already encoded in its representations, rather than providing new task-relevant information.

1 INTRODUCTION

Today's multimodal large language models (MLLMs) show remarkable capabilities in jointly processing visual and textual inputs. However, they are typically trained once on large-scale datasets and then deployed as static models. As a result, they lack the ability to continuously adapt to changes in the world around them. One promising solution to this limitation is in-context learning (ICL), which enables foundation models to learn new concepts or tasks from a few examples provided in context, without updating their weights Brown et al. (2020); Agarwal et al. (2024). Based on this, extending text-based ICL to multimodal in-context learning (MICL) by including visual demonstrations as examples in context has recently emerged as an active area of research Alayrac et al. (2022).

Existing work mostly studies MICL in standard few-shot learning scenarios in which the similarity criterion used to classify images is fixed by the dataset Wang et al. (2026); Agarwal et al. (2024), meaning that there is only one valid way to classify the same images. However, we argue that adapting this similarity criterion is precisely where MICL could offer its greatest advantage by inducing visual concepts solely from visual context. Conventional visual representations are learned as fixed embedding functions and therefore implicitly encode a single notion of similarity. Consequently, nearest-neighbour retrieval or linear probing on pretrained embeddings is biased toward whichever features dominate the embedding space. Recent approaches address this limitation by conditioning the representation on a natural-language prompt (Vaze et al., 2023a; Ruthardt et al., 2026). However, language-based conditioning assumes that the relevant concept can be explicitly named, whereas many visual concepts, such as a particular texture, are difficult to describe in words. Here, MICL offers an alternative by using visual demonstrations to condition the model's query representation. Thus, the same image can be encoded differently depending on the context, allowing the relevant similarity criterion to be induced at inference time. To investigate this capability, we

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.

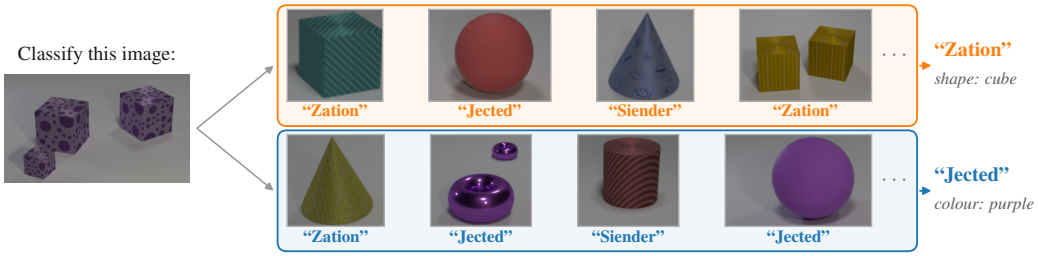


Figure 1: **Few-shot visual concept induction.** The same query image (left) can be classified based on different visual concepts (e.g. object shape or colour). The correct visual concept and class must be induced from in-context demonstrations. Demonstrations are labelled with nonsense words such that the concept cannot be induced from textual knowledge.

study a classification problem in which a model observes a set of images and must induce the visual concept they share, as illustrated in Fig. 1. This enables an intuitive form of visual learning, where the user demonstrates a visual concept without verbal description.

In this work, we investigate whether current MLLMs can induce visual concepts from context during MICL in a controlled setting, where images can be classified by multiple visual concepts, such as object shape, colour, count and texture. Our key insight is that understanding MICL requires *disentangling what information the model represents from what its frozen output head can read*. Although demonstrations influence a high-dimensional query representation, only a low-dimensional subspace spanned by differences between the label unembedding vectors of the output head can affect the classification decision. We therefore refer to this subspace as readout frame and analyse MICL within it. From this perspective, demonstrations can improve performance through two mechanisms, either increasing the *signal power* by making the demanded concept more separable, or by decreasing the *readout loss* by making that signal better aligned with the fixed output head.

Our analysis shows that conditioning the query representation on visual in-context demonstrations during MICL primarily improves the readability for the fixed language head, but may fail to provide sufficient signal power within the readout frame to separate classes. Importantly, demonstrations do not generally adapt the notion of similarity encoded in the query representation. We find that relevant information is already encoded in the query representation in the zero-shot setting, before any demonstrations are provided, and remains largely unchanged when demonstrations are introduced. Additionally, we show that undesired concepts co-occurring in the demonstrations manifest as interference in the readout frame, limiting the separability of the target concept. Suppressing this interference via a post-hoc intervention can improve concept induction by up to 11 pp for sufficient strongly represented concepts. Consequently, we conclude that MICL performance depends on both the *suppression of co-occurring concepts* and the *extent to which the desired concept-similarity criterion is already encoded in the unconditioned query representation*.

Our contributions can be summarised as follows:

- Our analysis disentangles what an MLLM represents from what its frozen output head can read during MICL;
- Through this decomposition we find that visual demonstrations leave the similarity information already encoded in the query representation largely unchanged;
- Undesired concepts co-occurring in visual demonstrations can interfere with target concept separation, but post-hoc interference suppression can improve concept induction accuracy up to 11 pp.

2 Related Work

Prior works on few-shot MICL mainly focus on measuring task performance and use standard few-shot benchmarks and semantic class names Agarwal et al. (2024); Baldassini et al. (2024); Höfer et al. (2025) where the MLLM can rely on textual knowledge or memorisation from pretraining, instead of inducing relevant visual concepts from context. (Tsimpoukelli et al., 2021) showed that

MICL can bind nonsense labels to image classes. In these tasks, however, each image has a single salient visual concept or information Chen et al. (2025), so the model only binds names to classes it already separates. We present the same images under four independent visual concepts, so the demonstrations must also fix the partition. Allen et al. (2025); Gavrikov et al. (2025) have shown an implicit bias inside MLLMs to generalise by object shape when provided with ambiguous visual demonstrations. We find that this implicit bias manifests as interference and actively harms visual concept induction, since MLLMs require a significant number of demonstrations to even weakly overwrite this bias. A similar task to our concept induction is proposed by Stracke et al. (2026), where a dedicated generative architecture is trained to generate new images from a small in-context set of images sharing a concept. In contrast, we ask whether general-purpose MLLMs can do visual concept induction in context, with no training. Here, demonstrations are labelled image–word pairs, so the visual concept must be induced from how the labels co-vary with the images across classes, rather than from a single concept the set holds constant, requiring the model to bind the classes to arbitrary nonsense words. In text-based ICL, prior works have studied mechanisms of how large language models (LLMs) internally perform few-shot learning. Yang et al. (2026a;b) attribute ICL to attention heads that align hidden states with the label-unembeddings. In contrast, we decompose task performance into signal power and readout loss inside the decision-relevant readout frame and a controlled setup in which we have information about confounding visual concepts.

3 METHOD

We present our approach for analysing few-shot learning in MICL as a prompt-conditioned classification problem. We first describe MICL as a frozen classifier whose decision function is conditioned on the demonstration set (Sec. 3.1), and define visual concepts as labelling functions over a shared image space (Sec. 3.2). We then formulate visual concept induction as the problem of adapting the decision function to the concept demonstrated by the in-context demonstrations (Sec. 3.3). Finally, we introduce the model's low-dimensional readout frame (Sec. 3.4), which enables us to analyse the adaptation by isolating the subspace of the query representation that can affect the decision.

3.1 MULTIMODAL IN-CONTEXT CLASSIFICATION

Let $\mathcal{X}$ be an image space and M a frozen MLLM with hidden dimension d, vocabulary unembedding matrix U, and L layers. A prompt consists of a demonstration set $\mathcal{D}$ of size n and a query $x_q \in \mathcal{X}$, where

$$\mathcal{D} = (x_i, \ell_i)_{i=1}^{n}, \qquad x_i \in \mathcal{X},\ \ell_i \in \mathcal{L}, \tag{1}$$

and $\mathcal{L} = \{\ell_1, \ldots, \ell_C\}$ is a set of C label words. The model encodes the prompt and produces a hidden state $h = M(\mathcal{D}, x_q) \in \mathbb{R}^d$ at the answer position in layer L. Restricting the prediction to the label vocabulary, the decision is

$$\hat{y}(\mathcal{D}, x_q) = \arg\max_{c \in [C]} u_c^\top h, \tag{2}$$

where $u_c \in \mathbb{R}^d$ is the unembedding row corresponding to the first token of label ℓ_c. For simplicity, we assume that the label words are chosen such that their first tokens uniquely identify the corresponding classes. All model parameters, including the unembedding rows $\{u_c\}_{c=1}^{C}$, are fixed and the demonstration set is the only free input. Thus, MICL can be viewed as a prompt-conditioned classifier: a map $\mathcal{D} \mapsto \hat{y}(\mathcal{D}, \cdot)$ from demonstration sets to decision functions on $\mathcal{X}$.

3.2 VISUAL CONCEPTS

A concept is a labelling function $t : \mathcal{X} \to [C]$ that partitions the image space into C classes. We consider a family $\mathcal{T} = t^{(1)}, \ldots, t^{(m)}$ of m concepts defined over the same images, representing different ways of classifying them, such as for example by object shape or colour.

Let $t^* \in \mathcal{T}$ be a target concept. Given a class-to-word assignment $\pi : [C] \to \mathcal{L}$, we construct a demonstration set by labelling images according to t^*:

$$\mathcal{D}_{t^*} = (x_i,\ \pi(t^*(x_i)))_{i=1}^{n}. \tag{3}$$

We assume that $\mathcal{D}_{t^*}$ uniquely identifies the target concept t^* within $\mathcal{T}$, meaning that no other concept $t' \neq t^*$, together with any alternative class-to-word assignment π', produces exactly the same labels

on all demonstration images. Intuitively, the demonstrations should contain enough variation to distinguish the target concept from the other concepts.

3.3 Visual Concept Induction

The model M succeeds at visual concept induction if it predicts the correct label ℓ_c for held-out queries x_q in an episode. Here, an episode is defined by a target concept t^*, a class-to-word assignment π, a demonstration set $\mathcal{D}_{t^*}$ generated from them, and a query set. Performing visual concept induction requires two steps: identifying the target partition t^* that generated the demonstrated pairing in $\mathcal{D}_{t^*}$, and binding its C classes to the words $\mathcal{L}$ via π. Based on these two steps, we distinguish two settings by the choice of $\mathcal{L}$:

Concept retrieval: The labels are natural words that are semantically associated with the classes (e.g. colour names for a colour concept). Therefore, the demonstrated pairing can be interpreted using prior semantic associations between the labels and concepts. Demonstrations primarily serve to select among concepts that the model already associates with the label words.

Concept induction: The label words are novel nonsense words with no prior association to any class. The demonstration pairing in Eq. 3 is then the only source of information about both the target concept t^* and the class-to-word mapping π. Therefore, the required concept cannot be determined from the query image, the label semantics, or the model's prior associations alone. It must be inferred from the demonstrations.

3.4 Readout Frame

The model does not use all dimensions of the high-dimensional hidden state h to make its decision. Recall from Eq. 2 that the decision is determined by the label logits $u_c^\top h$. Consequently, changes to h in directions orthogonal to all unembedding vectors u_c have no effect on the decision. Likewise, adding the same offset to all label logits does not change their relative ordering, so the component of h that shifts all logits equally is also irrelevant. This means that only a low-dimensional subspace of the hidden state is actually visible to the final decision. Let $\bar{u} = \frac{1}{C}\sum_c u_c$ and let $Q \in \mathbb{R}^{d\times r}$ be an orthonormal basis for the subspace spanned by the centred unembedding vectors $\{u_c - \bar{u}\}_{c=1}^{C}$, where $r \leq C-1$. We define the corresponding coordinates of the hidden state as

$$z = Q^\top h \in \mathbb{R}^r, \tag{4}$$

where we call $\mathrm{span}(Q)$ the *readout frame*. Intuitively, the readout frame is the part of the hidden state that the label unembedding rows can actually see. Two hidden states with the same z produce the same decision, regardless of how they differ outside this subspace. Thus, z is an exact sufficient representation for the model's decision. Any change to the hidden state outside the readout frame is invisible to the prediction, and any effect of the demonstrations on the final decision must therefore appear as a change in z. Our analysis therefore studies how demonstrations influence the geometry of this low-dimensional space.

4 Analysis

In this section, we first investigate visual concept induction by performing MICL with both semantically meaningful and nonsense labels (Sec. 4.1). Next, to determine why concept induction succeeds or fails, we decompose MICL decision quality into the *signal power* available in the readout frame and how *readable* that signal is by the models' frozen output head (Sec. 4.2). Finally, we examine the limitations of concept induction by quantifying how strongly the decisions are affected by interference from other visual concepts and demonstrate that reducing this interference can improve visual concept induction (Sec. 4.3).

Experimental Setup. We conduct our analysis with five modern MLLMs: Qwen3-VL (4B, 32B) Bai et al. (2025), Gemma3 (12B, 27B) Team et al. (2025) and GPT-5 mini Singh et al. (2025) as a closed-source reference. Unless specified, we report results on the Qwen3-VL family, due to its larger context window (256k vs. 128k tokens). The results of other models can be found in App. D.

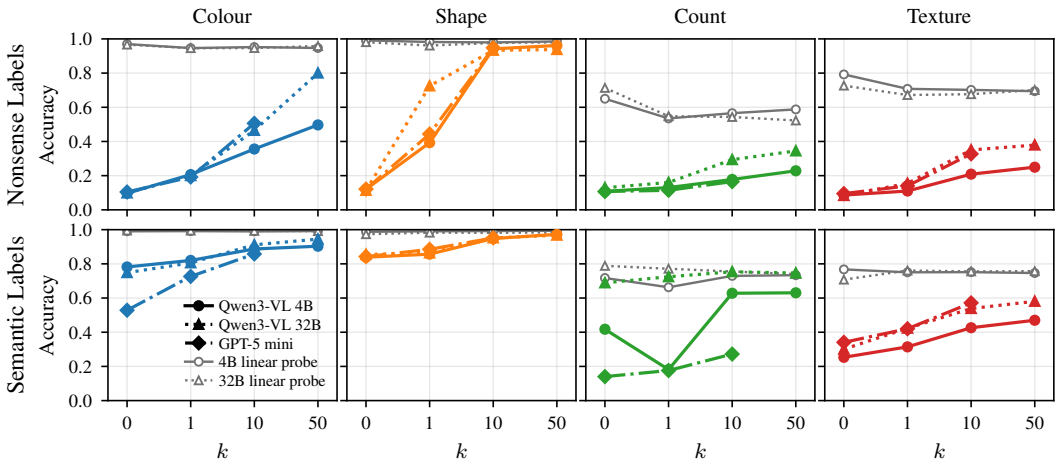


Figure 2: **MICL performance on Clevr-4.** Quantitative comparison of model scale (4B vs. 32B) across all four visual concepts (Colour, Shape, Count, Texture) using nonsense labels (top) and semantic labels (bottom), for $k \in \{0, 1, 10, 50\}$ shots per class. Grey curves show the accuracy of a linear probe fit on the models' full hidden state. GPT-5 mini ($k \leq 10$) is included for comparison. We report the mean accuracy over three episodes with randomly sampled demonstration sets.

We investigate visual concept induction using the Clevr-4 dataset Vaze et al. (2023b). It provides four equally valid ways of classifying the same images based on object Texture, Shape, Colour, and Count, with these four visual concepts being statistically independent. Each concept consists of 10 classes, corresponding to 10 possible values of the respective attribute. To construct the demonstration sets, we randomly sample k demonstrations per class from the training split and use the test split as the query set. Where necessary, we further divide the query set into training and test subsets, which we use to fit linear probes or estimate the quantities required for our analysis. For details regarding the prompt template and implementation, please refer to App. B.

4.1 Visual Concept Induction

We consider two different few-shot settings in which the same images must be classified according to four different visual concepts, using only k demonstrations per class. First, we consider a semantic setting, in which semantically meaningful class labels (e.g. Blue, Brown, ..., Yellow) are provided alongside the in-context demonstrations. Second, we consider an anonymised setting, in which the demonstrations are labelled only with meaningless nonsense words[1] (e.g. Zation, Siender, ..., Molonged), preventing the model from relying on textual knowledge Wang et al. (2026). Because the same query image can be evaluated under different target concepts, and because these concepts in Clevr-4 are independent, we can test whether the model conditions its predictions on the demonstrations. This setup enables us to isolate the models' ability to learn from visual demonstrations alone, since the MLLM must induce the relevant visual concept and bind the corresponding classes to the provided nonsense labels. Furthermore, because the output head is essentially a frozen linear classifier applied to the models' hidden state, we compare the accuracy achieved by MICL to a freely fitted linear classifier on the same hidden states.

Nonsense labels. In the nonsense-label condition, Fig. 2 shows a strong asymmetry in concept induction. The accuracy at $k = 0$ is near chance, confirming that the labels themselves provide no information about the four visual concepts. As the number of demonstrations increases, performance consistently follows the ordering Shape $>$ Colour $>$ Texture $\approx$ Count, with model scale substantially improving induction. Notably, performance on Shape is superior in all respects. It reaches near-ceiling accuracy with few demonstrations and at $k = 50$ accuracy is only marginally lower then when using semantic labels. The remaining concepts fall into two categories. Colour is inducible, but requires a substantial number of demonstrations. On the other hand, Texture and

[1]Taken from a nonsense-word generator: `https://www.soybomb.com/tricks/words/`

Count saturate substantially lower, only weakly improving accuracy, when increasing k from 10 to 50. Notably, GPT-5 mini shows a similar behaviour, suggesting that even closed-source models do not easily solve visual context induction. These findings show that while concept induction is possible in principle, *MLLMs struggle to classify the same images solely based on different visual concepts demonstrated in-context*.

Semantic labels. With semantic labels the accuracy at $k = 0$ is already high, indicating that the semantic class names alone induce the visual concept. Interestingly, performance increases further with more demonstrations[2], helping to refine the predictions. The gap between model scales is reduced, showing that semantic names enable the smaller model to perform similarly across conditions when the concept does not need to be induced from visual demonstrations alone. For Texture, we find that both models systematically exploit semantic textual knowledge, rather than following the demonstrated visual concept, confusion matrices are provided in App. D. In particular, at $k = 50$, they classify 72% (4B) and 63% (32B) of star-shaped objects whose true texture is not *star* as having texture *star*, see App. B for qualitative examples of the visual concepts. Thus, even 50 demonstrations of the *star* Texture do not reliably override the Shape-based semantic association. Additionally, both models struggle to separate the *chessboard* Texture from the coarser *checkered* Texture, indicating that concept induction is limited by the fidelity of the underlying representations. In summary, these findings show that semantic labels provide a better starting point, *but can also induce undesired semantic association, which visual demonstrations can only weakly override.*

Linear probing. The linear probing accuracy at $k = 0$ is much higher than the MICL accuracy. Notably, for the nonsense labels, all four visual concepts are decodable from the same hidden state, *indicating that information about all four visual concepts is encoded before the model even knows what the labels mean*. Moreover, with semantic labels the decodability of Count improves relative to the nonsense label setting, suggesting that textual cues can add decodable information to the hidden state. Interestingly, linear probing accuracy is flat over k and on some concepts even decreasing, while MICL performance is significantly improving with more demonstrations. In the nonsense label setting, only MICL on Shape is able to close the gap to the probing accuracy with sufficiently many demonstrations. To understand this gap better, we will next make use of the fact that the models' decision only depends on the hidden state through its projection onto the readout frame (Sec. 3.4).

4.2 From Demonstrations to Decisions

We analyse how demonstrations influence the models' decisions and quantify MICL decision quality. We begin by deriving a pairwise sufficient statistic Picinbono (1995); Bishop & Nasrabadi (2006) for each class pair and decomposing it into two terms that quantify class separability and alignment with the frozen output head. This decomposition allows us to isolate and distinguish different failure modes of MICL. All metrics in this section are derived from the class-conditional first- and second-order moments of the readout-frame variable z defined in Eq. 4. For a target concept t^*, we denote the class-conditional means by $\mu_c = \mathbb{E}[z \mid t^*(x_q) = c]\ \forall c$ and the average within-class covariance of z by Σ. Both quantities are estimated from a pool of queries under a fixed demonstration set.

Signal power and readability. For any pair of classes (c, c'), the output head's choice between labels ℓ_c and $\ell_{c'}$ is determined by the sign of the scalar statistic

$$y_{cc'} = (u_c - u_{c'})^\top h = \delta^\top z, \qquad \delta = Q^\top (u_c - u_{c'}), \tag{5}$$

where the second equality holds because $u_c - u_{c'}$ lies in the readout frame by construction. How well this statistic discriminates c and c' can be quantified by

$$d^2_{cc'} = \frac{\left(\delta^\top (\mu_c - \mu_{c'})\right)^2}{\delta^\top \Sigma \delta}. \tag{6}$$

Intuitively, this so-called deflection of $y_{cc'}$ measures the squared distance between the class-conditional means in units of its within-class standard deviation. Finally, after introducing the whitened signal and the whitened readout direction,

$$v = \Sigma^{-1/2}(\mu_c - \mu_{c'}), \qquad r = \Sigma^{1/2}\delta, \tag{7}$$

[2]The 4B model's Count accuracy drops significantly at $k = 1$, with 88–94% of predictions collapsing onto a single Count word within each episode, although the dominant word differs across episodes.

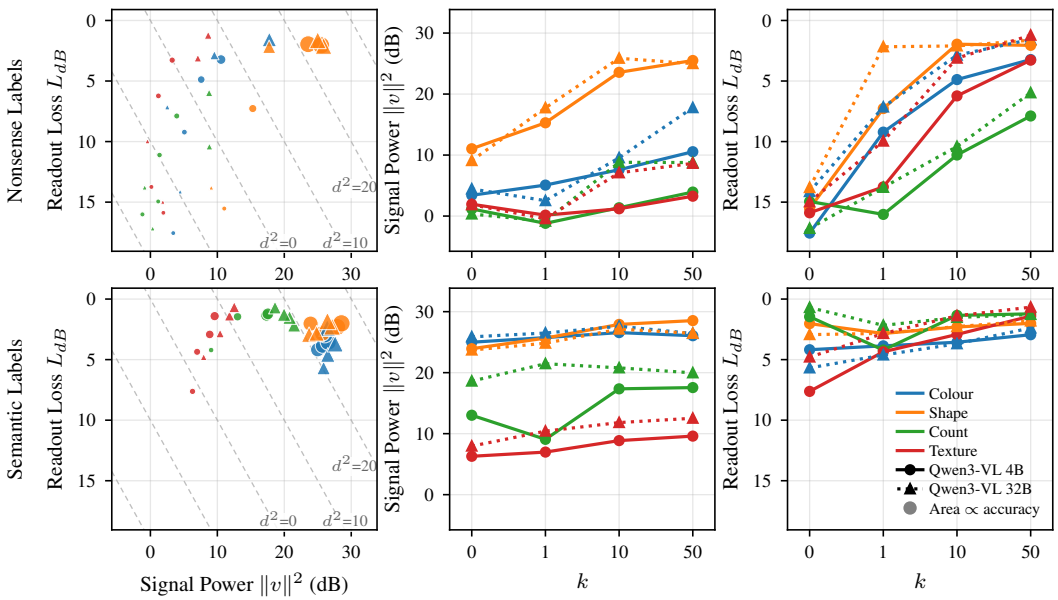


Figure 3: **Effect of demonstrations on model decisions.** Quantitative comparison of model scale (4B vs. 32B) across all four visual concepts (Colour, Shape, Count, and Texture), using nonsense labels (top) and semantic labels (bottom). The left column shows the two-factor plane of signal power $\|\boldsymbol{v}\|^2$ versus readout loss L_{dB}, with iso-d^2 lines (in dB) and marker size proportional to accuracy. The middle and right columns show signal power and readout loss, respectively, as a function of the number of shots per class, $k \in \{0, 1, 10, 50\}$.

the deflection can be factorised exactly:

$$d^2_{cc'} = \frac{(r^\top v)^2}{\|r\|^2} = \underbrace{\|v\|^2}_{\text{signal power}} \cdot \underbrace{\cos^2 \angle(r, v)}_{\text{readability } \eta}. \tag{8}$$

The two factors answer different questions. Signal power $\|v\|^2 = (\mu_c - \mu_{c'})^\top \Sigma^{-1} (\mu_c - \mu_{c'})$ is the pair's squared Mahalanobis separation, the largest deflection any linear statistic on z can achieve, attained by $\delta \propto \Sigma^{-1}(\mu_c - \mu_{c'})$. It measures how much discriminative signal the state carries, independent of how it is read. Readability $\eta \in (0, 1]$ measures how well the head's fixed readout is oriented relative to that optimum.

Readout loss. Taking logarithms of Eq. 8 yields an easily interpretable additive decomposition,

$$10 \log_{10} d^2 = 10 \log_{10} \|v\|^2 - L_{\mathrm{dB}}, \qquad L_{\mathrm{dB}} = -10 \log_{10} \eta \geq 0, \tag{9}$$

where we call L_{dB} the *readout loss*. It measures the gap in dB between the separation achieved by the frozen readout and the maximum separation attainable by any linear readout of the same representation. Since a C-class episode consists of $\binom{C}{2}$ pairwise signal-loss decompositions, we summarise the episode-level decomposition by averaging Eq. 9 over all class pairs.[3] Demonstrations can affect either term in Eq. 9: they may increase the signal power by making the demanded concept more separable, or decrease the readout loss by making that signal better aligned with the fixed readout.

Nonsense labels. Fig. 3 shows the decomposition applied to every episode from Sec. 4.1. In the nonsense label setting, at $k = 0$, the signal power $\|v\|^2$ is close to 0 dB for most concepts, indicating that the classes are not well separated within the readout frame. Notably, Shape is an

[3] Arithmetic averaging the logarithmic quantities is equivalent to taking the geometric mean of the corresponding linear quantities, which is more sensitive to low efficiencies of individual class pairs than the arithmetic mean.

exception again, as it already has non-zero signal power at $k = 0$, before the context can induce any concept. This highlights that the models' have an implicit bias toward classifying objects by Shape. However, as the models cannot know which label corresponds to which class, the readout loss is high for all concepts. This demonstrates that the frozen head does not utilise the separation available in the frame. As shown in the left panel of Fig. 3, both terms change with increasing number of demonstrations, moving episodes from the lower left towards the upper right across the iso-d^2 contours. The marker areas further confirm that d^2 tracks accuracy. Interestingly, however, the two terms do not necessarily move by the same amount. Demonstrations reliably reduce the readout loss, which reaches similarly low values at $k = 50$ under both nonsense and semantic labels. In contrast, the signal power lags behind. This difference is particularly visible for the 32B model on Colour and Texture. Here, although the readout loss becomes as low as that for Shape, signal power remains substantially lower, explaining the lower accuracy in Fig. 2. In the case of Texture and Count further demonstrations do not help improve significantly performance, since they primarily reduce readout loss but do not improve separability.

Semantic labels. In the semantic setting, we similarly observe Texture and Count being almost perfectly readable, but lacking in signal power, indicating that the lower performance of both in Sec 4.1 comes from a lack of separability. Showing that decodability in the full hidden state does not imply separability in the decision relevant readout frame. Furthermore, we observe that signal power remains mostly flat over k, demonstrating that semantic labels make the visual concept better separable already in the zero-shot setting and demonstrations mainly reduce the readout loss.

Findings. In summary, in-context demonstrations can effectively reduce the readout loss, but leave a gap in signal power. Together with the linear probing results in Sec. 3.3 these findings show that *visual demonstrations do not improve separability in the hidden state, but steer signal into the readout frame*. The readout loss is the main quantity that MICL successfully reduces, but the big differences across the visual concepts show in the signal power. This shows that while classes are in principle separable in the full hidden state, the model struggles to separate the classes inside the readout frame. Demonstrations reliably make the concepts readable for the output head, but far less reliably make the target concept separable in the first place. The most limiting factor for visual concept induction during MICL is signal power in the readout frame, which will be investigated in the next section.

4.3 Alternative Concepts act as Interferers

The within-class covariance Σ in Eq. 6 has been treated as noise, representing the unexplained variance around the class means. However, since Clevr-4 provides labels for all four visual concepts of each image, we can determine how much of this apparent noise is in fact *interference*, i.e. structured variance caused by the class structure of a competing concept. In App. C we provide a detailed derivation of the variance decomposition and show how the signal power $\|v\|^2$ can be further decomposed into interference-free signal power and interference loss. The interference-free signal power γ_N is the signal power achievable without interference from the competing concept and the interference loss $L_I = 10 \log_{10} \frac{\gamma_N}{\|v\|^2}$ is the loss due to that competitor. In Fig. 4 each curve is based on the dominant competitor, i.e. the one with the largest L_I. For Colour, Count, and Texture the strongest competitor is Shape, while for Shape it is Colour, which demonstrates the implicit bias to classify by Shape even when in-context demonstrations are provided. Additionally, Fig. 4 shows the normalised mutual information (NMI) between incorrect predictions of non-Shape concepts and Shape. This intuitively measures how much prediction errors are structured by Shape, showing that interference indeed leads to incorrect predictions.

Table 1: **LEACE erasure of the dominant interfering concept.** Each block gives baseline accuracy and the accuracy after erasure, averaged over three seeds with best highlighted in **bold**.

		4B, final layer		4B, interm. layer		32B, final layer	
Concept	k	Base	LEACE	Base	LEACE	Base	LEACE
Colour	1	0.21	**0.24**	0.21	**0.25**	0.20	**0.27**
	10	0.36	**0.47**	0.36	**0.51**	0.47	**0.54**
	50	0.50	**0.56**	0.50	**0.58**	0.80	**0.81**
Shape	1	**0.40**	**0.40**	**0.40**	**0.40**	0.73	**0.74**
	10	0.94	**0.95**	0.94	**0.95**	0.93	**0.94**
	50	**0.96**	**0.96**	0.96	**0.97**	**0.94**	**0.94**
Texture	1	**0.11**	**0.11**	**0.11**	0.10	0.16	**0.18**
	10	0.21	**0.23**	0.21	**0.23**	0.35	**0.40**
	50	0.25	**0.27**	0.25	**0.26**	0.38	**0.39**
Count	1	**0.13**	**0.13**	0.13	**0.14**	0.16	**0.17**
	10	**0.18**	**0.18**	0.18	**0.20**	0.29	**0.31**
	50	0.23	**0.25**	0.23	**0.25**	0.34	**0.36**

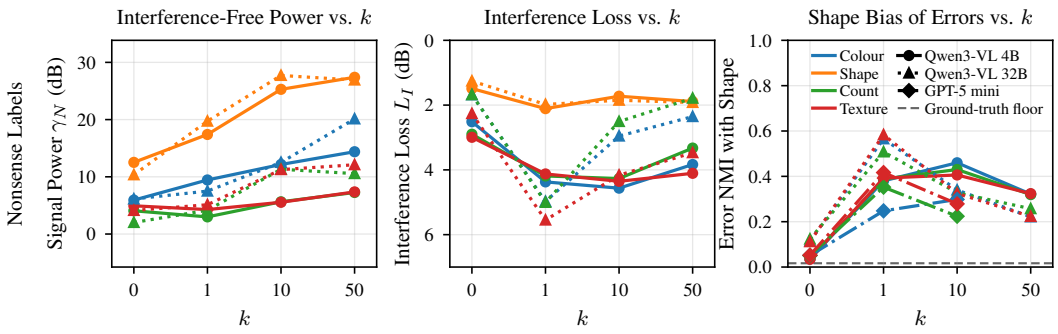


Figure 4: **Interference from the competing concept.** Quantitative comparison of interference under model scale (4B vs. 32B) across all four visual concepts (Colour, Shape, Count, and Texture), as a function of the number of shots per class, $k \in \{0, 1, 10, 50\}$. The left column shows the interference-free signal power γ_N (in dB) and the middle column shows the interference loss $L_I = 10 \log_{10}(\gamma_N / \|\boldsymbol{v}\|^2)$ (in dB). The right column shows the NMI between the model's incorrect predictions and the query's Shape. The dashed line marks the NMI between Shape and the true labels.

Notably, the larger model reaches the largest interference loss L_I at $k = 1$ and suppresses it more rapidly afterwards. This shows resemblance to a nearest-demonstration classifier strongly keyed on Shape. *The model simply copies the label of a demonstration with matching shape*. The decrease in L_I with k reflects the confound averaging out, as each label word accumulates demonstrations of many Shapes, such that the concept consistent for each label becomes the only one that remains.

Interference cancellation. We propose to use linear concept erasure (LEACE) Belrose et al. (2023) to remove the interfering concept from the query representation and report performance in Tab. 1. We find that removing interference improves concept induction. If the target concept has sufficient interference-free signal power, the improvement can be significant (e.g. Colour). Moreover, when LEACE is applied on an intermediate layer, improvements can be even bigger, details in App. D. This post-hoc intervention only requires labels of the dominant interferer (Shape in our case), but not the target labels. *Therefore, it may be possible to improve concept induction when information about the confounder is available*.

5 Conclusion

In this work, we investigated whether MICL enables MLLMs to induce visual concepts from demonstrations without explicitly naming them. By introducing the notion of a readout frame, we disentangled the visual information encoded in the query representation from the information that can be accessed by the frozen language output head. Our analysis shows that visual demonstrations primarily improve the readability of pre-existing visual signals rather than substantially changing what the query representation encodes. At the same time, co-occurring visual concepts, above all Shape, can introduce interference within the readout frame and thus limit the separability of the target concept. Suppressing this interference post hoc can improve concept induction by up to 11 pp for sufficiently strongly represented concepts. These findings suggest that the effectiveness of MICL in current MLLMs is constrained by the concepts already encoded in the unconditioned representation and by how well the model is able to suppress co-occurring concepts in the demonstrations.

AI use statement

We have not used generative AI tools for any of the tasks with required disclosure, including developing theoretical models or conceptual frameworks, formulating or proving mathematical claims, proposing hypotheses, designing research methodology or experiments, or interpreting results. The tasks of generating synthetic datasets, cleaning or reformatting datasets, translation, and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for three tasks with recommended disclosure: (i) polishing the writing, i.e. improving grammar,

clarity, and readability of text drafted by the authors, (ii) retrieval and discovery, i.e. searching for and identifying potentially relevant related work, and (iii) implementing methods, i.e. using coding assistance during the implementation of our evaluations. We have reviewed all AI-assisted work. All wording suggestions were checked by the authors to ensure they did not alter the intended technical meaning, and every reference surfaced with AI assistance was verified by an author to exist and to support the claim it is cited for. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, et al. Many-shot in-context learning. *Advances in Neural Information Processing Systems*, 37:76930–76966, 2024.

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. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Kelsey Allen, Ishita Dasgupta, Eliza Kosoy, and Andrew K Lampinen. The in-context inductive biases of vision-language models differ across modalities. *arXiv preprint arXiv:2502.01530*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. What makes multimodal in-context learning work? In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1539–1550. IEEE, 2024.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36:66044–66063, 2023.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Shuo Chen, Jianzhe Liu, Zhen Han, Yan Xia, Daniel Cremers, Philip Torr, Volker Tresp, and Jindong Gu. True multimodal in-context learning needs attention to the visual context. *arXiv preprint arXiv:2507.15807*, 2025.

Paul Gavrikov, Jovita Lukasik, Steffen Jung, Robert Geirhos, Muhammad Jehanzeb Mirza, Margret Keuper, and Janis Keuper. Can we talk models into seeing the world differently? In *International Conference on Learning Representations*, volume 2025, pp. 61715–61746, 2025.

Sebastian Höfer, Dorian F Henning, Artemij Amiranashvili, Douglas Morrison, Mariliza Tzes, Ingmar Posner, Marc Matvienko, Alessandro Rennola, and Anton Milan. Kaputt: A large-scale dataset for visual defect detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24224–24233. IEEE, 2025.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Bernard Picinbono. On deflection as a performance criterion in detection. *IEEE Transactions on Aerospace and Electronic Systems*, 31(3):1072–1081, 1995.

Jona Ruthardt, Manu Gaur, Deva Ramanan, Makarand Tapaswi, and Yuki M Asano. Steerable visual representations. *arXiv preprint arXiv:2604.02327*, 2026.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Nick Stracke, Kolja Bauer, Stefan Andreas Baumann, Miguel Angel Bautista, Josh Susskind, and Björn Ommer. Show me examples: Inferring visual concepts from image sets. *arXiv preprint arXiv:2607.02402*, 2026.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

Sagar Vaze, Nicolas Carion, and Ishan Misra. Genecis: A benchmark for general conditional image similarity. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6862–6872. IEEE, 2023a.

Sagar Vaze, Andrea Vedaldi, and Andrew Zisserman. No representation rules them all in category discovery. *Advances in Neural Information Processing Systems*, 36:19962–19989, 2023b.

Yunhan Wang, Eshika Khandelwal, Edson Araujo, Walid Bousselham, Nina Shvetsova, and Hilde Kuehne. Decompose, compare, and decide: Multimodal llms are implicit few-shot learners. *arXiv preprint arXiv:2607.00125*, 2026.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Haolin Yang, Hakaze Cho, and Naoya Inoue. Localizing task recognition and task learning in in-context learning via attention head analysis. In *International Conference on Learning Representations*, volume 2026, pp. 93780–93839, 2026a.

Haolin Yang, Hakaze Cho, Yiqiao Zhong, and Naoya Inoue. Unifying attention heads and task vectors via hidden state geometry in in-context learning. *Advances in Neural Information Processing Systems*, 38:21701–21759, 2026b.

A Appendix

The appendix is structured as follows. First, we provide additional implementation details of our analysis (App. B). Second, detailed derivation of the variance decomposition into noise and interference (App. C). Finally, we show complementary findings and additional results on the Gemma3 family (App. D).

B Implementation Details

We use Hugging Face `transformers` Wolf et al. (2019) with Qwen3-VL-Instruct (4B/32B) and Gemma 3 (12B/27B), in `bfloat16` precision with SDPA attention. Images are resized to 224×224. Every episode consists of prompts with three parts:

1. a system prompt listing the 10 candidate labels, as shown in B;
2. k support examples per class, drawn from the train split with a fixed seed and shown as image/label pairs;
3. the query image.

Table 2: **Label words used during MICL**

Type	Label Words
Semantic Shape	Cone, Cube, Tube, Diamond, Gear, Monkey, Sphere, Star, Teapot, Torus
Semantic Colour	Blue, Brown, Cyan, Gray, Green, Orange, Pink, Purple, Red, Yellow
Semantic Count	One, Two, Three, Four, Five, Six, Seven, Eight, Nine, Ten
Semantic Texture	Brick, Checkered, Chessboard, Circles, Emojis, Metal, Rubber, Star, Wave, Zigzag
Nonsense	Zation, Siender, Theously, Jected, Maniness, Inadul, Divills, Squenter, Pytheon, Molonged

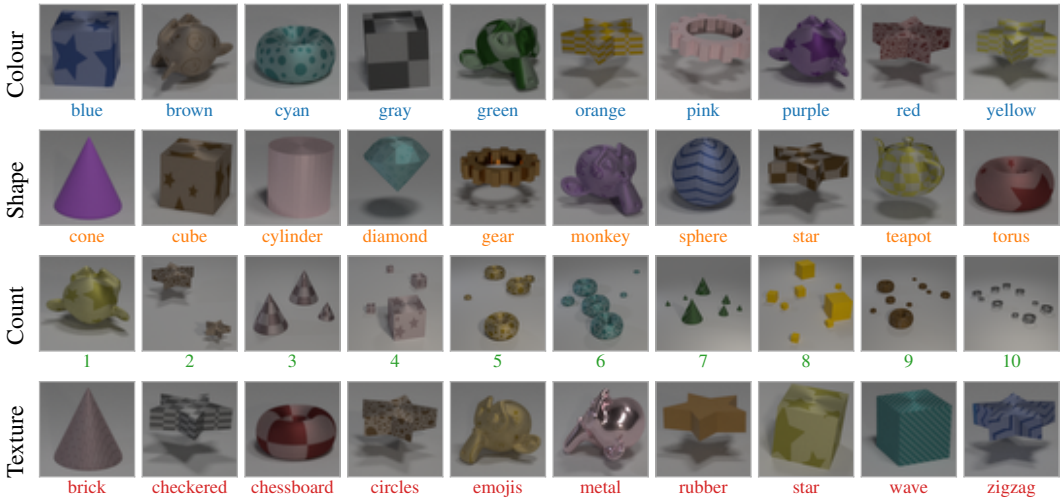


Figure 5: **Examples of each visual concept and corresponding classes in Clevr-4.**

Queries are taken from the validation split, which we split 50/50 into probe-train and probe-test halves. The system prompt and support examples are shared by every query in an episode, so we use KV caching to encode them only once. For each query we then run only its own tokens through the model, with episode's KV cache. At the final query token we record the hidden state of every transformer layer. The model's prediction is the label whose first answer token scores highest under the frozen output head. Linear probes are a standardised logistic regression (`scikit-learn` Pedregosa et al. (2011), `max_iter=1000`) fit on the train half and evaluated on the test half.

We query GPT-5 mini through the OpenAI Responses API with reasoning effort set to `minimal` and a 2,048-token output limit.

System Prompt

```
You are a helpful assistant.  Your task is to classify images into
one of these categories:  <LABELS>.  Based on the support examples
provided, respond with the single-word label that best describes the
query image.
```

Details on LEACE erasure. The eraser is fit on the train half and evaluated on the test half. In Tab. 1 we also report performing LEACE as an intervention during the forward pass on an intermediate layer. The reported performance corresponds to layer 23 of Qwen-3 VL which we found to perform best during a layer sweep.

C Theoretical Background

Interference and residual noise. The within-class covariance Σ averages two different sources of variability. First, the structure that is explained by a competing concept t' (e.g. shape when colour is

demanded), and second, residual variability that neither labelled concept accounts for. To separate them, we further condition on the competitor and denote the subgroup means by $\mu_{cg} = \mathbb{E}[z \mid t^*(x_q) = c,\, t'(x_q) = g]$, estimated from the n_{cg} queries of subgroup (c, g), with $n = \sum_{c,g} n_{cg}$. An analysis of variance then splits the average within-class variance exactly into a between-concept and a within-concept term,

$$\Sigma = \underbrace{\frac{1}{n-C}\sum_{c}\sum_{g} n_{cg}\,(\mu_{cg}-\mu_c)(\mu_{cg}-\mu_c)^\top}_{\text{interference } \Sigma_I} + \underbrace{\frac{1}{n-C}\sum_{c}\sum_{g}\sum_{i\in(c,g)} (z_i-\mu_{cg})(z_i-\mu_{cg})^\top}_{\text{residual noise } \Sigma_N}, \tag{10}$$

where the cross terms vanish, because residuals are centred within each subgroup. Σ_I collects the part of the within-class spread that is predictable from the competitor's label, while Σ_N is what remains once both concepts are fixed. Since the split is linear in Σ, it carries over directly to the denominator of the deflection in Eq. 6,

$$d^2_{cc'} = \frac{\left(\delta^\top(\mu_c - \mu_{c'})\right)^2}{\delta^\top \Sigma_I\, \delta + \delta^\top \Sigma_N\, \delta}, \tag{11}$$

so that the error along the decision-relevant direction δ is resolved into an interference term $I = \delta^\top \Sigma_I \delta$ and a noise term $N = \delta^\top \Sigma_N \delta$.

Interference-free signal power and interference loss. We define the interference-free signal power, i.e. what a receiver would see if the competing concept were known and cancelled:

$$\gamma_N = \Delta\mu^\top \Sigma_N^{-1} \Delta\mu. \tag{12}$$

This lets you define an interference loss

$$L_I = 10 \log_{10} \frac{\gamma_N}{\|v\|^2} \;\geq\; 0, \tag{13}$$

which extends our decomposition from Eq. 9:

$$10 \log_{10} d^2_{cc'} = 10 \log_{10} \gamma_N \;-\; L_I \;-\; L_{\mathrm{dB}}. \tag{14}$$

If γ_N itself is low, the target concept is weakly encoded in the readout frame, and the problem is noise and weak signal. Therefore, if γ_N is strong but L_I is large, competing concepts are masking it.

D Additional Results

Table 3: **Gemma3 LEACE erasure of the dominant interfering concept.** Each block gives baseline accuracy and the accuracy after erasure (nonsense labels), averaged over three seeds with best highlighted in **bold**.

		Gemma3 12B		Gemma3 27B	
Concept	k	base	LEACE	base	LEACE
Colour	10	0.63	**0.68**	0.25	**0.29**
Shape	10	**0.90**	**0.90**	**0.95**	**0.95**
Texture	10	**0.20**	0.19	0.19	**0.21**
Count	10	**0.23**	**0.23**	0.31	**0.33**

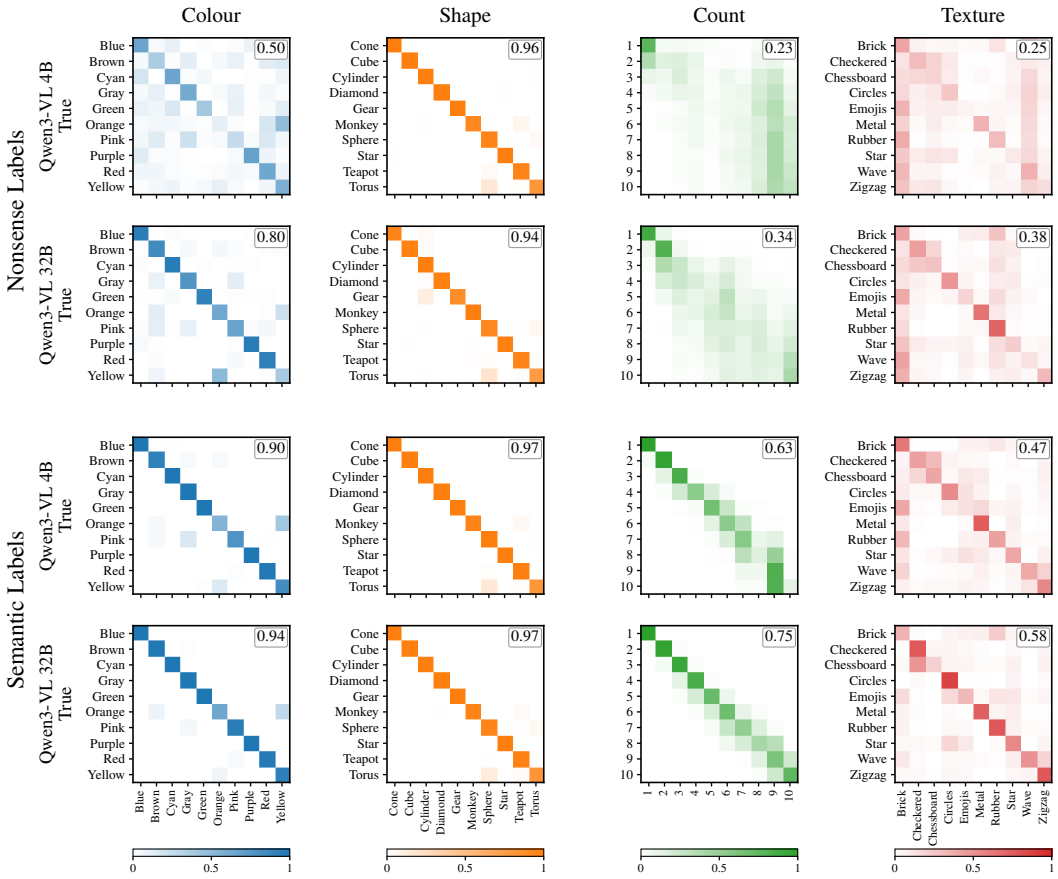


Figure 6: **Confusion matrices on Clevr-4.** Quantitative comparison of Qwen3-VL model scale (4B vs. 32B) across all four visual concepts (Colour, Shape, Count, Texture) using nonsense labels (top) and semantic labels (bottom), for $k = 50$ shots per class.

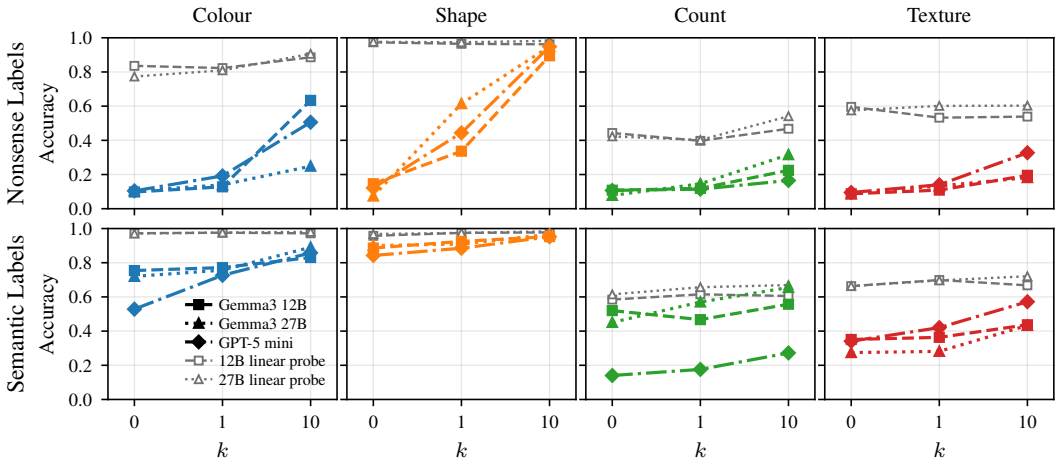


Figure 7: **Gemma3 MICL performance on Clevr-4.** Quantitative comparison of model scale (12B vs. 27B) across all four visual concepts (Colour, Shape, Count, Texture) using nonsense labels (top) and semantic labels (bottom), for $k \in \{0, 1, 10\}$ shots per class. Grey curves show the accuracy of a linear probe fit on the models' full hidden state. GPT-5 mini is included as a reference. We report the mean accuracy over three episodes with randomly sampled demonstration sets.

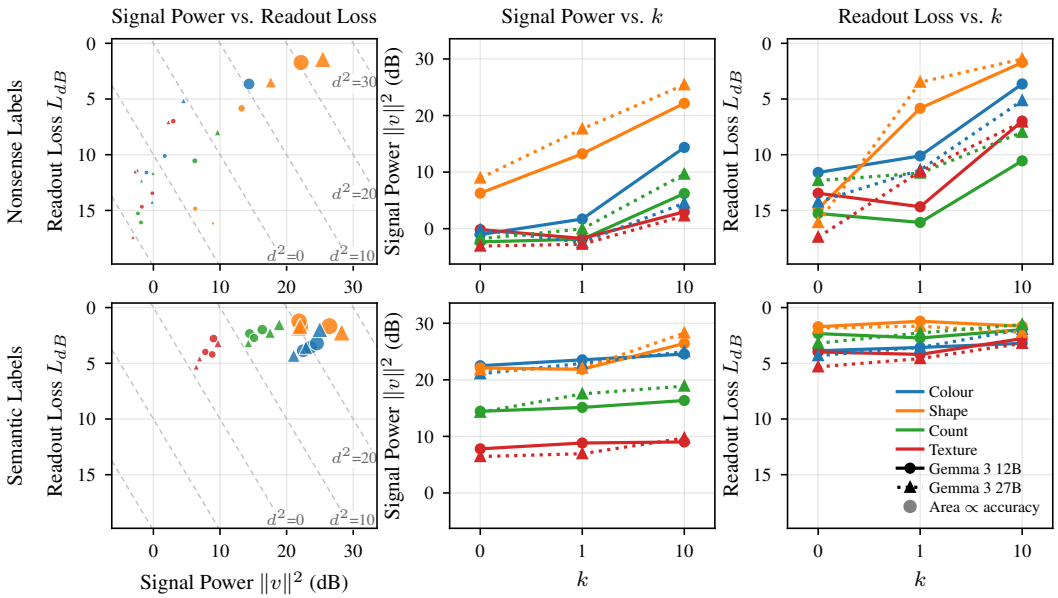


Figure 8: **Effect of demonstrations on Gemma3 decisions.** Quantitative comparison of model scale (12B vs. 27B) across all four visual concepts (Colour, Shape, Count, and Texture), using nonsense labels (top) and semantic labels (bottom). The left column shows the two-factor plane of signal power $\|\boldsymbol{v}\|^2$ versus readout loss L_{dB}, with iso-d^2 lines (in dB) and marker size proportional to accuracy. The middle and right columns show signal power and readout loss, respectively, as a function of the number of shots per class, $k \in \{0, 1, 10\}$.

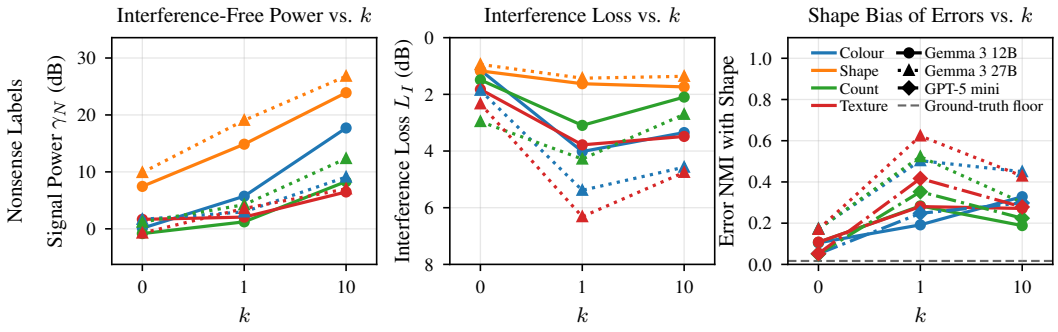


Figure 9: **Interference from the competing concept in Gemma3.** Quantitative comparison of interference under model scale (12B vs. 27B) across all four visual concepts (Colour, Shape, Count, and Texture), as a function of the number of shots per class, $k \in \{0, 1, 10, 50\}$. The left column shows the interference-free signal power γ_N (in dB) and the middle column shows the interference loss $L_I = 10 \log_{10}(\gamma_N / \|\boldsymbol{v}\|^2)$ (in dB). The right column shows the NMI between the model's incorrect predictions and the query's Shape, with GPT-5 mini ($k \leq 10$) for comparison. The dashed line marks the NMI between Shape and the true labels.