Pixels Versus Priors: Controlling Knowledge Priors in Vision-Language Models through Visual Counterfacts

Michal Golovanevsky[*1], **William Rudman**[*1], **Michael Lepori** [1], **Amir Bar**[2],
Ritambhara Singh[1], **Carsten Eickhoff**[3]
[1]Brown University, [2]Tel Aviv University, [3]University of Tübingen
{michal_golovanevsky, william_rudman}@brown.edu

Abstract

Multimodal Large Language Models (MLLMs) perform well on tasks such as visual question answering, but it remains unclear whether their reasoning relies more on memorized world knowledge or on the visual information present in the input image. To investigate this, we introduce Visual CounterFact, a new dataset of visually-realistic counterfactuals that put world knowledge priors (e.g, red strawberry) into direct conflict with visual input (e.g, blue strawberry). Using Visual CounterFact, we show that model predictions initially reflect memorized priors, but shift toward visual evidence in mid-to-late layers. This dynamic reveals a competition between the two modalities, with visual input ultimately overriding priors during evaluation. To control this behavior, we propose *Pixels Versus Priors* (PvP) steering vectors, a mechanism for controlling model outputs toward either world knowledge or visual input through activation-level interventions. On average, PvP successfully shifts 99.3% of color and 80.8% of size predictions from priors to counterfactuals. Together, these findings offer new tools for interpreting and controlling factual behavior in multimodal models. [*] [†]
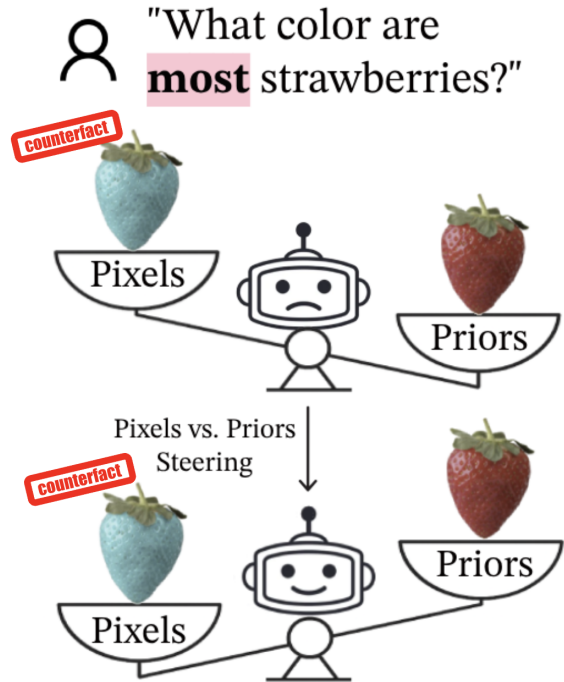
Figure 1: **Pixels Versus Priors Steering.** We introduce a framework for controlling whether a vision-language model relies on visual input or memorized knowledge. Counterfactual visual evidence often overrides world knowledge priors.

1 Introduction

As multimodal large language models (MLLMs) demonstrate increasing success in real-world vision-language tasks (Li et al., 2024; Wang et al., 2024; Chen et al., 2025), it is becoming increasingly important to understand their internal mechanisms in order to ensure the reliability and safety of these systems (Golovanevsky et al., 2025; Jiang et al., 2024; Luo et al., 2024; Rudman et al., 2025). Despite recent advances, interpretability in MLLMs remains underdeveloped compared to progress in natural language processing (NLP), where the encoding of world knowledge facts is

well-researched and methods exist for systematically editing factual associations (Meng et al., 2022). In NLP, *counterfactual datasets* consist of minimally altered input pairs that isolate specific factual changes, such as swapping one entity or relation while holding others constant. These datasets enable causal analysis of model behavior and have been central to understanding how factual associations are stored, retrieved, and manipulated (Geva et al., 2020, 2023; Dai et al., 2021; Yu et al., 2023; Meng et al., 2022). Unlike language, where factual associations are well understood, there is no visual equivalent for locating or modifying stored associations in MLLMs. In particular, there is no

---

24837

counterfactual dataset for testing how these models balance visual perception against memorized priors, nor any method for controlling their responses when the two sources of information conflict. To address this gap, we introduce Visual CounterFact, the first dataset designed to study world knowledge priors related to visual attributes in MLLMs, and use it to develop *Pixels Versus Priors* steering (PvP), a method for controlling whether the model relies on pixel-level information or on world knowledge.

Visual CounterFact modifies visual attributes, color and size, of everyday objects to create direct conflicts between memorized facts and input pixels. In our framework, *world knowledge priors* refer to linguistic associations between visual attributes and objects that the model has memorized during pretraining. In contrast, *visual perception* is defined by the in-context visual input, which we manipulate to create counterfactual images. These counterfactuals are designed to challenge the model's world knowledge of visual attributes by presenting plausible but contradictory visual evidence. For example, we contrast the size-related knowledge prior "strawberries are bigger than flies" with the counterfactual "flies are bigger than strawberries," violating expected size relations (see Figure 2).

Using Visual CounterFact, we find that MLLMs often ignore world knowledge when shown counterfactual images, favoring perceptual input even when prompted for general facts. We then trace where in the forward pass predictions shift from in-weight knowledge (e.g., strawberries are red) to in-context perception (*this* strawberry is blue), finding that this transition consistently emerges in mid-to-late layers. During this transition, models frequently flip between the two answers, revealing a competition between in-context pixel and in-weights prior information, with pixels often overriding priors in the model's output. To control this behavior, we use our *Pixels Versus Priors* steering to control whether a vision-language model relies on knowledge priors or pixel information. PvP is a novel framework to construct steering vectors for vision-language models that control whether the model responds based on memorized knowledge or in-context visual input. Through this steering, we successfully shift an average of 99.3% of color predictions and 80.8% of size predictions from memorized priors to counterfactual answers.

Together, these contributions provide a new visual counterfactual benchmark and a mechanism for interpreting and controlling the behavior of vision-language models. We present the necessary foundation for a mechanistic understanding of how MLLMs integrate image input with prior knowledge of visual attributes, bridging the gap between interpretability research in language models and the emerging needs of multimodal models.

2 Related Works

Studies in mechanistic interpretability have shown that LLMs encode factual associations grounded in world knowledge within their weights, enabling precise manipulation through targeted interventions. In particular, feedforward layers often act as key-value memories, injecting factual knowledge into subject representations (Geva et al., 2020, 2023), while clusters of "knowledge neurons" have been shown to store and control specific facts (Dai et al., 2021; Yu et al., 2023). These internal representations can be edited by introducing counterfactuals through weight-level interventions(Meng et al., 2022), or by tracing how attention mechanisms recover or suppress modified content during inference (Jin et al., 2024). More recently, activation-level interventions have emerged as an alternative to weight-level editing. In NLP, *steering vectors* are computed by subtracting internal representations from contrasting prompts to isolate meaningful activation directions (Subramani et al., 2022; Turner et al., 2023). These directions can be added at inference to shift a model's behavior without altering the model's weights.

Although model editing and steering have been successful in shifting model outputs, Gekhman et al. (2025) show that even when a model produces incorrect outputs, it may still internally represent the correct fact, highlighting a disconnect between stored knowledge and in-context generation. Investigating this disconnect further, there is a growing body of NLP literature that seeks to understand how language models flexibly deploy both in-context and memorized in-weight knowledge (Chan et al., 2022; Singh et al., 2023; Anand et al., 2025; Reddy; Lampinen et al., 2024; Zucchet et al., 2025; Park et al., 2024). These studies suggest that models often switch between relying on memory and adapting to context, depending on training dynamics and task structure. This inherent conflict *within*-modality motivates our mechanistic analysis *across*-modalities.

In order to study the conflict of world knowledge

priors across vision and language, a visual counterfactual dataset is needed. While many benchmarks test visual-textual alignment, none directly evaluate a model's reliance on visual world knowledge. VL-Checklist (Zhao et al., 2022) and VALSE (Parcalabescu et al., 2021) vary captions over fixed images, while FOIL-COCO (Shekhar et al., 2017), SVO-Probes (Hendricks and Nematzadeh, 2021), and Winoground (Thrush et al., 2022) alter semantic content in real images. However, these methods often suffer from uncontrolled visual artifacts. COCO-Counterfactuals (Le et al., 2023) uses generative models to edit images to replace a single object in the image with a new object. For example, for an image reflecting the caption "A large black ball sitting next to a glass of **milk**", they generate a "counterfactual" image from the prompt "A large black ball sitting next to a glass of **water**". While these images represent minimally altered pairs, they are not true counterfactuals designed to contradict visual world knowledge priors, such as the expectation that "strawberries are red" or that "strawberries are larger than flies".

Similarly, model editing methods developed for LLMs have proven difficult to adapt to vision-language models. Early work shows that multi-modal neurons can encode visual-textual concepts (Pan et al., 2023), and that vision and language encoders often share object-level semantics (Sammani and Deligiannis, 2024). Yet, attempts to localize or edit factual knowledge in MLLMs, such as MMEdit (Cheng et al., 2023) and VLKEB (Huang et al., 2024) face challenges with generalization and control. Model edits can introduce unintended side effects, such as altering predictions on unrelated inputs, and often fail to generalize across paraphrased prompts or unseen contexts. In addition to the side effects of model editing, in this work, we show that MLLMs tend to override memorized knowledge priors when presented with conflicting visual input. Given that pixels override priors, applying model editing to steer the model towards "strawberries are blue" would be overridden when presented with an image of a red strawberry. Instead, steering provides a reliable mechanism to modulate model responses between pixels and priors.

Steering has seen little adoption in the multi-modal setting. Liu et al. (2024) propose latent-space steering to reduce hallucinations by stabilizing vision features at inference time. Zhang et al. (2024) show that embedded image prompts can act as hidden meta-instructions, influencing output

style or sentiment. Luo et al. (2024) demonstrate that vision-language models learn shared task vectors that generalize across modalities, suggesting that internal representations can be steered with compact task encodings. Similarly, Hojel et al. (2024) identify visual task vectors in the activation space of prompting models, showing that these representations can be patched into attention heads to guide model behavior across tasks. However, these works do not examine how models handle conflicts between visual input and stored knowledge, nor do they provide mechanisms for explicitly controlling which source of information the model relies on. To our knowledge, no prior work applies steering vectors to vision-language models for the purpose of modulating reliance on memorized visual priors versus image inputs.

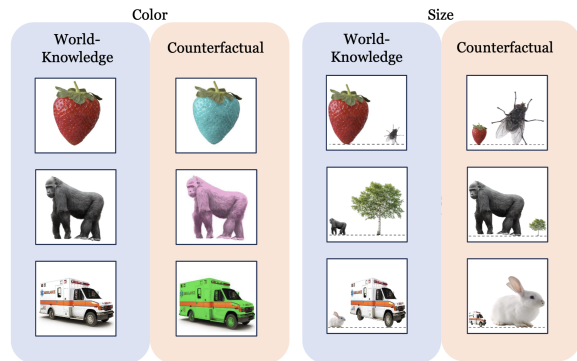## 3 Creating Visual CounterFact

Figure 2: **Visual CounterFact.** A new benchmark to study how VLMs utilize world knowledge compared to visual inputs. (Left) images created using color relations, (right) images created using size relations.

First, we describe the creation of *Visual CounterFact*, a dataset designed to examine how MLLMs use visual input and world knowledge when presented with controlled counterfactual examples. *Visual CounterFact* contains images that deliberately introduce conflicts between visual input and world knowledge, spanning two tasks: color and size. Each image is created through a four-step pipeline designed to preserve realism and control for background noise while introducing counterfactual evidence. Additional details on each step are provided in Appendix Section A.

(Step 1) Identifying objects with strong visual priors. We begin by selecting objects that have widely known visual attributes, such as canonical colors (e.g., "strawberries are red") or typical size relationships (e.g., "strawberries are larger than

flies"). These objects are sourced from human-annotated datasets (McRae norms (McRae et al., 2005)) and extended with GPT-4o estimates of typical attributes for CIFAR-100 (Krizhevsky, 2012) and ImageNet (Deng et al., 2009) categories.

(Step 2) Retrieving world knowledge images. For each object, we collect images using the Google Images API, specifying that the object should appear on a white background to reduce spurious visual cues. We aim to retrieve images that match the canonical visual prior (e.g., a red strawberry rather than a pink or green one). Each image is filtered and scored by GPT-4o for color accuracy, object correctness, and realism, and the highest-scoring image is selected.

(Step 3) Generating counterfactual relations. We construct counterfactuals that intentionally conflict with typical visual priors for each object. For the color task, we first prompt the LLaVA-Next model to generate likely colors for a given object (e.g., "What color is a strawberry?"), then sample from the five least likely common colors (e.g., blue, orange, purple) to select a counterfactual color. To maintain visual clarity, we constrain these counterfactuals to be visually distinct from the original (e.g., avoiding red/orange or gray/black swaps). For the size task, we use GPT-4o to estimate the real-world dimensions of objects and compute their total size. We select object pairs that differ by at least a factor of 10 and generate two counterfactual relations per object by inverting the expected size ordering. For example, if object A is smaller than object B and object B is smaller than object C, we create counterfactuals such as "A is bigger than B" and "B is bigger than C."

(4) Editing images to reflect counterfactual attributes. We use SAM2 (Ravi et al., 2024) segmentation masks to apply controlled, localized transformations. In the color task, we modify hue values while preserving texture and shading (Figure B Color); in the size task, we resize object masks and align them on a dashed line to reflect altered size relations without introducing depth ambiguity (Figure B Size). The final dataset contains 575 color exemplars, 575 color counterfactuals, and 877 original and 877 counterfactual size images, for a total of 2,904 visually grounded examples.

4   Methods

We use Visual CounterFact to evaluate how MLLMs store world knowledge priors of visual characteristics using three models: LLaVA-Next-7B (Li et al., 2024), Qwen2-VL-7B (Wang et al., 2024), and DeepSeek Janus Pro-7B (Chen et al., 2025). These models were selected to cover a range of current state-of-the-art multimodal architectures, including established families like LLaVA, and emerging MLLMs like QwenVL and Janus Pro. We first apply early decoding to trace the evolution of the model's prediction across layers and identify the point at which visual information overtakes linguistic priors or vice versa. We then develop PvP steering vectors that can actively shift model behavior toward either image-grounded or world knowledge responses. The results of these methods are presented in Sections 5.2 and 5.3, respectively.
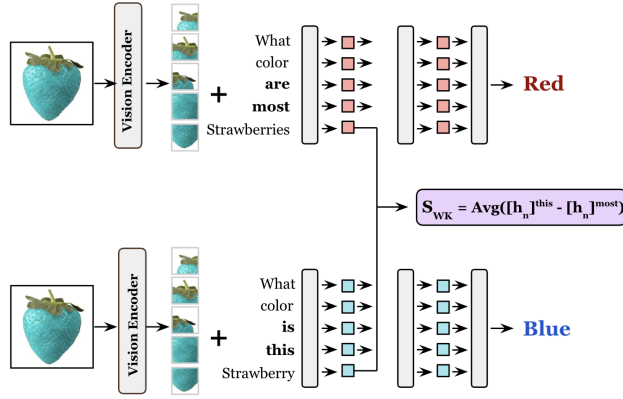
4.1   Early Decoding

Early decoding is a technique for probing the intermediate computations of a model by decoding hidden states before the final output layer. Originally introduced by nostalgebraist (2020) and extended in follow-up work (Belrose et al., 2023; Pal et al., 2023; Ghandeharioun et al., 2024; Vilas et al., 2023), this method applies the final layernorm, σ, to the intermediate hidden states h_l at layer l and then projects this representation onto the vocabulary space using the unembedding matrix W_U, yielding $W_U(\sigma(h_l))$. This produces a probability distribution over tokens, effectively allowing us to observe what the model "believes" at a given stage in its forward pass.

We use early decoding to identify when the model's prediction shifts from being guided by knowledge stored in weights to being grounded in visual perception. By decoding the model's predictions layer by layer, we observe how the probability distribution over possible output tokens evolves, allowing us to pinpoint where the model begins to favor a counterfactual (image-based) answer over the memorized world knowledge alternative.

4.2   Pixel Versus Prior Steering

Using Visual CounterFact, we introduce *Pixels Versus Priors* (PvP) steering vectors by calculating the difference in activations with contrasting prompts. Specifically, we present the model with a counterfactual image accompanied by one prompt that encourages the retrieval of world knowledge priors and another that directs it to analyze the image pixels. Consider the example in Figure 3. The prompt "What color *is this* strawberry?" encourages a visually grounded response, while "What
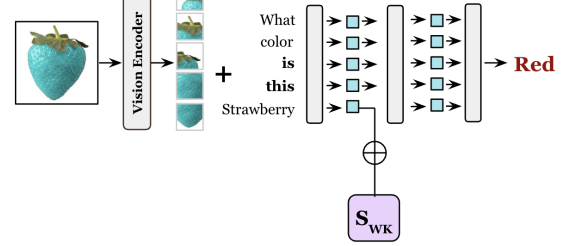
Figure 3: Pixel Versus Prior steering vectors are created by contrasting representations of prompts that emphasize pixel-level information ("this") versus priors ("most"), using the last hidden state.

color *are most* strawberries?" draws on memorized world knowledge priors about the color of a strawberry. When paired with a counterfactual image (e.g., a blue strawberry), the model should ideally answer "blue" in the first case and "red" in the second. When computing PvP steering vectors, the visual input is always the *counterfactual image*. For a given layer, l, we extract the hidden representations at the MLP block for both prompts at each layer and compute two steering vectors, S_{CF}^l and S_{WK}^l:

$$S_{CF}^l = \frac{1}{D} \sum_i^D ([h_n^l]_i^{\text{this}} - [h_n^l]_i^{\text{most}})$$

$$S_{WK}^l = \frac{1}{D} \sum_i^D ([h_n^l]_i^{\text{most}} - [h_n^l]_i^{\text{this}}).$$

Here, $i \in \{1, 2..., D\}$ represent the text-image pairs in Visual CounterFact and h_n represents the hidden state of the last text token in the sequence of a sample, typically an instruction token, which has been shown to store more important information when compared to specific subject tokens (Golovanevsky et al., 2025). After computing the world knowledge (S_{WK}^l) and counterfactual (S_{CF}^l) steering vectors, we control the model's output by modifying the hidden state of the final token in the sequence at a given layer in the language decoder. Formally, let h_n^l denote the hidden state of the last token at layer l in the language decoder of an MLLM. To steer the representation toward pixel-level information from the image, we apply the following intervention:

$$\hat{h}_n^l = h_n^l + S_{CF}^l$$

To instead steer the model toward world knowledge priors, we apply:

$$\hat{h}_n^l = h_n^l + S_{WK}^l$$

These interventions are applied for all $l \in [l, l+w]$. Our method for calculating multimodal steering vectors captures the representational shift needed to modulate the model's reliance on vision input versus world knowledge priors.

5 Results

5.1 MLLMs are Distracted By Counterfactual Images

We begin by analyzing how MLLMs behave when presented with counterfactual (CF) images that intentionally contradict common object priors, alongside baseline world knowledge (WK) images that reflect real-world visual expectations (Figure 2). To test whether models rely more on memorized knowledge or on the current image, we use two types of prompts: "What color are **most** <objects>?" and "What color is **this** <object>?"

All models perform well on "this" prompts, achieving over 80% accuracy even when the input image presents a counterfactual. This indicates that MLLMs are highly effective at grounding their answers in the current visual input. Errors in this setting typically involve subtle hue disagreements such as gold versus orange or yellow, rather than confusion about the underlying object property.

In contrast, the "most" prompts reveal a critical weakness. When asked about what is generally true, models are expected to retrieve world knowledge rather than attend to the current image. This

24841

Model	Task	CF + "this"	WK + "this"	CF + "most"	WK + "most"
LLaVA-Next	Color	85.19	87.22	47.26	92.09
	Size	82.12	96.42	40.30	95.60
Qwen2-VL	Color	84.79	85.40	60.65	90.87
	Size	91.20	98.21	28.34	96.29
Janus-Pro	Color	86.00	88.03	59.23	90.47
	Size	85.14	96.84	18.02	96.01

Table 1: Accuracy (%) for color and size tasks under "this" (e.g., "What color is this strawberry?") and "most" (e.g., "What color are most strawberries?") questions with counterfactual (CF) and world knowledge (WK) images. Models perform well when grounded in the current image, but accuracy drops sharply in the "most + CF" setting, indicating that MLLMs are overly influenced by misleading visual input.

behavior holds when WK images are shown, but accuracy drops sharply when the same question is paired with CF images. In these cases, models often abandon their prior knowledge in favor of what is visually presented, even though the prompt clearly targets a generic concept. This suggests that MLLMs are easily distracted by the current image, even when instructed to generalize.

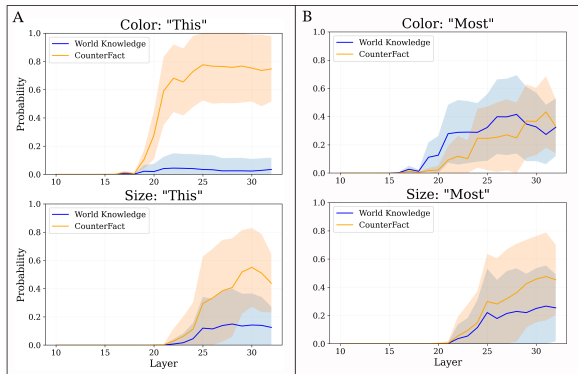## 5.2 Localizing Visual Perception Shifts through Early Decoding

Figure 4: Early decoding results on LLaVA-Next show a conflict between answering "world knowledge" using priors or answering "counterfact".

To understand how this visual override emerges during inference, we apply *early decoding* to track model predictions across layers. This reveals when the model transitions from relying on memorized priors to integrating counterfactual visual input. Figure 4 shows that in the color task when the model is prompted for the world knowledge answer but given a counterfactual image, the probability of the world-knowledge answer rises in mid to late layers, then flips to the counterfactual answer (orange) in the final layers. This "flipping behavior"

is most common when the model is prompted to respond with the world knowledge answer and provided with a counterfactual image (Figure 4 Panel B). This delayed integration of visual input leads to errors when the image contradicts memorized associations, matching the results in Table 1.

In contrast, when using "this" prompt (e.g., "what color is **this** strawberry?") for identifying the counterfactual attribute, models are confident in the counterfactual answer by the middle layers and rarely flip to the world-knowledge alternative (Figure 4 panel A and Table 2). This confidence is supported by the high inference accuracies seen in Table 1. Despite their confidence in the counterfactual answer, there is still a slight spike in world knowledge answer probability in mid-to-later layers. This slight spike shows that memorized knowledge does not fully disappear from the model, even when presented with contradicting inputs.

Table 2 shows how often models alternate between world knowledge and counterfactual answers on the color and size tasks when provided with a counterfactual image and a "most" prompt. On average, LLaVA-Next flips from world knowledge to counterfact 1.24 times on samples where a flip occurs, compared to 0.79 in the reverse direction. This indicates that MLLMs are prone to overriding prior knowledge when presented with a counterfactual image. These results suggest a consistent pattern: models initially rely on linguistic priors rooted in world knowledge, and only later override these with visual evidence as processing progresses through the layers.

	Size			Color		
	LLaVA-Next	Qwen2-VL	Janus-Pro	LLaVA-Next	Qwen2-VL	Janus-Pro
% Samples w/o Flip	45%	69%	70%	35%	71%	88%
% Samples w/ Flip	55%	31%	30%	65%	29%	12%
Avg. # CF → WK	1.02	0.47	0.76	0.84	0.18	0.56
Avg. # WK → CF	1.12	0.70	0.86	1.31	0.90	0.56

Table 2: Flip statistics for size and color attributes with counterfactual images and prompts designed to elicit world knowledge responses. A "flip" occurs when the initially less probable response later exceeds the alternative by at least 5%.
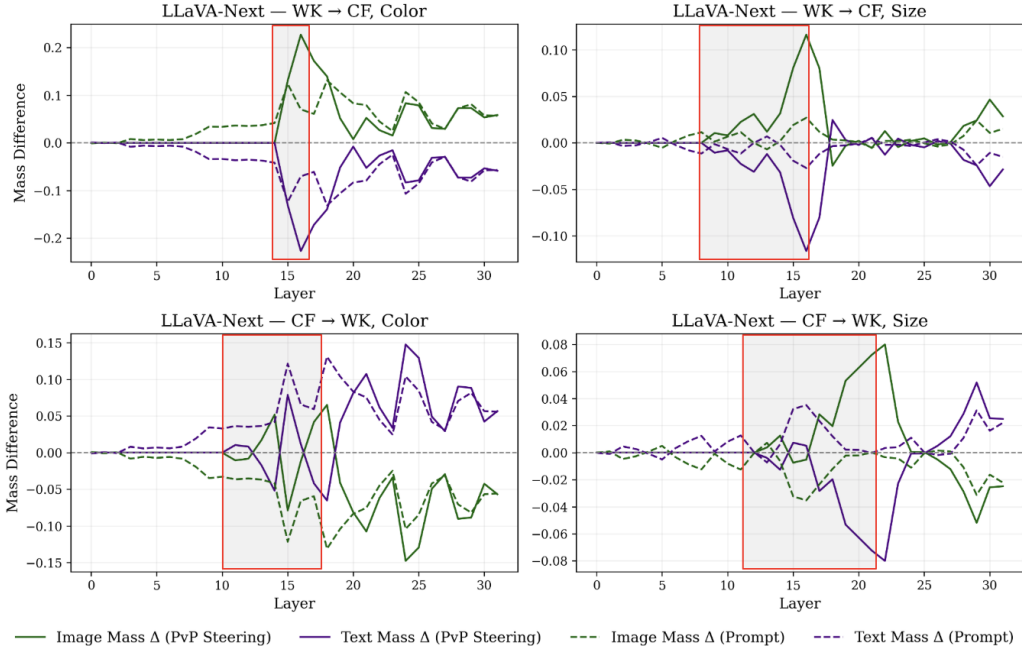

Figure 5: Effect of prompt changes and interventions on attention mass across layers for LLaVA-Next in the color and size tasks. Solid lines show changes when applying the steering vector; dashed lines show the effect of modifying the prompt. Green and purple lines represent attention shifts toward image and text tokens, respectively. The red shaded region highlights the layers where the intervention was applied (corresponding to Table 3). We see that intervention has a much stronger effect than changing the prompt.

5.3 Controlling World Knowledge Associations with PvP Steering Vectors

In Section 5.2, we show that MLLMs tend to rely on world knowledge in early layers and shift to visual information later, often flipping between the two. This delayed integration of visual input often results in unstable predictions when images conflict with prior knowledge (seen in Table 1). To stabilize predictions and control whether the model attends to the image or draws from prior knowledge, we use Pixel Versus Prior Steering (see Section 4.2). Practically, PvP steering offers an interpretable method to causally intervene in model processing, revealing the layers and temporal windows where the balance between vision and world knowledge can be effectively manipulated.

Table 3 shows the effectiveness of our steering approach across tasks and models, highlighting both the percentage of successful steering of the model and the key layers at which intervention has the highest impact. We apply PvP steering vectors to the set of inputs that the model originally gets incorrect, meaning without PvP-steering, the model performance on this subset of data is 0%. Remarkably, we achieve at least a 98% success rate in flipping model predictions from world knowledge to counterfactual answers in the color task for all models. This demonstrates that MLLMs are not only steerable but highly responsive to targeted interventions, particularly when guided away from strongly encoded world knowledge priors. The

Model	Task	Direction	Flips %	Layers
LLaVA	Color	WK → CF	**99.5**	[14-16]
		CF → WK	86.4	[10-17]
	Size	WK → CF	**71.3**	[8-16]
		CF → WK	33.5	[12-21]
QwenVL	Color	WK → CF	**99.7**	[17-19]
		CF → WK	78.8	[12-17]
	Size	WK → CF	**89.9**	[16-22]
		CF → WK	61.8	[13-23]
Janus-Pro	Color	WK → CF	**98.6**	[14-16]
		CF → WK	78.2	[15-18]
	Size	WK → CF	**81.2**	[16-19]
		CF → WK	70.37	[16-20]

Table 3: Performance of models under Color and Size tasks with two flip directions: WK → CF and CF → WK. The key layers are shown for each flip direction.

size task is more difficult by nature: it requires detecting two objects and reasoning about their size relationship, making it more dependent on deeper, distributed visual processing. This is reflected in lower flip rates and broader intervention windows.

Across models, we observe that the most effective interventions tend to occur within specific mid-to-late layer ranges, typically requiring sustained influence over multiple layers (Table 3). In general, steering the model from world knowledge to a counterfactual (WK → CF) demands less intervention than reversing that shift (CF → WK), suggesting that overriding memorized priors is easier than restoring them once suppressed by a counterfactual input.

5.4 Impact of Prompting and Steering Vectors on Attention Mass
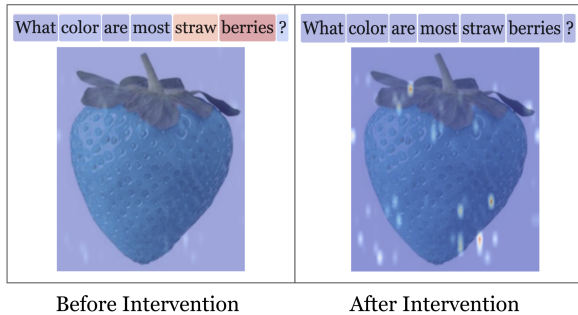

Figure 6: Impact of intervention on attention distribution. We observe more attention on text before intervention, and see a shift in attention towards the image and away from the text after intervention.

While our results demonstrate that PvP steering vectors can reliably shift model outputs, we do not

know how this shift is implemented internally. We hypothesize that this shift is implemented in the attention layers, as these components gather information from either the image or text tokens. To study the impact of steering vectors on model predictions, we analyze their impact on the model's attention patterns. We compare two settings: (1) changing the prompt from "most" to "this" (dashed lines), and (2) applying our PvP steering vector (solid lines) that steers the model toward or away from a counterfactual response. The first setting explores how changing the prompt shifts the model's attention. Asking about the color of "most" strawberries should encourage the model to focus on prior knowledge (red), while asking about "this" strawberry directs attention to the specific pixels (blue). The second setting shows how injecting the PvP steering vector guides the model's internal attention beyond the effect of the prompt.

As shown in Figure 5, changing the prompt from "most" to "this" yields a modest shift in attention toward image tokens. For example, on the color task, LLaVA-Next increases the attention mass to image tokens by 13%. In contrast, the PvP intervention vector causes a much stronger shift, increasing attention mass to image tokens to 40% (see Table 5 in Appendix C for all models).

Among all models, LLaVA-Next shows the strongest and most consistent shifts in attention when interventions are applied, followed by Qwen2-VL and then Janus-Pro. For the size task, steering vectors must act earlier in the network and across more layers to be effective, reflecting the fact that size requires integrating more visual features than color. In contrast, the color task is more localized and easier to influence with a smaller intervention window. Figure 5 illustrates how attention moves across layers in response to both prompt changes and steering task interventions in LLaVA-Next. To illustrate the same effect but with a concrete example, Figure 6, shows the attention before intervention being heavily focused on text, with "strawberries" highlighted in red (highest attention). After intervention, attention shifts to the image, with most red regions being inside the image, rather than over the text.

These findings show that PvP steering vectors reshape internal attention mechanisms more effectively than prompt changes alone. They offer precise control over how models allocate attention to visual inputs, especially in tasks like size comparison that require broader spatial reasoning. By

intervening directly in the model's representation space, PvP steering enables deeper interpretability and control over MLLM behavior.

6 Conclusion

In this work, we investigate how multimodal large language models (MLLMs) reconcile memorized world knowledge and visual input. Understanding this balance is essential for building reliable models that can correctly choose between conflicting sources of information. To study this, we introduce *Visual CounterFact*, a dataset that constructs realistic visual counterfactuals targeting familiar attributes like object color and size. These examples violate learned priors while preserving visual plausibility, enabling precise comparisons between perception and memory. Using this dataset, we find that MLLMs often default to perception, even when prompted to retrieve general knowledge. In these cases, performance on knowledge-based prompts drops significantly, suggesting that models are overly influenced by visual inputs, even when the question targets memorized facts. Through studying the forward-pass, we observe that model predictions initially reflect stored priors, then transition to visually grounded answers in mid-to-late layers. This transition is often unstable, with models flipping between the two sources of information across layers. To control this behavior, we introduce *Pixels Versus Priors* steering vectors, which allow us to edit model behavior toward preferring either world knowledge priors or visual input. These activation-level interventions produce significant attention shifts towards or away from the image, depending on our steering vector direction. Our findings offer a new framework for interpreting and controlling MLLMs, advancing our ability to understand and control the interaction between memory and perception in multimodal models.

7 Limitations

Our framework focuses on three state-of-the-art models: LLaVA-Next, Qwen2-VL, and Janus-Pro, which, while diverse, do not represent the full spectrum of multimodal architectures, such as monolithic MLLMs. However, this level of focus is consistent with standard practice in interpretability research, where analyses typically target one or two models to enable detailed, mechanism-level insights across both LLMs and MLLMs (Meng et al., 2022; Dai et al., 2021; Luo et al., 2024; Hojel

et al., 2024). Despite architectural differences, our findings consistently generalize across the models studied, supporting the robustness of our approach. In future work, we plan to expand our analysis to a broader range of models to explore how architectural design impacts reliance on perception versus prior knowledge.

Additionally, through our analysis we find that steering models from visual perception back to world knowledge is more difficult than the reverse, suggesting an asymmetry in how MLLMs prioritize in-context versus memorized information. Understanding this distinction further remains an open direction for future work.

References

Bang An, Sicheng Zhu, Michael-Andrei Panaitescu-Liess, Chaithanya Kumar Mummadi, and Furong Huang. 2024. Perceptionclip: Visual classification by inferring and conditioning on contexts. *Preprint*, arXiv:2308.01313.

Suraj Anand, Michael A Lepori, Jack Merullo, and Ellie Pavlick. 2025. Dual process learning: Controlling use of in-context vs. in-weights strategies with weight forgetting. *International Conference on Learning Representations*.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *Preprint*, arXiv:2303.08112.

Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. 2022. Data distributional properties drive emergent in-context learning in transformers. *Advances in neural information processing systems*, 35:18878–18891.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *Preprint*, arXiv:2501.17811.

Siyuan Cheng, Bozhong Tian, Qingbin Liu, Xi Chen, Yongheng Wang, Huajun Chen, and Ningyu Zhang. 2023. Can we edit multimodal large language models? *arXiv preprint arXiv:2310.08475*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2021. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference*

on Computer Vision and Pattern Recognition, pages 248–255.

Zorik Gekhman, Eyal Ben David, Hadas Orgad, Eran Ofek, Yonatan Belinkov, Idan Szpektor, Jonathan Herzig, and Roi Reichart. 2025. Inside-out: Hidden factual knowledge in llms. *arXiv preprint arXiv:2503.15299*.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. *arXiv preprint arXiv:2304.14767*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2020. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patch-scopes: A unifying framework for inspecting hidden representations of language models. *Preprint*, arXiv:2401.06102.

Michal Golovanevsky, William Rudman, Vedant Palit, Ritambhara Singh, and Carsten Eickhoff. 2025. What do vlms notice? a mechanistic interpretability pipeline for gaussian-noise-free text-image corruption and evaluation. *Preprint*, arXiv:2406.16320.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*.

Alberto Hojel, Yutong Bai, Trevor Darrell, Amir Globerson, and Amir Bar. 2024. Finding visual task vectors. In *European Conference on Computer Vision*, pages 257–273. Springer.

Han Huang, Haitian Zhong, Tao Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2024. Vlkeb: A large vision-language model knowledge editing benchmark. *arXiv preprint arXiv:2403.07350*.

Nick Jiang, Anish Kachinthaya, Suzie Petryk, and Yossi Gandelsman. 2024. Interpreting and editing vision-language representations to mitigate hallucinations. *arXiv preprint arXiv:2410.02762*.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. *arXiv preprint arXiv:2402.18154*.

Alex Krizhevsky. 2012. Learning multiple layers of features from tiny images. *University of Toronto*.

Andrew Kyle Lampinen, Stephanie CY Chan, Aaditya K Singh, and Murray Shanahan. 2024. The broader spectrum of in-context learning. *arXiv preprint arXiv:2412.03782*.

Tiep Le, Vasudev Lal, and Phillip Howard. 2023. Coco-counterfactuals: Automatically constructed counterfactual examples for image-text pairs. *Advances in Neural Information Processing Systems*, 36:71195–71221.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *Preprint*, arXiv:2407.07895.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2024. Reducing hallucinations in vision-language models via latent space steering. *Preprint*, arXiv:2410.15778.

Grace Luo, Trevor Darrell, and Amir Bar. 2024. Task vectors are cross-modal. *arXiv preprint arXiv:2410.22330*.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Mazda Moayeri, Wenxiao Wang, Sahil Singla, and Soheil Feizi. 2023. Spuriosity rankings: Sorting data to measure and mitigate biases. *Preprint*, arXiv:2212.02648.

nostalgebraist. 2020. interpreting gpt: the logit lens. *LessWrong*.

Koyena Pal, Jiuding Sun, Andrew Yuan, Byron Wallace, and David Bau. 2023. Future lens: Anticipating subsequent tokens from a single hidden state. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics.

Haowen Pan, Yixin Cao, Xiaozhi Wang, Xun Yang, and Meng Wang. 2023. Finding and editing multi-modal neurons in pre-trained transformers. *arXiv preprint arXiv:2311.07470*.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2021. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*.

Core Francisco Park, Ekdeep Singh Lubana, Itamar Pres, and Hidenori Tanaka. 2024. Competition dynamics shape algorithmic phases of in-context learning. *arXiv preprint arXiv:2412.01003*.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura

Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. 2024. Sam 2: Segment anything in images and videos. *Preprint*, arXiv:2408.00714.

Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *The Twelfth International Conference on Learning Representations*.

William Rudman, Michal Golovanevsky, Amir Bar, Vedant Palit, Yann LeCun, Carsten Eickhoff, and Ritambhara Singh. 2025. Forgotten polygons: Multimodal large language models are shape-blind. *arXiv preprint arXiv:2502.15969*.

Fawaz Sammani and Nikos Deligiannis. 2024. Interpreting and analyzing clip's zero-shot image classification via mutual knowledge. *arXiv preprint arXiv:2410.13016*.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. Foil it! find one mismatch between image and language caption. *arXiv preprint arXiv:1705.01359*.

Aaditya Singh, Stephanie Chan, Ted Moskovitz, Erin Grant, Andrew Saxe, and Felix Hill. 2023. The transient nature of emergent in-context learning in transformers. *Advances in Neural Information Processing Systems*, 36:27801–27819.

Nishant Subramani, Nivedita Suresh, and Matthew E Peters. 2022. Extracting latent steering vectors from pretrained language models. *arXiv preprint arXiv:2205.05124*.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Martina G Vilas, Timothy Schaumlöffel, and Gemma Roig. 2023. Analyzing vision transformers for image classification in class embedding space. *Advances in neural information processing systems*, 36:40030–40041.

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. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. 2021. Noise or signal: The role of image backgrounds in object recognition. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yu Yang, Besmira Nushi, Hamid Palangi, and Baharan Mirzasoleiman. 2023. Mitigating spurious correlations in multi-modal models during fine-tuning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. Characterizing mechanisms for factual recall in language models. *arXiv preprint arXiv:2310.15910*.

Tingwei Zhang, Collin Zhang, John X Morris, Eugene Bagdasarian, and Vitaly Shmatikov. 2024. Soft prompts go hard: Steering visual language models with hidden meta-instructions. *arXiv preprint arXiv:2407.08970*.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. 2022. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*.

Nicolas Zucchet, Jörg Bornschein, Stephanie Chan, Andrew Lampinen, Razvan Pascanu, and Soham De. 2025. How do language models learn facts? dynamics, curricula and hallucinations. *arXiv preprint arXiv:2503.21676*.

A   Creating Visual CounterFact

Step 1: Sourcing Objects. We begin by identifying a set of real-world *objects* (e.g., *strawberry*, *squirrel*, *cherry*) from the McRae feature norms dataset (McRae et al., 2005), ImageNet (Deng et al., 2009), and CIFAR-100 (Krizhevsky, 2012). The McRae dataset tasks 100 human participants with listing common attributes for an object. If at least 30% of participants respond with a specific color as a key attribute of that object, we include it in our dataset. After filtering with this 30% threshold, we obtain 190 objects. To increase the number of objects, we use the GPT-4o to elicit typical colors for ImageNet 1000 and CIFAR-100 classes and discard any categories lacking strong color ground-truth priors (e.g., clothing items). After both procedures, we obtain **622** unique objects.

Step 2: Sampling Real-World Images. We use the Google Images API to retrieve three candidate images of each object, specifying that they must be on a white background (a `<world knowledge color> <subject>` on a white background). Since prior work has shown that background information
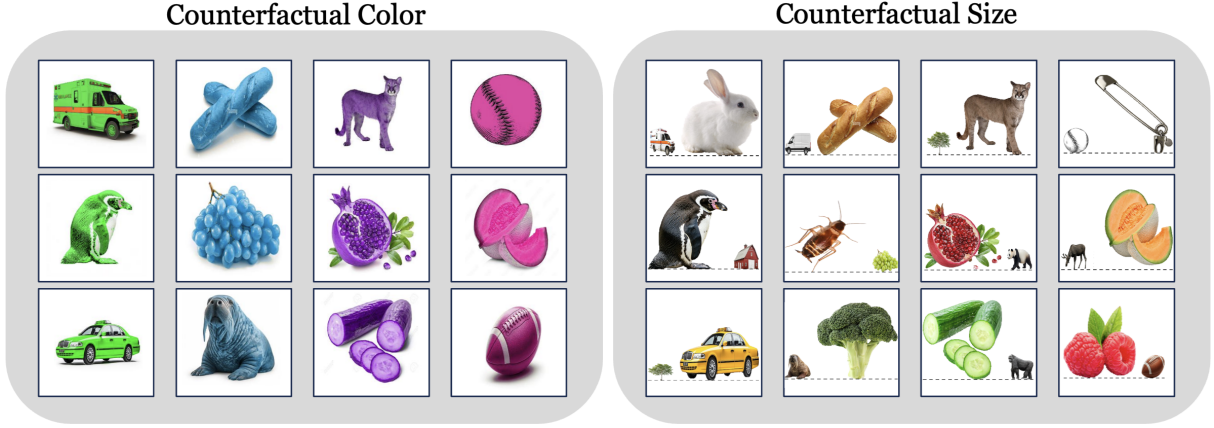
Figure 7: Examples from our dataset. (Left) images created using color relations, (right) images created using size relations.

can heavily bias object detection models, we extract objects against a white background to ensure that model decisions are not influenced by spurious background features (Xiao et al., 2021; Moayeri et al., 2023; Yang et al., 2023; An et al., 2024). Sampled images are then evaluated using GPT-4o, which scores them based on object correctness, color accuracy, presence of a white background, and overall realism. Specifically, we prompt GPT-4o with the following questions:

1. Is this an image of a <color> <object>? (yes/no)
2. Is this image on a white background? (yes/no)
3. Is this image an illustration or a realistic image? (illustration/realistic)
4. On a scale from 1 to 10, how realistic is this <object>? (numerical score)

We retain the highest-scoring image of the three to ensure visual fidelity with our inclusion criteria. For each yes/no question, the image receives a score of 0 for "no" or a score of 10 for "yes". For images scoring 0 overall, we repeat the querying process but remove the world knowledge color (query: a <subject> on a white background), as most of the images resulting in a score of 0 are multi-colored (e.g., zebra, bee). After the second round of querying, we drop any remaining images with score 0, resulting in **575** unique objects.

Step 3: Constructing Object-Size and Object-Color Counterfactuals.

(1) Color: To generate color counterfactuals, we prompt the LLaVA-Next model (Mistral-7B backbone) with "What color is a <object>?" and randomly sample from the five least likely color predictions (using common colors such as red, blue,

green, pink, orange, etc.). This ensures that counterfactuals challenge the model's linguistic priors, encouraging reliance on visual input rather than memorized associations. We constrain counterfactual colors to be perceptually distinct from the original color (e.g., avoiding red/orange or gray/black swaps). **(2) Size:** For the size task, we use the same set of objects from the color task and estimate their typical real-world dimensions using GPT-4o. The model provides height and width in inches, which we multiply to compute a total size metric. We then identify object pairs that differ in size by at least a factor of 10. For each object, we create two counterfactual images. Given three objects that satisfy $object_1$ < $object_2$ < $object_3$, where "<" denotes increasing real-world size, we generate two counterfactual images containing $object_2$. Namely, ($object_1$, bigger_than, $object_2$) and ($object_2$, bigger_than, $object_3$). For example, if a squirrel is typically bigger than a cherry and smaller than an alligator, we create the counterfactual images (cherry, bigger_than, squirrel) and (squirrel, bigger_than, alligator). This creates twice the number of samples since we construct two size relations for one object. After manual filtering of sizes that GPT-4o reported incorrectly, we are left with 877 unique size-object relations.

Step 4: Creating Counterfactual Images. Given a retrieved object on a white background, the first step in creating counterfactual images is to use SAM2 (Ravi et al., 2024) to obtain a segmentation mask. After we obtain segmentation masks, we use two separate pipelines to create color and size counterfactuals, respectively.
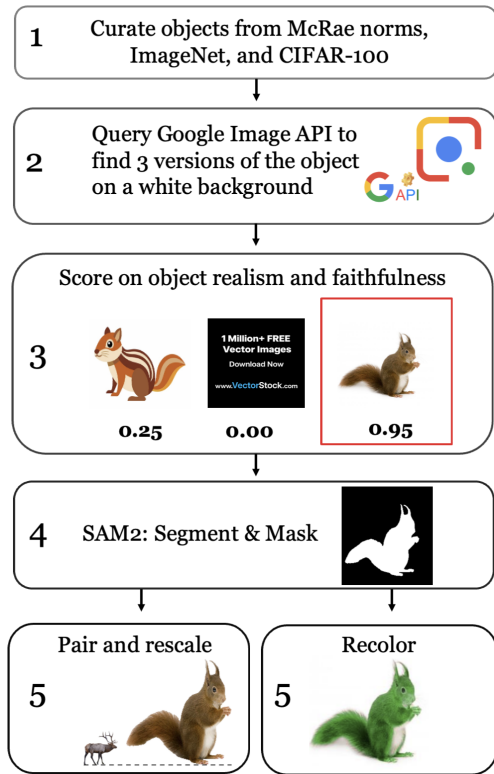
24848

Figure 8: Construction pipeline for the Visual Counterfactual Dataset. We identify typical traits using semantic knowledge sources, retrieve realistic visual exemplars, and apply transformations to create perceptually plausible counterfactuals that conflict with language priors.

(1) Color: Given a color-object relation with the relation "[object] is [color]" and its counterfactual relation "[object] is [counterfactual color]", we apply a segmentation mask to isolate the object and modify only hue values in the HSV color space in order to change the color of the object to the counterfactual color while preserving the original saturation and brightness. This produces realistic and semantically surprising transformations (e.g., turning a red strawberry blue) while maintaining texture and shading. For objects with minimal hue (e.g., gray or black), we apply a set of hand-written remapping rules.

(2) Size: Given two objects with the world knowledge relation "[object 1] is larger than [object 2]" we create the counterfactual image reflecting "[object 2] is larger than [object 1]" by combining the segmentation masks of $object_1$ and $object_2$. For the world knowledge image, we resize the masks so that $object_1$ appears significantly larger than $object_2$, and in the counterfactual image, we resize the masks so that $object_2$ appears significantly larger than $object_1$. Specifically, 250×250

pixels versus 80×80 pixels. This size difference visually reflects the intended relation.

To make the size comparison clear and avoid depth ambiguity, we place both objects on the same horizontal baseline and add a black dashed reference line that both objects touch. This helps ensure that differences in perceived size are interpreted as scale changes rather than perspective shifts.

Visual CounterFact consists of 575 original (world knowledge) object images, 575 color counterfactual images, and 877 size original and 877 counterfactual images, totaling 2,904 unique images. Figure 2 shows examples from each split of the dataset. These transformations yield a dataset that explicitly conflicts with world knowledge priors of an object's color and size while preserving perceptual plausibility, enabling targeted evaluation of visual reasoning models under counterfactual conditions. In Appendix Section B, we provide additional examples of images as well as dataset statistics on the kinds of objects we include in Visual Counterfact.

B Dataset Statistics and Examples

Category	Count
Animals	218
Household Items and Furniture	59
Fruits and Vegetables	58
Vehicles and Transportation	42
Electronics and Appliances	29
Tools and Hardware	22
Food and Drink (non-produce)	21
Buildings and Structures	21
Plants and Trees	19
Musical Instruments	15
Clothing and Accessories	15
Weapons and Military Items	13
Medical and Hygiene Items	11
Toys and Recreational Items	11
Natural Objects (non-living)	10
Office Supplies	6
Miscellaneous	5

Table 4: Distribution of object categories in the dataset.

Figure 2 illustrates representative examples from the Visual CounterFact dataset, including counterfactual edits based on object color (left) and relative size (right). Each example maintains visual realism while introducing semantically meaningful contra-

dictions to typical object properties.

Table 4 summarizes the distribution of object categories in the dataset. The majority of counterfactuals involve animals, followed by a diverse set of objects spanning furniture, produce, vehicles, tools, and more. This broad coverage ensures the dataset tests model reliance on both visual input and memorized associations across varied semantic domains.

C Attention Mass Details

To better understand how steering vectors and prompt changes affect the internal attention mechanisms of MLLMs, we visualize the change in attention mass over layers for each model and task. Figures 10 and 11 show these effects across color and size tasks, respectively. Each subplot compares attention mass difference toward image tokens (green) and text tokens (purple), with solid lines indicating PvP steering interventions and dashed lines indicating prompt-only changes.

We find that across all models and both tasks, interventions consistently produce stronger shifts in attention mass compared to prompt changes alone. In the color task (Figure 10), steering from WK to CF reliably increases image attention, while the reverse direction decreases it (as expected). The effect is particularly pronounced in LLaVA-Next, with peaks around the intervention window. The size task (Figure 11) shows a similar but more muted pattern, consistent with the task's higher visual complexity. These trends reinforce that PvP steering vectors exert precise, causal control over how attention is allocated between vision and language streams.
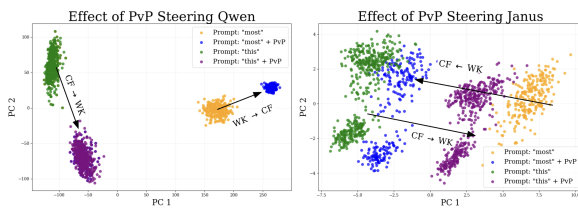

Figure 9: First two principal components of sentence embeddings of Qwen2.5-VL and Janus-Pro before and after steering from priors to pixels and from pixels to priors.

Table 5 reports the maximum change in attention mass directed toward image tokens across models and tasks, comparing the effects of prompt changes ("most" to "this") and PvP steering interventions. As described in Section 4.2, we measure the peak

LLaVA-Next		Intervention	Prompt
Color	WK → CF	40.0%	13.1%
Color	CF → WK	−15.6%	−13.1%
Size	WK → CF	10.9%	3.2%
Size	CF → WK	−7.6%	−3.2%
Qwen2-VL			
Color	WK → CF	21.8%	12.8%
Color	CF → WK	−25.2%	−12.4%
Size	WK → CF	14.1%	4.7%
Size	CF → WK	−10.8%	−4.7%
Janus-Pro			
Color	WK → CF	19.5%	11.7%
Color	CF → WK	−11.0%	−10.2%
Size	WK → CF	2.4%	1.1%
Size	CF → WK	−1.1%	−1.1%

Table 5: Max change in image attention mass (Δ) under intervention and prompt changes for each model and task.

increase or decrease in image attention during inference.

Across all models, PvP steering consistently produces larger attention shifts than prompt modifications. This effect is most pronounced in the color task, where interventions increase image attention by up to 40% in LLaVA-Next, compared to 13% from prompting. As visualized in Figure 5, prompting leads to moderate reallocation of attention, while steering vectors induce strong and targeted redistribution.

We also observe an asymmetry between WK → CF and CF → WK directions: steering toward perception (WK → CF) is generally more effective than restoring attention to prior-based information (CF → WK). This aligns with accuracy results in Section 5.3, where interventions that shift models away from priors are more successful than those that attempt to recover them.

These findings provide further evidence that PvP steering offers fine-grained control over internal attention dynamics in MLLMs, outperforming prompt-based techniques in both strength and specificity.

D Early Decoding

Figure 12 shows early decoding traces for Qwen2.5-VL and Janus-Pro, extending our main-layer analysis from Figure 4. Consistent with the behavior observed in LLaVA-Next, both models initially assign high probability to the world knowledge answer when prompted with a "most" question and shown a counterfactual image. However, as the forward
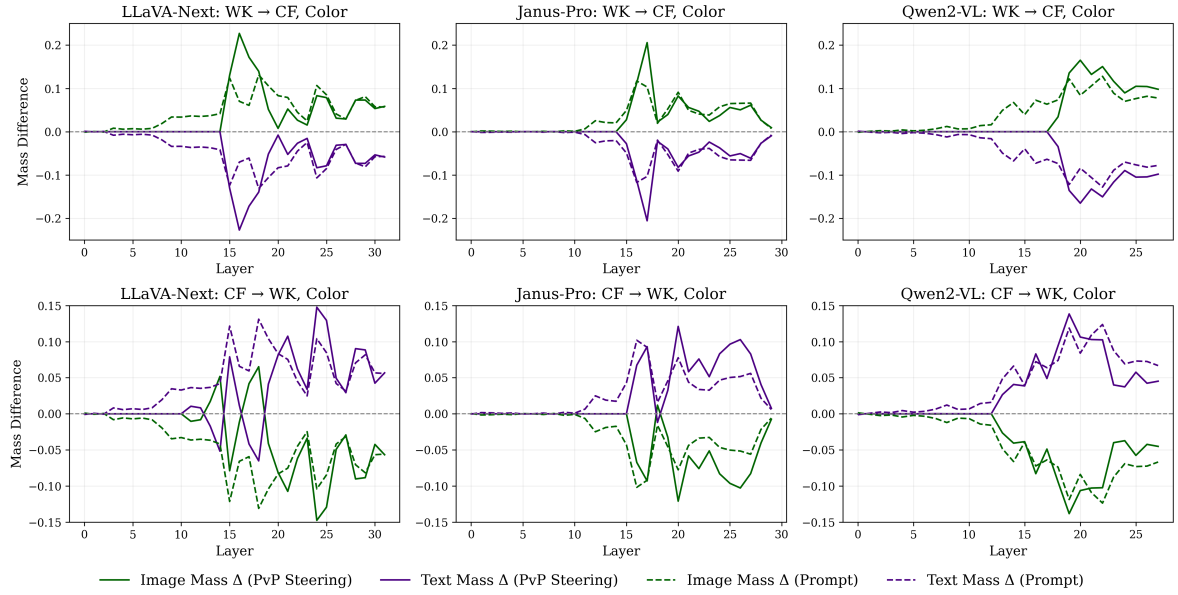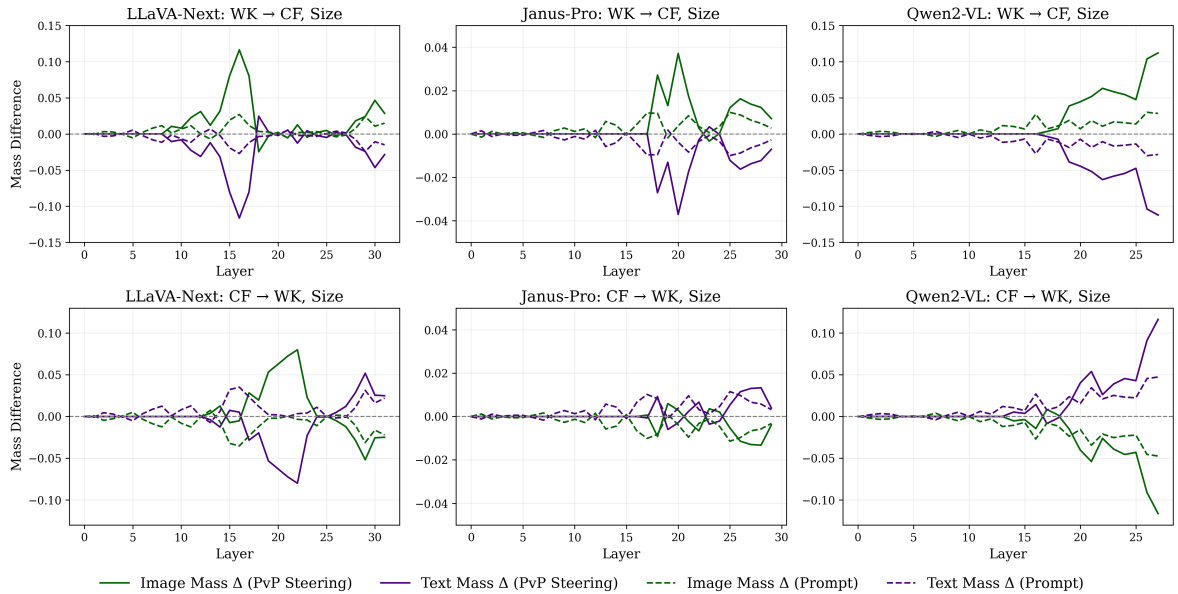
Figure 10: Attention mass difference across layers for all models on the color task. Solid lines show changes from PvP steering vectors; dashed lines show prompt-only changes. Green represents attention to image tokens, purple to text tokens. Each subplot shows one model and intervention direction (WK → CF or CF → WK).

Figure 11: Attention mass difference across layers for all models on the size task. Solid lines show changes from PvP steering vectors; dashed lines show prompt-only changes. Each subplot shows one model and intervention direction (WK → CF or CF → WK).

pass progresses, the probability of the counterfactual answer rises and ultimately dominates by the final layer.

For both models, this flipping behavior illustrates the delayed integration of visual information, often leading the model to override its prior with perceptual evidence late in the forward pass.

When prompted with "this" questions, both models quickly favor the counterfactual answer and rarely flip to world knowledge. These early decoding results across all three models reinforce our central finding: MLLMs are highly sensitive to visual input and tend to prioritize perception over memorized priors when the two conflict, particularly in
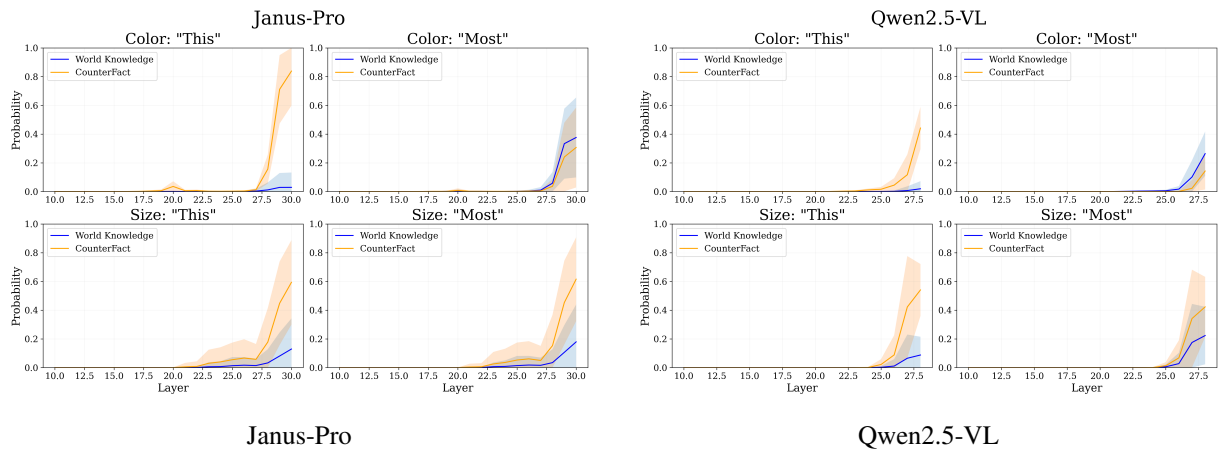
24851

Figure 12: Early decoding results on Qwen2.5-VL and Janus-Pro show a conflict between answering "world knowledge" using in-weight memorization or answering "counterfact" using visual perception.

later layers of processing.