When seeing Overrides Knowing: Disentangling Knowledge Conflicts in Vision-Language Models

Francesco Ortu [1,2] *

Zhijing Jin [3,4,5]

Diego Doimo[2]†

Alberto Cazzaniga[2]†

[1] University of Trieste [2] Area Science Park
[3] MPI [4] University of Toronto [5] Vector Institute

Abstract

Vision-language models (VLMs) increasingly combine both visual and textual information to perform complex tasks. However, conflicts between their internal knowledge and external visual input can lead to hallucinations and unreliable predictions. In this work, we investigate the mechanisms that VLMs use to resolve cross-modal conflicts by introducing a dataset of multimodal counterfactual queries that deliberately contradict internal commonsense knowledge. Through logit inspection, we identify a small set of attention heads that mediate this conflict. By intervening in these heads, we can steer the model towards its internal knowledge or the visual inputs. Our results show that attention from these heads effectively locates image regions that influence visual overrides, providing a more precise attribution compared to gradient-based methods. [1]

1 Introduction

Vision–language models (VLMs) [Alayrac et al., 2022, Li et al., 2022, Liu et al., 2023, Team, 2024, Deitke et al., 2024] have shown remarkable versatility in various multimodal tasks, from image understanding to image generation. They draw on their ability to combine two key sources of information: a rich set of world knowledge acquired during pretraining, and contextual cues provided in the input prompts. However, these two sources can sometimes contradict each other, for example, when the pretraining knowledge becomes outdated [Lazaridou et al., 2021, Luu et al., 2022] or when prompts include intentionally misleading visual information [Liu et al., 2024d]. Such conflicts often lead to hallucinations in model responses [Cui et al., 2023, Liu et al., 2024a, Guan et al., 2024], yet the internal mechanisms responsible for these errors remain poorly understood [Xu et al., 2024].

In this work, we analyze how VLMs resolve conflicts between visual input and internal knowledge by framing the problem through counterfactual image-text pairs. We prompt the VLMs with images depicting unusual or absurd scenes taken from the WHOOPS! dataset [Guetta et al., 2023], followed by a sentence encouraging a typical knowledge-based continuation. As shown in fig. 1, each input prompt is associated with a counterfactual pair of completions. For instance, the model may be shown an image of a wolf howling at the sun, a scene that contradicts commonsense knowledge, and asked to complete the prompt accordingly (see top-left panel). We construct the dataset such that

*Correspondence to: francesco.ortu@phd.units.it, alberto.cazzaniga@areasciencepark.it
†Equal supervision.
[1]Code and Dataset: francescortu/Seeing-Knowing

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Mechanistic Interpretability.

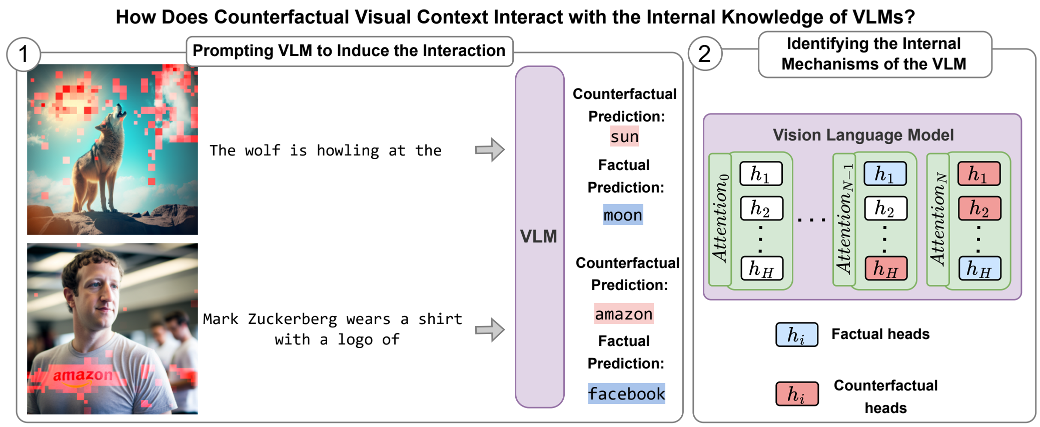

Figure 1: **Overview of our approach.** (**Left**) We construct prompts that induce a conflict between a vision-language model's internal factual knowledge and counterfactual visual context. (**Right**) We then analyze which components in the model mediate this tension, identifying attention heads and visual patches that favor factual or visually grounded predictions.

VLMs, when prompted with text alone, generate commonsense responses while in the presence of the image, change their prediction to align with the visual context, even when it contradicts their internal knowledge. Building on the approach of Ortu et al. [2024], we identify which internal components of the model contribute the most to factual versus counterfactual predictions. We find that a small subset of attention heads mediates this competition, and targeted interventions on these heads reliably alter the model's outputs. We also show that these heads are more effective than gradient-based methods in identifying the most important parts of an image to resolve multimodal conflicts in VLMs.

In summary, our contributions are as follows:

1. We construct WHOOPS-AHA!, a dataset that combines images containing counterfactual scene elements and commonsense textual queries, designed to analyze conflicts between visual context and internal knowledge (sec. 5.1);
2. We identify the attention heads that promote factual and counterfactual responses, ranking their importance with logit attribution (sec. 5.2);
3. By reweighting these heads, we show that we can control the tendency of the model to rely on the visual evidence or its internal knowledge and vice versa (sec. 5.3);
4. We demonstrate that direct attention attribution from conflict-resolution heads provides more accurate identification of counterfactual image regions than traditional gradient-based attribution methods (sec. 5.4).

2 Related Work

Knowledge Conflicts in Vision-Language Models VLMs frequently encounter situations where visual input contradicts their internal parametric knowledge, yet the mechanisms governing conflict resolution remain poorly understood [Xu et al., 2024]. Early work on multimodal conflicts focused primarily on behavioral evaluation through benchmark construction. Han et al. [2024] introduced datasets that probe contextual knowledge conflicts with deceptive visual elements, while Liu et al. [2024c] developed ConflictVis to evaluate conflicts between visual input and parametric knowledge. Le et al. [2023] created COCO-Counterfactuals using minimally edited counterfactual image pairs to study model behavior under visual contradictions. However, these studies limit their analysis to evaluating model outputs and prompt structures without investigating the internal mechanisms by which models resolve conflicts.

Mechanistic interpretability in VLMs. Mechanistic interpretability, which seeks to reverse engineer deep neural networks, has made significant strides in text-only models [Elhage et al., 2021, Geva et al., 2023, 2021, Hanna et al., 2023, inter-alia]. Recently, attention has shifted to VLMs. For instance, Neo et al. [2025] explores how LLaVA processes visual information, and Basu et al. [2024]

examines knowledge retrieval in VLMs for visual question answering. Despite these advances, the mechanistic investigation of how VLMs resolve conflicts between modalities remains underexplored.

Internal dynamics of multimodal conflicts. Although interest in VLM interpretability is growing, mechanistic studies of how these models process conflicting information remain limited. In the context of LLMs, research has focused on understanding how models resolve conflicts between contextual and internal knowledge [Ortu et al., 2024, Yu et al., 2023, Jin et al., 2024]. Recent work has started to explore the internal mechanisms of VLMs in handling multimodal conflicts. Recent work has begun exploring internal mechanisms in VLMs: Hua et al. [2025] analyze conflicts between text and image, while Golovanevsky et al. [2025b] introduced NOTICE, using semantically corrupted image pairs to study attention heads in LLaVA and BLIP. Although they identified attention heads with distinct functional roles in tasks, their focus was not specifically on the resolution of conflict between visual and internal knowledge. Recently, Golovanevsky et al. [2025a] introduced a method that uses steering vectors to control model predictions. They examined how varying visual input affects the competition between modalities, utilizing image pairs, one aligned with the model's internal knowledge and the other modified to introduce a counterfactual variation. Their approach focuses on simple object attributes, such as color or size, as a means of probing these conflicts.

3 Dataset

3.1 Requirements for Mechanistic Analysis of Multimodal Conflicts

Mechanistic interpretability of VLMs requires datasets that enable precise analysis of internal information flow. To support this goal, we identified four key requirements for a suitable dataset:

- **Controlled Conflict Induction**: Conflicts between visual input and internal knowledge must be systematically induced and verifiable, enabling causal analysis.
- **Token-Level Precision**: The dataset should allow token-level inspection and interventions, with prompts designed to elicit specific, predictable continuations.
- **Commonsense Knowledge Grounding**: Scenarios must rely on the model's internal parametric knowledge, providing strong, consistent priors that can be challenged by visual input.
- **Topical Generality**: To test broad knowledge and contextual understanding, the dataset should cover a wide range of topics rather than narrow or highly specific domains.

To meet these requirements, we construct `WHOOPS-AHA!`, a dataset specifically designed to support mechanistic interpretability techniques for VLMs. To the best of our knowledge, no existing dataset combines these characteristics, making `WHOOPS-AHA!` a necessary resource for studying controlled knowledge conflicts in multimodal models. Although designed for our experiments, it may also benefit the broader community interested in the mechanistic analysis of multimodal conflicts.

3.2 Dataset Construction

WHOOPS-AHA! addresses these requirements by building on the WHOOPS! collection [Guetta et al., 2023], which features 500 visually implausible, semantically rich scenes annotated with textual descriptions and explanations of their underlying anomalies Each example in WHOOPS-AHA! consists of (i) a counterfactual image depicting an unusual scene, (ii) a sentence referring to the image, and (iii) two sets of plausible continuations: (S_{fact}) reflecting common sense knowledge, and (S_{cofa}) consistent with the counterfactual scene represented in the image.

Construction pipeline. For each image in WHOOPS!, we use GPT-4o to generate a sentence that references the anomaly, while remaining consistent with commonsense (factual) completion without visual input. GPT-4o is also prompted to produce a set of plausible factual tokens S_{fact} and visually-grounded counterfactual continuations S_{cofa}. For instance, given an image representing a wolf howling at the sun (see fig. 1), the sentence proposed by GPT-4o is `"The wolf is howling at the"`, S_{fact} = {`"moon", "night",...`} S_{cofa} = {`"sun", "daylight","morning",..`}. Full prompt details are provided in app. G.

Quality control and validation. To ensure dataset quality, we implemented an LLM-as-a-judge approach [Zheng et al., 2023], using GPT-4.1 [OpenAI, 2025] and Gemini-2.5-Flash [Comanici et al., 2025]. Models evaluated each completion for grammatical correctness (1–3 scale) and for alignment with common knowledge or visual anomalies (1–5 scale). Across the dataset, the average grammatical score was 2.94 ± 0.25 for completions of inner knowledge and 2.93 ± 0.28 for completions aligned with the image. Alignment with knowledge or visual anomalies received a mean score of 4.43 ± 0.97 and 4.69 ± 0.92, respectively.

To validate this setup, we compared LLM ratings with those of 2 human evaluators on a 20-item subset. Full details, including prompts, scoring instructions, and agreement results, are provided in app. B.

4 Background and Methods

4.1 Model Architectures

A VLM encodes image-text tokens with a visual encoder and text embeddings, propagating the resulting residual stream through layers with attention and MLP blocks. The final output is projected to the vocabulary space. We focus our analysis on the residual stream, attention, and MLP blocks, and individual attention heads.

We focus on two models: LLaVA-NeXT-7b [Liu et al., 2024b] and Gemma3-12b [Kamath et al., 2025]. LLaVA-NeXT has 32 layers with 32 attention heads per layer, while Gemma3 has 48 layers with 16 attention heads per layer. Both models use a visual encoder to process image features, but generate only textual output.

4.2 Analytical Tools

Logit inspection. To identify the internal components of VLMs responsible for the competition between inner knowledge and conflicting visual context, we apply the *Logit Lens* technique [Nostalgebraist, 2020], which projects intermediate hidden representations into the vocabulary space. This approach has been used in previous work to analyze token-level information flow [Nanda et al., 2023, Halawi et al., 2023, Yu et al., 2023, Ortu et al., 2024] in LLMs. In our setting, we apply the Logit Lens to the last token of the prompt and extract the logits corresponding to the tokens in S_{fact} and S_{cofa} in the output of the MLP and Attention blocks, and for all the attention heads of the model to identify the components that contribute to the promotion of one mechanism over the other.

Targeted intervention on attention heads. To test the causal role of specific attention heads in promoting predictions aligned with either factual inner knowledge or counterfactual visual context, we intervene on their attention patterns during inference. We define two groups of heads based on Logit Inspection: factual heads ($\mathcal{H}_{\text{fact}}$), which favor predictions based on inner knowledge, and counterfactual heads ($\mathcal{H}_{\text{cofa}}$), which favor visually grounded alternatives. We apply a multiplicative intervention to their attention weights at the final token position (i.e., the last row of the attention matrix), after the softmax operation. Let $\mathbf{A}^{hl}_{\text{last}} = [\mathbf{A}^{hl}_{\text{last,img}}, \mathbf{A}^{hl}_{\text{last,text}}]$ denote the last row of the attention weights for head h at layer l, divided between image and text tokens. The intervention is defined as

$$\mathbf{A}^{hl}_{\text{last,img}} \leftarrow (1 + \lambda) \cdot \mathbf{A}^{hl}_{\text{last,img}} \tag{1}$$

if $(h, l) \in \mathcal{H}_{\text{cofa}}$, and

$$\mathbf{A}^{hl}_{\text{last,text}} \leftarrow (1 - \lambda) \cdot \mathbf{A}^{(hl)}_{\text{last,text}} \tag{2}$$

if $(h, l) \in \mathcal{H}_{\text{fact}}$.

This targeted and bidirectional intervention alters the flow of information in a controlled way, allowing us to test whether modulating the influence of these heads changes the model predictions toward the factual or counterfactual outcome.

Identification of conflict-inducing visual tokens. To isolate the visual tokens responsible for introducing counterfactual information that conflicts with the inner knowledge of the model, we apply

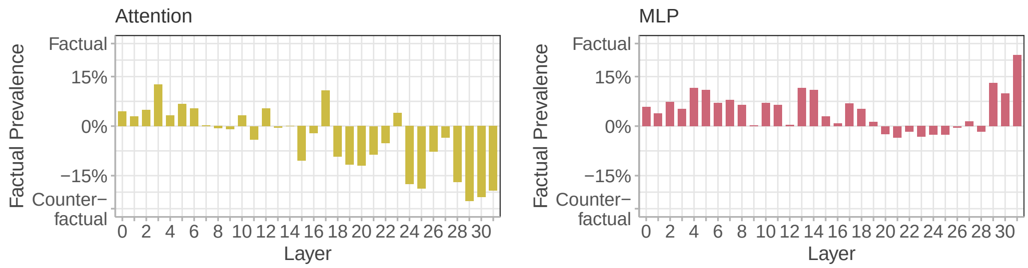

Figure 2: **Factual prevalence in attention and MLP blocks.** Factual prevalence of LLaVA-NeXT shows whether each block favors predictions aligned with factual knowledge (positive) or counterfactual visual context (negative). The results reveal a functional distinction: attention blocks tend to support counterfactual information (**top**), whereas MLP blocks frequently promote the model's internal knowledge (**bottom**).

two methods. Both are based on a threshold parameter $\tau \in [0, 1]$, which controls the sensitivity of token selection.

1. **Most-Attended Visual Tokens:** Given a set of attention heads, we select the visual tokens that receive at least τ times the maximum attention weight within each head. We then take the union of these tokens across all heads.
2. **Gradient-Based Token Importance:** We compute the gradient of the logit associated with a target token (e.g., from S_{fact} or S_{cofa}) with respect to the input visual token embeddings. Visual tokens whose gradient magnitudes exceed τ times the maximum are selected as influential.

By varying τ, we control how many image patches are selected—from none when τ is 1, to all when τ is 0. This allows us to ablate different image portions and analyze how they affect the model predictions.

5 Experimental Results

5.1 Inducing the Conflict between Inner Knowledge and Visual Context

Given a model, we select t_{fact} as the highest probability token in S_{fact} using text-only prompts, and t_{cofa} as the highest probability token from S_{cofa} using multimodal input. Selecting from these sets ensures that we capture the completions most aligned with the model's internal knowledge (text-only) or most influenced by visual information (multimodal), allowing us to reliably study the interaction and potential conflicts between the two sources of information. For example, `"The wolf is howling at the"` yields $t_{\text{fact}} =$ "moon" (with probability of 78% and 100% in LLaVA-NeXT and Gemma3 respectively) in text-only mode, but shift to $t_{\text{fact}} =$ "sun" (26% LLaVA-NeXT, 44% Gemma3) when the image is included, while the probability of `moon` drops to 17% and 0.02%. After filtering ambiguous cases where counterfactual tokens dominate in text-only scenarios, we retain 436 examples for LLaVA-NeXT and 432 for Gemma3. The systematic shift from factual to counterfactual predictions (factual accuracy drops to 27% and 24%, respectively) confirms that visual input successfully overrides internal knowledge. This setup ensures that the image introduces a counterfactual signal that conflicts with the model's inner knowledge, allowing us to analyze how visual input alters the model's prediction compared to its default behavior based on factual knowledge alone.

5.2 The Tension Between Inner Knowledge and Visual Context is Localized

Building on the controlled knowledge conflict, we apply Logit Lens to identify which model components mediate the competition between t_{fact} and t_{cofa}. For attention and MLP blocks, we report factual preference strength—the deviation from a random baseline (0.5) in factual accuracy—where positive values indicate bias toward internal knowledge and negative values toward visual context.

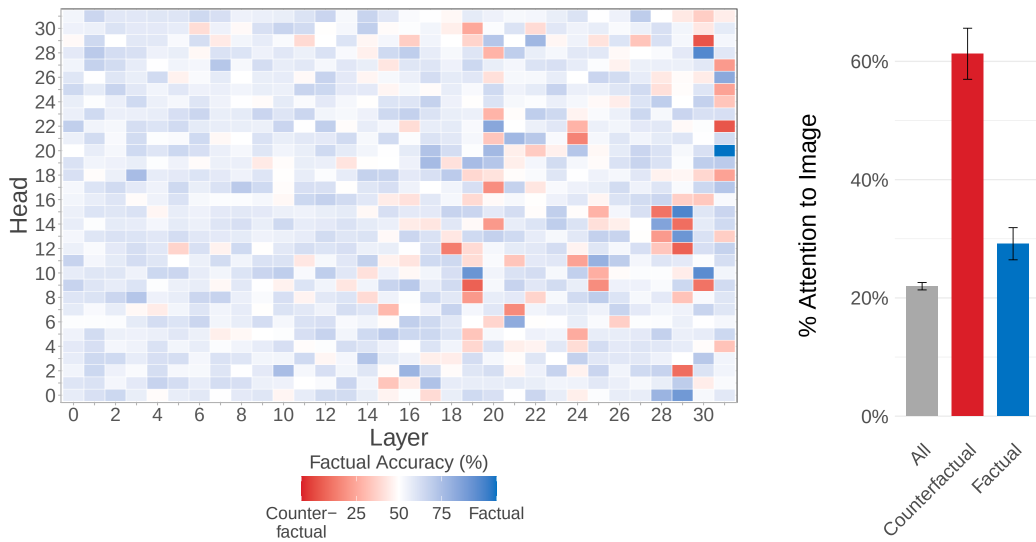

Figure 3: **Contribution of attention heads to factual and counterfactual predictions.** (**Left**) Factual accuracy of individual attention heads in LLaVA-NeXT, based on Logit Lens projections at the final token position. Blue indicates heads that tend to favor the factual token (reflecting inner knowledge), while red indicates heads that favor the counterfactual token (introduced by the visual context). (**Right**) Mean attention to image tokens at the final generation step for heads in each group. Each group contains 20 attention heads. Counterfactual heads attend significantly more to the image (60%) than factual heads (28%) or the model-wide average (22%), indicating that visual information is directly propagated to the output and plays a key role in counterfactual predictions.

For individual attention heads, we report raw factual accuracy (the fraction of examples where factual logits exceed counterfactual logits) to identify heads with strong directional preferences.

Functional separation between attention and MLP layers. We first compare attention and MLP contributions to predicting t_{fact} and t_{cofa} (Figure 2 for LLaVA-NeXT; see app. D for Gemma3). Attention blocks exhibit a stronger tendency to favor the counterfactual visual context, whereas MLP blocks are more aligned with the internal factual knowledge. In particular, the influence of attention blocks increases from the middle layers (around layer 15), peaking in the final four layers. MLP blocks similarly show their strongest alignment to factual knowledge in the upper layers, with a peak at the final layer, consistent with prior findings on upper-layer MLPs retrieving factual knowledge [Geva et al., 2021, Meng et al., 2022, Dai et al., 2022].

Localization of the modality conflict to individual attention heads. We next examine the role of individual attention heads. Figure 3-left shows the tendency for each attention head to promote or suppress the factual token in LlaVa-NeXT (see fig. 9 for Gemma3). The distribution shows that only a small subset of heads exhibit a strong, consistent alignment with t_{fact} or t_{cofa}. Moreover, consistent with the results at the block level, these factual and counterfactual heads are concentrated in the final layers of the model, indicating that the conflict between inner knowledge and visual context is resolved late in the forward pass. In the analyses that follow, we focus on the 20 attention heads that most strongly promote the factual and counterfactual tokens. We chose 20 heads as this provides an optimal balance: it maximizes factual accuracy while minimizing potential disruptions to model stability that could arise from intervening in too many heads (see app. E). On average, the factual heads favor the t_{fact} 85% of the time, and the counterfactual ones t_{cofa} 15% of the time, indicating strong alignment with their respective information sources.

Factual and counterfactual heads exhibit distinct visual attention patterns. We then investigate whether heads associated with the factual mechanism or the counterfactual visual context exhibit distinct attention patterns – specifically, whether they attend to different token modalities (image or text). Since the counterfactual information is introduced through the image, a natural hypothesis

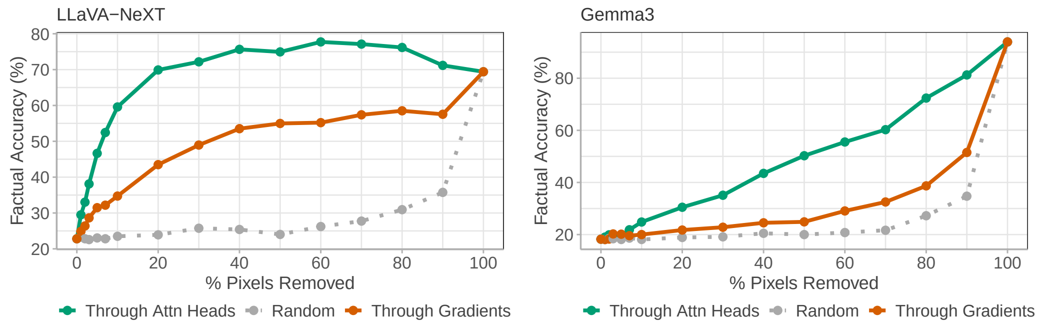

Figure 4: **Intervention on target attention heads.** Change in factual accuracy under different levels of intervention strength (λ). For $\lambda < 0$, we boost the counterfactual heads (on image tokens) and weaken the factual heads (on text tokens); for $\lambda > 0$, we do the opposite. The intervention is applied at the final token position, modifying only the relevant attention values in the last row.

is that counterfactual heads attend more strongly to visual tokens, while factual heads rely more on textual content. To test this hypothesis, for each group of heads, we sum the attention weights assigned to visual tokens in the last row of each head and average across the dataset. Figure 3-right reports the average amount of attention to the image for the two groups of heads. Heads favoring the counterfactual token t_{cofa} attend to image tokens significantly more (61%) than those aligned with inner knowledge (29%) or the model-wide average (22%). Although the counterfactual signal originates in the image, it is not a priori clear that this information is transmitted directly to the final token. The model could, in principle, diffuse or encode this signal in different positions across intermediate layers. However, the observed attention patterns suggest that the visual context influences the output token directly in late layers of the model, with limited intermediate processing. These findings are consistent for Gemma3, and we report the analysis in app. D.

5.3 Targeted Intervention on Selected Attention Heads Causally Shifts Model Behavior

Having identified attention heads aligned with either factual knowledge or counterfactual visual context, we next examine whether these components play a causal role in shaping model predictions. To this end, guided by our earlier observation that counterfactual heads attend more to visual tokens, we apply the targeted bidirectional intervention strategy described in sec. 4.2 that selectively adjusts attention values based on head type and token modality, modifying the attention weights to steer the output of the model towards one mechanism or the other. As a control experiment to isolate the effect of targeted interventions, we randomly select 100 attention heads and apply the same intervention for varying λ values. This manipulation does not produce a substantial deviation from the baseline. The complete results for the control experiment are reported in app. E.

Figure 4 shows the results of our intervention for LLaVA-NeXT (orange profile) and Gemma3 (green profile). When we increase attention from factual heads and decrease it from counterfactual heads using LLaVA-NeXT, the factual accuracy increases to 74%, indicating a strong shift towards predictions of inner knowledge. Conversely, reversing the intervention reduces the accuracy to 16%, confirming that these heads causally influence whether the model favors factual or counterfactual content. A similar trend can be observed for Gemma3, with an even stronger relative effect driven by its lower baseline factual accuracy of 18% and a peak of 83%. Comprehensive details about the choice of the parameter λ are reported in app. F.

5.4 Counterfactual Predictions Depend on Localized Image Regions

The previous analysis reveals that specific attention heads at the final token position mediate the conflict between contextual information and internal knowledge, with heads aligned with the visual context strongly attending to image tokens, injecting visually grounded information into the generation process. However, two key questions remain open. (i) Is the counterfactual visual signal localized to specific image regions or spread across the input? (ii) Is the visual signal passed directly to the last token position, or is it mediated by successive layers and tokens before reaching the output in

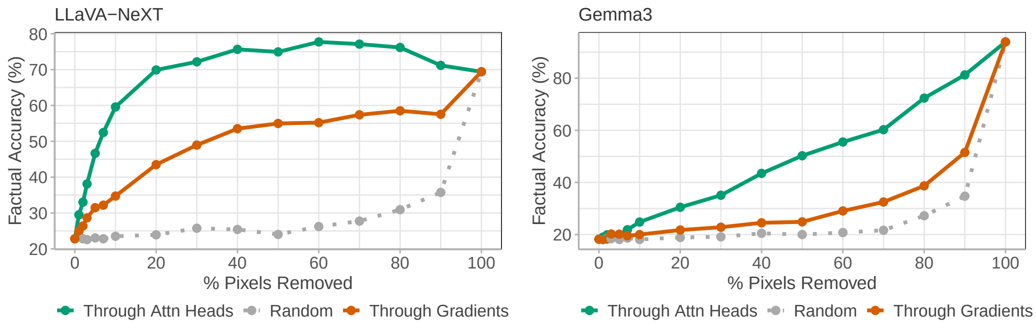

Figure 5: **Ablation of relevant pixels.** The plot shows the effect of ablating different percentages of image pixels in LLaVA-NeXT. The green line corresponds to pixels selected based on the highest attention from counterfactual heads, while the orange line corresponds to pixels with the highest gradient magnitude with respect to the counterfactual token. The gray line shows a random baseline where pixels are removed uniformly at random.

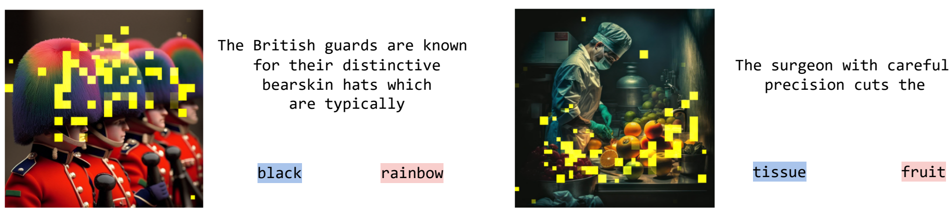

Figure 6: **Visual regions driving counterfactual predictions.** Highlighted image regions, identified through attention-based attribution, show the most influential visual patches for counterfactual predictions. In both examples, the model generates visually grounded but factually incorrect tokens (e.g., rainbow, fruit) instead of commonsense alternatives (black, tissue). The highlighted areas align with semantically meaningful, visually anomalous content, indicating that counterfactual outputs are grounded in localized image features.

the upper layers? To address these, we conduct two analyses: (i) using attention and gradient-based attribution to identify the image patches driving counterfactual predictions, as described in sec. 4.2; and (ii) ablating these patches by setting their visual token embeddings to zero and measuring the change in factual accuracy. A control experiment is also performed where an equivalent number of randomly selected patches are ablated.

Quantitative analysis of patch attribution and ablation. The results (Figure 5) show that ablating patches identified through attention-based attribution leads to a sharp and consistent increase in factual accuracy as more pixels are removed (green profiles). For LLaVA-NeXT, factual accuracy improves markedly with the ablation of just 10–30% of the top-ranked patches and eventually plateaus around 80%. Gradient-based attribution (shown in red) also yields a substantial increase in factual accuracy, but with less pronounced effects, suggesting lower precision in identifying counterfactual-driving regions. In contrast, ablating an equivalent number of randomly selected patches results in only minor fluctuations in accuracy. These findings confirm the causal role of the identified regions and support the hypothesis that counterfactual signals are spatially localized and semantically specific.

Qualitative analysis of visual attribution. To assess the semantic coherence of the identified visual regions, we also qualitatively examine examples where attribution methods highlight specific patches as responsible for counterfactual predictions (see fig. 6. In many cases, these regions correspond to intuitive scenes that directly contradict commonsense knowledge, such as unusual objects, implausible substitutions, or visual features that override typical textual expectations. For instance, when the model predicts "rainbow" instead of "black" for a bearskin hat, the highlighted patches focus on the hat's unrealistic coloring (fig. 6-left). Similarly, when "fruit" replaces "tissue" in a surgical scene, the attention centers on the bright, unexpected presence of oranges on the operating table (fig. 6-right). These observations confirm that counterfactual outputs are grounded in localized, semantically meaningful image features.

6 Discussion

Our work extends the interpretability framework of Ortu et al. [2024] to VLMs, exploring how they handle conflicts between visual input and internal knowledge. Previous studies, such as Liu et al. [2024c] and Han et al. [2024], have constructed diagnostic benchmarks to measure the susceptibility of the model to conflicting visual and textual cues; however, these works primarily focus on evaluating the model output without analyzing in-depth the internal conflict resolution mechanisms. Likewise, datasets such as HallusionBench [Guan et al., 2024] and PhD [Liu et al., 2025] differ significantly from our dataset in their goals and methodologies: HallusionBench utilizes carefully controlled image-question pairs to systematically induce hallucinations, while PhD employs extensive synthetic generation to broadly cover diverse hallucination patterns. In contrast, our WHOOPS-AHA! dataset deliberately induces controlled knowledge conflicts by pairing visually anomalous scenes with commonsense prompts, emphasizing mechanistic analysis over general hallucination detection for deeper mechanistic insights. Our approach aligns with Golovanevsky et al. [2025b] in terms of mechanistic interpretability but diverges in its focus on resolving conflicts between visual input and internal knowledge. While Golovanevsky et al. [2025b] identify attention heads with broad roles across tasks, we pinpoint specific attention heads involved in conflict resolution and demonstrate their causal role through targeted interventions. Unlike broad hallucination mitigation methods [Liu et al., 2024a, Leng et al., 2023], our work emphasizes precise localization of conflict triggers and inference-time interventions. Although not a ready-to-deploy solution, our analysis lays the groundwork for developing targeted, interpretable interventions in multimodal models.

7 Conclusion

We investigated how VLMs internally resolve conflicts between visual input and parametric knowledge through mechanistic analysis. Our WHOOPS-AHA! dataset enables controlled study of these conflicts by pairing counterfactual images with commonsense textual prompts. Three key findings emerge from our analysis: (i) multimodal conflicts localize to a small set of attention heads in the model's upper layers, with distinct functional roles. Counterfactual heads attend primarily to visual tokens while factual heads focus on textual content; (ii) targeted interventions on these heads causally shift model predictions between knowledge-based and visually-grounded outputs; (iii) attention patterns from conflict-resolution heads provide more precise visual attribution than gradient-based methods, identifying semantically meaningful image regions responsible for counterfactual predictions. These results contribute to a mechanistic understanding of multimodal reasoning and provide a foundation for developing more interpretable and controllable VLMs. Our approach demonstrates that complex multimodal behaviors can be traced to specific, identifiable components, opening pathways for targeted interventions in scenarios where visual input conflicts with model knowledge.

Limitations

Methodological limitations. The analysis relies on the Logit Lens technique to project intermediate hidden states into token logits. Although this method has been widely adopted for interpretability, it is known to introduce distortions due to projection from non-final residual states [Belrose et al., 2023], and should be interpreted as an approximate diagnostic rather than a precise decoding proxy. In our setting, we use a representative factual and counterfactual token per example to enable controlled comparisons. Although this simplifies the generative landscape of the model, it offers a practical and interpretable probe of the underlying mechanisms. Future work could explore more model behavior across full generations to complement this approach. Our attribution and intervention methods focus on attention heads and target the final token position. This design isolates interpretable causal signals while remaining tractable, though it does not capture the possible contributions of other components, such as MLP layers or visual encoders. Extending this framework to broader architectural elements is a promising direction.

Scope and generalizability. We focus on late-fusion, LLaVA-style architectures, which are particularly well-suited for controlled image-understanding tasks. These models are among the best open-source architectures for visual understanding, making them ideal for the interpretability methods employed in our study. Our interest is specifically in how visual input interacts with internal knowledge during textual generation, so we chose models that are designed with a focus on image understanding. While early or mid-fusion models also use attention to integrate visual features into the language stream, they may differ significantly in how information is communicated between the modalities [Serra et al., 2024]. The point of injection of visual features varies, but the underlying mechanism of cross-modal communication through attention remains consistent across these models. By focusing on late-fusion models, we ensure a more controlled and traceable examination of visual-to-text interactions in widely used open source multimodal models, though this choice limits the generalizability of our findings to models with different fusion strategies. Finally, the WHOOPS-AHA! dataset is constructed from synthetic and curated inputs, which allow precise manipulation of visual-textual conflict. Although this setting facilitates analysis, future extensions to more naturalistic data could further validate the findings in less constrained contexts.

Ethical Considerations

This work aims to improve our understanding of how VLMs resolve conflicts between internal factual knowledge and contradictory visual context. Our analysis is intended to contribute to foundational research in model interpretability, with the broader goal of developing more transparent and controllable multimodal systems. The techniques presented are diagnostic and exploratory in nature, designed to support responsible development and analysis of multimodal systems. We believe that studying the dynamics of conflicting information sources is essential for anticipating model failure modes, mitigating unintended behaviors, and building more robust AI systems. All models and data are used in accordance with their intended research licenses, and WHOOPS-AHA! is released solely for non-commercial, research purposes under compatible terms. We used AI assistants (e.g., GitHub Copilot) to support code completion during experiment implementation; all generated code was manually reviewed and supervised by the authors.

Acknowledgments

The authors acknowledge the AREA Science Park supercomputing platform ORFEO made available for conducting the research reported in this paper and the technical support of the Laboratory of Data Engineering staff. Francesco Ortu, Diego Doimo and Alberto Cazzaniga were supported by the project “Supporto alla diagnosi di malattie rare tramite l’intelligenza artificiale" CUP: F53C22001770002 and “Valutazione automatica delle immagini diagnostiche tramite l’intelligenza artificiale", CUP: F53C22001780002. Alberto Cazzaniga was supported by the European Union – NextGenerationEU within the project PNRR “PRP@CERIC" IR0000028 - Mission 4 Component 2 Investment 3.1 Action 3.1.1.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf.

Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding Information Storage and Transfer in Multi-modal Large Language Models, June 2024. URL http://arxiv.org/abs/2406.04236. arXiv:2406.04236.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *CoRR*, abs/2303.08112, 2023. doi: 10.48550/ARXIV.2303.08112. URL https://doi.org/10.48550/arXiv.2303.08112.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit S. Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilaï Deutel, Nam Nguyen, Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé, Sagar Waghmare, Helen Miller, Nathan Byrd, Ashrith Sheshan, Raia Hadsell Sangnie Bhardwaj, Pawel Janus, Tero Rissa, Dan Horgan, Sharon Silver, Ayzaan Wahid, Sergey Brin, Yves Raimond, Klemen Kloboves, Cindy Wang, Nitesh Bharadwaj Gundavarapu, Ilia Shumailov, Bo Wang, Mantas Pajarskas, Joe Heyward, Martin Nikoltchev, Maciej Kula, Hao Zhou, Zachary Garrett, Sushant Kafle, Sercan Arik, Ankita Goel, Mingyao Yang, Jiho Park, Koji Kojima, Parsa Mahmoudieh, Koray Kavukcuoglu, Grace Chen, Doug Fritz, Anton Bulyenov, Sudeshna Roy, Dimitris Paparas, Hadar Shemtov, Bo-Juen Chen, Robin Strudel, David Reitter, Aurko Roy, Andrey Vlasov, Changwan Ryu, Chas Leichner, Haichuan Yang, Zelda Mariet, Denis Vnukov, Tim Sohn, Amy Stuart, Wei Liang, Minmin Chen, Praynaa Rawlani, Christy Koh, JD Co-Reyes, Guangda Lai, Praseem Banzal, Dimitrios Vytiniotis, Jieru Mei, and Mu Cai. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261, 2025. doi: 10.48550/ARXIV.2507.06261. URL https://doi.org/10.48550/arXiv.2507.06261.

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. Holistic analysis of hallucination in gpt-4v(ision): Bias and interference challenges. *CoRR*, abs/2311.03287, 2023. doi: 10.48550/ARXIV.2311.03287. URL https://doi.org/10.48550/arXiv.2311.03287.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.581. URL https://doi.org/10.18653/v1/2022.acl-long.581.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Jen Dumas, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024.

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

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5484–5495. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EMNLP-MAIN.446. URL https://doi.org/10.18653/v1/2021.emnlp-main.446.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. *CoRR*, abs/2304.14767, 2023. doi: 10.48550/ARXIV.2304.14767. URL https://doi.org/10.48550/arXiv.2304.14767.

Michal Golovanevsky, William Rudman, Michael Lepori, Amir Bar, Ritambhara Singh, and Carsten Eickhoff. Pixels versus priors: Controlling knowledge priors in vision-language models through visual counterfacts. *CoRR*, abs/2505.17127, 2025a. doi: 10.48550/ARXIV.2505.17127. URL https://doi.org/10.48550/arXiv.2505.17127.

Michal Golovanevsky, William Rudman, Vedant Palit, Carsten Eickhoff, and Ritambhara Singh. What do VLMs NOTICE? a mechanistic interpretability pipeline for Gaussian-noise-free text-image corruption and evaluation. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11462–11482, Albuquerque, New Mexico, April 2025b. Association for Computational Linguistics. ISBN 979-8-89176-189-6. URL https://aclanthology.org/2025.naacl-long.571/.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models, 2024. URL https://arxiv.org/abs/2310.14566.

Nitzan Bitton Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. Breaking common sense: Whoops! A vision-and-language benchmark of synthetic and compositional images. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 2616–2627. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00247. URL https://doi.org/10.1109/ICCV51070.2023.00247.

Danny Halawi, Jean-Stanislas Denain, and Jacob Steinhardt. Overthinking the truth: Understanding how language models process false demonstrations. *CoRR*, abs/2307.09476, 2023. doi: 10.48550/arXiv.2307.09476. URL https://doi.org/10.48550/arXiv.2307.09476.

Tianyang Han, Qing Lian, Rui Pan, Renjie Pi, Jipeng Zhang, Shizhe Diao, Yong Lin, and Tong Zhang. The instinctive bias: Spurious images lead to illusion in MLLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16163–16177, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.904. URL https://aclanthology.org/2024.emnlp-main.904/.

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *CoRR*, abs/2305.00586, 2023. doi: 10.48550/arXiv.2305.00586. URL https://doi.org/10.48550/arXiv.2305.00586.

Tianze Hua, Tian Yun, and Ellie Pavlick. How do vision-language models process conflicting information across modalities? *CoRR*, abs/2507.01790, 2025. doi: 10.48550/ARXIV.2507.01790. URL https://doi.org/10.48550/arXiv.2507.01790.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1193–1215, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.70. URL https://aclanthology.org/2024.findings-acl.70/.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Róbert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin

Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucinska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, and Ivan Nardini. Gemma 3 technical report. *CoRR*, abs/2503.19786, 2025. doi: 10.48550/ARXIV.2503.19786. URL https://doi.org/10.48550/arXiv.2503.19786.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomáš Kočiský, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. Mind the gap: Assessing temporal generalization in neural language models. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=73OmmrCfSyy.

Tiep Le, Vasudev Lal, and Phillip Howard. Coco-counterfactuals: Automatically constructed counterfactual examples for image-text pairs. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/e14e4cb8266184ceb234973dfe07faed-Abstract-Datasets_and_Benchmarks.html.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding, 2023. URL https://arxiv.org/abs/2311.16922.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/li22n.html.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=J44HfH4JCg.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. Phd: A chatgpt-prompted visual hallucination evaluation dataset, 2025. URL https://arxiv.org/abs/2403.11116.

Xiaoyuan Liu, Wenxuan Wang, Youliang Yuan, Jen tse Huang, Qiuzhi Liu, Pinjia He, and Zhaopeng Tu. Insight over sight? exploring the vision-knowledge conflicts in multimodal llms, 2024c. URL https://arxiv.org/abs/2410.08145.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. Prompt injection attack against llm-integrated applications, 2024d. URL https://arxiv.org/abs/2306.05499.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. Time waits for no one! analysis and challenges of temporal misalignment. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5944–5958, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.435. URL https://aclanthology.org/2022.naacl-main.435/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=9XFSbDPmdW.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=chanJGoa7f.

Nostalgebraist. interpreting gpt: the logit lens. Accessed: Nov 2023, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

OpenAI. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/, April 14 2025. Accessed: 2025-09-22.

OpenRouter, Inc. Openrouter: The unified interface for llms, 2025. URL https://openrouter.ai/.

Francesco Ortu, Zhijing Jin, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, and Bernhard Schölkopf. Competition of mechanisms: Tracing how language models handle facts and counterfactuals. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 8420–8436. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.458. URL https://doi.org/10.18653/v1/2024.acl-long.458.

Alessandro Serra, Francesco Ortu, Emanuele Panizon, Lucrezia Valeriani, Lorenzo Basile, Alessio Ansuini, Diego Doimo, and Alberto Cazzaniga. The narrow gate: Localized image-text communication in vision-language models. *CoRR*, abs/2412.06646, 2024. doi: 10.48550/ARXIV.2412.06646. URL https://doi.org/10.48550/arXiv.2412.06646.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.486. URL https://aclanthology.org/2024.emnlp-main.486/.

Qinan Yu, Jack Merullo, and Ellie Pavlick. Characterizing mechanisms for factual recall in language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9924–9959, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.615. URL https://aclanthology.org/2023.emnlp-main.615.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

A Reproducibility

We run the experiments on one NVIDIA H100 GPU, and two GPUs for the gradient-based attribution tests. We use the HuggingFace Transformers library [Wolf et al., 2020] with public implementations of LLaVA-NeXT and Gemma3. The total compute time is 15 GPU hours. The WHOOPS! dataset was released with a CC-By 4.0 license.

B LLM-as-a-Judge: Detailed Validation and Analysis

B.1 Evaluation Setup

We used GPT-4.1 (gpt-4.1-2025-04-14) and Gemini-2.5-Flash (gemini-2.5-flash-image-preview) thorugh OpenRouter [OpenRouter, Inc., 2025] to evaluate each dataset completion along two dimensions:

- Grammatical correctness:
  - 1 (No) All completed sentences contain grammatical errors.
  - 2 (Some do not make sense) Some completed sentences have grammatical errors.
  - 3 (Yes) All completed sentences follow proper grammar rules.
- Knowledge/Anomaly Alignment (1–5 scale)
  - 1 = Poor alignment: Completion ignores or misrepresents common knowledge or visual anomalies
  - 5 = Excellent alignment: Completion clearly reflects correct knowledge or accurately captures anomalies in the image.

B.2 Aggregate Statistics

Metric	Mode	Gemini-2.5-Flash	GPT-4.1	Average	Exact Agreement
Grammatical Correctness	Text-only	2.95 ± 0.23	2.93 ± 0.26	2.94 ± 0.25	95.0%
	With Image	2.92 ± 0.30	2.93 ± 0.26	2.93 ± 0.28	92.8%
Knowledge/Anomaly Alignment	Text-only	4.55 ± 0.98	4.31 ± 0.94	4.43 ± 0.97	69.5%
	With Image	4.76 ± 0.93	4.60 ± 0.91	4.68 ± 0.92	80.2%

Table 1: **LLM-as-a-Judge evaluation results.** Gemini-2.5-Flash and GPT-4.1 for Text-only and Image-based Scenarios

We report the average results for all dimensions in Tab. 1. The results shown a that the judges rate the sentence and the completions mostly grammatical correct and aligned with the common knowledge or the anomaly in the image.

To measure agreement between the two judges, we calculate the exact agreement rate. We observed a strong ceiling effect in our data, with the distribution of responses being heavily imbalanced. Overall, the results demonstrate that the LLM as a judge setup produces consistent and interpretable ratings across both factual and counterfactual scenarios, with high alignment between models (Gemini vs. GPT-4.1). These findings support the reliability of LLM-based evaluation for large-scale dataset validation.

B.3 Human Validation

To assess how well the LLM-as-a-judge setup reflects human evaluation, we sampled 20 entries from our dataset and asked 2 human evaluators, both graduate students at a local institution, to complete the same tasks as the LLMs. To measure agreement between human evaluators and LLM judgments, we averaged the ratings of the human evaluators and rounded to the nearest integer for each category, and then measured the exact and adjacent agreement rates.

For grammatical correctness in text-based completions, we observed 60% agreement, and 70% for image-associated completions. For alignment, the agreement was 35% for text-only completions and

65% for image-associated completions. However, when considering adjacent category agreement (i.e., when responses were within one point of each other on the scale), the agreement increased to 85% for text-based completions and 100% for image-associated completions.

B.4 Prompt and Instruction

Prompt Used to Instruct LLM-as-a-Judge with Image

You will evaluate sentence completions that should describe UNUSUAL or ANOMALOUS elements visible in the provided image. Focus on identifying completions that capture strange, unexpected, or contradictory visual elements.
EVALUATION CRITERIA:
1. GRAMMATICAL CORRECTNESS: Are the completed sentences grammatically valid?
Check syntax, word order, agreement, and grammatical structure of each completed sentence.
→ "Yes" = ALL completed sentences follow proper grammar rules → "Some do not make sense" = SOME completed sentences have grammatical errors → "No" = ALL completed sentences contain grammatical errors
Focus on: subject-verb agreement, article usage, word order, tense consistency
2. KNOWLEDGE ALIGNMENT: How well do completions reflect unusual/anomalous visual elements in the image? IMAGE CONTEXT: Rate how accurately completions describe the strange/unusual elements you can SEE in the image.
SCORING RUBRIC: • 5 = Perfectly captures unusual/anomalous visual elements in the image - highly accurate and relevant • 4 = Mostly accurate - minor gaps or imprecisions • 3 = Partially accurate - some correct elements, some missing/wrong • 2 = Minimally accurate - mostly incorrect with few relevant elements • 1 = Completely inaccurate - fails to reflect unusual/anomalous visual elements in the image
SENTENCE: "{text}" COMPLETIONS TO EVALUATE: {tokens}
INSTRUCTION: Look carefully at the image and identify what appears unusual, unexpected, or anomalous. Then evaluate how well the completions capture these strange elements.
IMPORTANT: You must respond with valid JSON only. Do not include any text before or after the JSON.
Required JSON format: {{ "grammatically_correct": "Yes, all are grammatically correct" | "Some are not grammatically correct" | "No, they are not grammatically correct", "knowledge_reflection_score": 1-5, "reasoning": "Explain your evaluation for each criterion, citing specific completions" }}"""

Prompt Used to Instruct LLM-as-a-Judge without image

You will evaluate sentence completions that should reflect normal, expected real-world knowledge real-world scenarios based on common knowledge and typical experiences.
EVALUATION CRITERIA:
1. GRAMMATICAL CORRECTNESS: Are the completed sentences grammatically valid?
Check syntax, word order, agreement, and grammatical structure of each completed sentence.
→ "Yes" = ALL completed sentences follow proper grammar rules → "Some do not make sense" = SOME completed sentences have grammatical errors → "No" = ALL completed sentences contain grammatical errors
Focus on: subject-verb agreement, article usage, word order, tense consistency
2. KNOWLEDGE ALIGNMENT: How well do completions reflect "normal, expected real-world knowledge? REAL-WORLD CONTEXT: Rate how well completions reflect typical, widely-accepted real-world scenarios.
SCORING RUBRIC: • 5 = Perfectly captures normal, expected real-world knowledge - highly accurate and relevant • 4 = Mostly accurate - minor gaps or imprecisions • 3 = Partially accurate - some correct elements, some missing/wrong • 2 = Minimally accurate - mostly incorrect with few relevant elements • 1 = Completely inaccurate - fails to reflect normal, expected real-world knowledge

SENTENCE: "{text}" COMPLETIONS TO EVALUATE: {tokens}
Consider what would be normal, expected, and typical in real-world scenarios. Then evaluate how well the completions reflect this common knowledge.
IMPORTANT: You must respond with valid JSON only. Do not include any text before or after the JSON.
Required JSON format: {{ "grammatically_correct": "Yes, all are grammatically correct" | "Some are not grammatically correct" | "No, they are not grammatically correct", "knowledge_reflection_score": 1-5, "reasoning": "Explain your evaluation for each criterion, citing specific completions" }}"""

Instruction Given to Human Evaluator

Task Instructions You will be asked to evaluate sentences and their possible completions. Sometimes, an image will also be provided. Your job is to judge whether the completions are appropriate, whether the sentence is grammatically correct, and how well the sentence and completions reflect knowledge or the content of the image.
Please follow these criteria carefully for each question:
1. Is the sentence grammatically correct?
Ignore meaning here; only focus on grammar and syntax. Mark "Yes" if the base sentence with all the possible completions is grammatically well-formed. Mark "Some do not make sense" if at least one completion create a grammatically incorrect sentence. Mark "No" if the sentence with all the possible completions contains grammar errors. 2. How well do the sentence and completions reflect common knowledge (or reflect strange/anomalous things in the image)?
If the question refers to common knowledge: Judge how typical, reasonable, or widely accepted the sentence + completion is.
Example: *"The sun rises in the east"* should score high (5).
Example: *"The sun rises in the north"* should score low (1).
If the question refers to strange/anomalous things in the image: Judge how accurately the sentence and completions capture unusual, odd, or unexpected elements visible in the image. Score higher if the completion clearly reflects what is strange in the image.
Score lower if it ignores or misrepresents the anomaly. Use the 1–5 scale consistently:

1 = Not at all accurate/appropriate
3 = Neutral or partially accurate
5 = Very accurate and appropriate
Please read each question carefully and provide your evaluations with attention.

C MLP Intervention

We also tested whether intervening on MLP blocks could produce effects comparable to those observed with attention heads. Specifically, we applied interventions to the last three MLP blocks at the final token position in both LLaVA-NeXT and Gemma3. The results, reported in Figure 7, show only marginal changes in factual accuracy relative to the baseline. This effect is substantially weaker than the gains obtained from targeted attention-head interventions (Figure 4).

These findings reinforce two key conclusions. First, MLP interventions are less precise: they modify the residual stream broadly, affecting a much larger number of parameters and acting indiscriminately across modalities. This broad influence makes it harder to isolate causal mechanisms and increases the risk of introducing unintended side effects. Second, the limited efficacy of MLP interventions indicates that factual–counterfactual conflicts are primarily mediated by attention mechanisms, not by MLP transformations. This aligns with prior evidence that late-layer MLPs often retrieve factual associations, whereas attention heads are more directly responsible for integrating conflicting cross-modal signals.

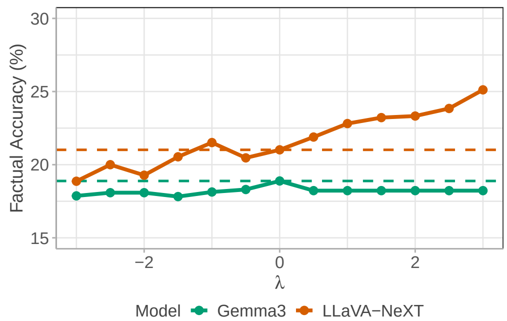

Figure 7: **Effect of MLP interventions.** Factual accuracy when intervening on the last three MLP blocks at the final token position in LLaVA-NeXT and Gemma3. The observed improvements are minor compared to targeted attention-head interventions (Figure 4). This suggests that MLP interventions are less effective and less precise, reinforcing our decision to focus on attention mechanisms as the main locus of conflict resolution.

For these reasons, our analysis centers on attention interventions, which provide both stronger causal leverage and more interpretable control over the balance between internal knowledge and visual input.

D Experimental Analysis for Gemma-12b

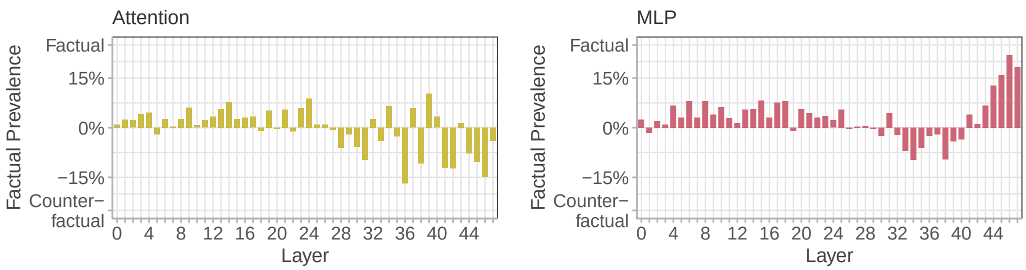

Figure 8: **Factual and counterfactual contributions of MLP and attention blocks in gemma3.** Layer-wise deviation from 50% factual accuracy for attention and MLP blocks, as measured by the relative logits of t_{fact} and t_{cofa} via Logit Lens. Positive values indicate a bias toward the factual token, while negative values indicate preference for the counterfactual token. Consistent with trends observed in LLaVA-NeXT, attention blocks in Gemma3 increasingly support counterfactual predictions in higher layers, while MLP blocks show stronger alignment with internal factual knowledge.

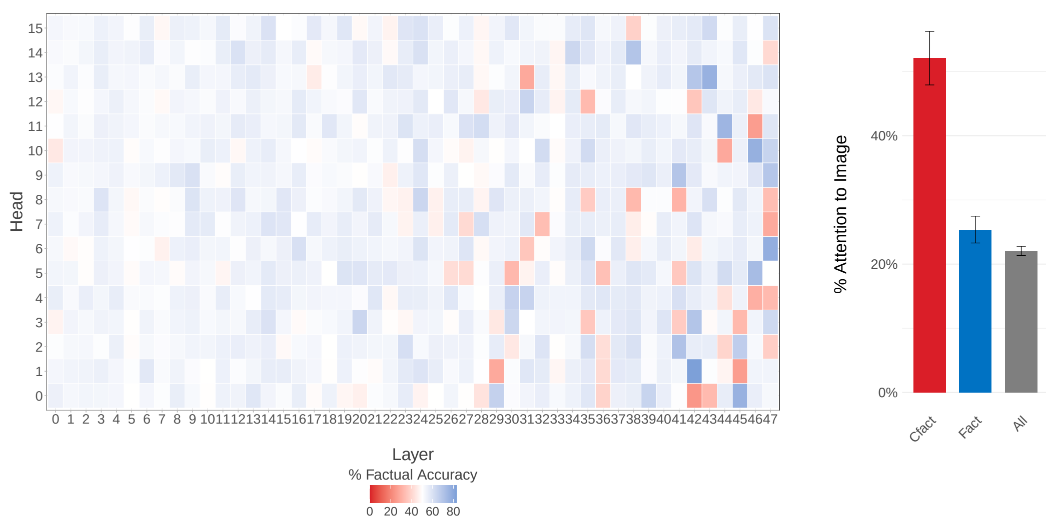

Figure 9: **Factual and counterfactual contributions of attention heads for gemma3.** (**Left**) Factual accuracy of individual attention heads in Gemma3, computed using Logit Lens projections of the final token's hidden state. Blue indicates heads that more frequently favor the factual token (t_{fact}), while red indicates those that favor the counterfactual token (t_{cofa}). As in LLaVA-NeXT, highly polarized heads are concentrated in the upper layers. (**Right**) Mean attention to image tokens at the final generation step. Counterfactual heads attend more strongly to image tokens (52%) than factual heads (25%) or the model-wide average (22%), highlighting the direct role of visual input in modulating counterfactual predictions.

Image	λ	Caption
	0	The image is a digital artwork of a young boy with a contemplative expression. He has short, light brown hair and striking blue eyes. The boy is wearing a striped shirt with a collar and a patterned tie.
	3	The image is a digital artwork of a young child. The child is depicted with a contemplative expression, looking slightly to the side with a thoughtful gaze. They are holding a piece of paper or a small object in their hand, which appears to...
	10	Jimmy Wooster spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr spr
	0	The image is a dramatic and evocative artwork depicting a young girl standing in the center, holding a flag with the colors of the French flag—blue, white, and red.
	3	The image depicts a young girl standing in the center, holding a small, tattered flag with the design of the French flag.
	10	The image shows a Telephone P p p p p p p p p p p p p p p p p p Telephone P p p p p p p p p p p p p p p p p p Telephone P p p p

Table 2: **Effect of intervention strength on caption generation quality.** Examples of captions generated by LLaVA-NeXT under different intervention strengths ($\lambda = 0, 3, 10$). As intervention magnitude increases, captions begin to diverge from the original output. At moderate levels ($|\lambda| = 3$), outputs remain coherent but show lexical and structural variations. At high levels ($|\lambda| = 10$), generations often degrade into repetitive or nonsensical sequences.

E Details on the Number of Heads Selected and Control Experiment

Figure 10 reports the control experiment, where the intervention was applied to 100 randomly selected attention heads. The results show no measurable change in factual accuracy, confirming that improvements are not due to random head selection but to the specific heads identified in our method.

Figure 11 examines the effect of varying the number of intervened attention heads, with intervention strength fixed at $\lambda = 3$. We observe that factual accuracy increases as the number of heads grows, reaching its peak at 20 heads. Beyond this point, further interventions do not yield additional gains and may introduce instability. This demonstrates that intervening on 20 heads provides the best balance between accuracy improvement and model robustness.

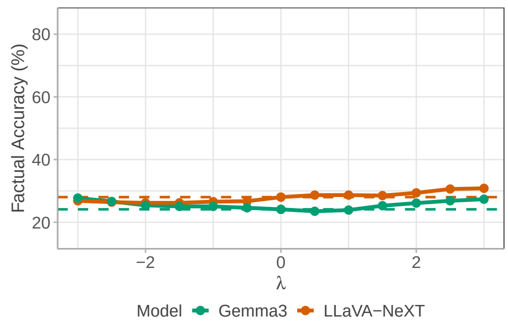

Figure 10: **Intervention on Random Attention Heads.** Change in factual accuracy under varying levels of intervention strength (λ) applied to 100 randomly selected attention heads. The results show no substantial deviation from baseline, confirming the specificity of the identified target heads.

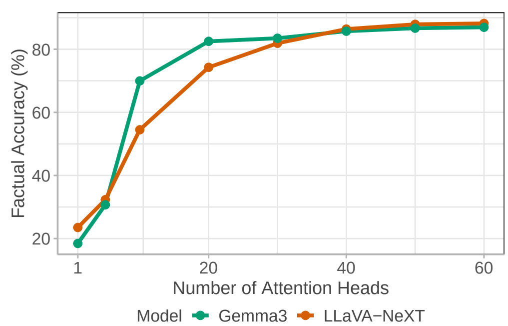

Figure 11: **Effect of intervening on varying numbers of attention heads.** Change in factual accuracy as a function of the number of attention heads involved in the intervention. Each value x indicates that x heads are selected from both the factual and counterfactual groups. Intervention strength is fixed at $\lambda = 3$. The results highlight that intervening on 20 heads provides the optimal trade-off, maximizing factual accuracy without excessively affecting model stability.

F Details on the Intervention Parameter Choice

To ensure plausible interventions, we constrain the scaling parameter to $\lambda \in [-3, 3]$ and monitor the position of the higher-logit token in the full next-token distribution. For example, using LLaVA-NeXT, the average rank of the token t_{fact} shifts from 3 at $\lambda = 0$ to 31 at $\lambda = 3$, indicating that while the intervention is highly effective, it introduces some deviation in the overall logit distribution, an expected effect when strongly modulating internal components.

To further support this choice of intervention range, we also prompt the model to generate captions with and without intervention, and manually inspect the quality of the outputs as we increase the intervention strength, $|\lambda|$. We empirically observe that for $|\lambda|$ greater than three, the quality of the generated captions degrades, and most of the time, they become agrammatical when $|\lambda| > 10$ (see Table 2).

We also attempted to quantify the quality of the generated text after the intervention with a KL-divergence with the generated text before the intervention ($|\lambda| > 0$), which we consider as a reference for a well-structured sentence. Figure 12 shows the average KL-divergence across all examples in `WHOOPS-AHA!` as we increase $|\lambda|$ in LLaVA-NeXT.

The KL divergence sharply increases for $|\lambda| < 3$, and then the growth slows down and stabilizes around $|\lambda| = 12$ for $\lambda < -20$ and 18 for $\lambda > 20$. When the KL is smaller than 10, for λ between -3 and 3, the output sentences have a similar quality to those generated before intervention.

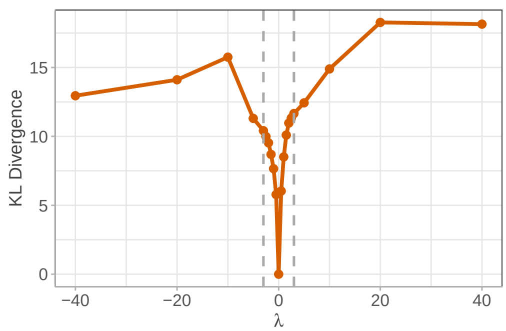

Figure 12: **KL divergence between generated captions at different intervention Strengths in LLaVA-NeXT.** Symmetric increase in KL divergence around $\lambda = 0$, with rapid divergence until $|\lambda| = 3$ and stabilization near $|\lambda| = 10$. Higher intervention magnitudes cause substantial shifts in the generated token distribution, indicating degradation in caption quality.

G Prompts For Dataset Generation

Prompt Used to Generate Dataset Instances.

You are a helpful assistant expert in LLMs research.
Counterfactual Dataset Generation Prompt
Objective: Generate captions for images that highlight a clear contrast between common (factual) and unusual (counterfactual) scenarios involving the subject depicted. Each caption must include the subject of the image and end with "___" indicating the blank space where a single-word token is placed.
Definitions: - **Factual token**: A single word that represents typical, expected behavior or attributes of the main subject shown in the image. - **Counterfactual token**: A single word introducing a surprising, unexpected, or unusual element related explicitly to the same main subject; it makes sense only if the image explicitly illustrates this twist.
Context Provided: For each image, you will receive the following textual information: - Selected Caption: A primary description identifying the main subject clearly. - Crowd Captions: Alternative descriptions from multiple annotators. - Designer Explanation: Explanation emphasizing the unusual or counterintuitive aspect involving the subject. - Crowd Explanations: Multiple explanations focusing on the unusual aspects related directly to the subject of the image.
Task Instructions:
Caption Construction: - Create exactly one neutral sentence (caption) clearly containing the main subject depicted in the image, but avoiding the description of unusual aspects contained in the image. - The sentence must end with an intentional blank ("___"). - Critical Requirement: The caption must compel the model to complete the blank differently based on the context: - **Without the image**: complete with a factual token (typical scenario involving the subject). - **With the image**: complete with a counterfactual token (unexpected scenario explicitly depicted). - Important Constraint: Use neutral language with NO textual hints indicating abnormality. The main subject must explicitly appear

in the caption to establish context clearly. Only the image content itself should disambiguate the scenario. - The caption should not contain any unusual or counterintuitive elements; the unusual aspect should be reflected solely in the image content and in the counterfactual tokens. - Make sure that if you substitute the blank with a factual or counterfactual token, the sentence is fluent and grammatically correct.

Explicit Single-Word Token Generation: - Generate exactly **ten single-word factual tokens** representing common scenarios involving the main subject that could complete in a grammatically correct way the sentence. - Generate exactly **ten single-word counterfactual tokens** representing surprising scenarios involving the same subject, justified solely by the provided image, and that could complete the sentence in a grammatically correct way. - Strictly enforce single-word tokens; no multi-word phrases or sentences. - Ensure clear differentiation without conceptual overlap between factual and counterfactual tokens.

JSON Output Format: Provide each caption and tokens following this exact schema:

{ "caption": "Neutral sentence explicitly containing the main subject and ending with an intentional blank ('___')", "factual_tokens": ["token1", "token2", "token3", "token4", "token5", ...], "counterfactual_tokens": ["token1", "token2", "token3", "token4", "token5", ...], "context": { "selected_caption": "Primary description clearly stating the main subject of the image", "crowd_captions": ["Caption 1", "Caption 2", "..."], "designer_explanation": "Explanation highlighting the unusual aspect directly involving the main subject", "crowd_explanations": ["Explanation 1", "Explanation 2", "..."] } }

Your role is to craft neutral captions explicitly containing the main subject of each image, along with precisely differentiated factual and counterfactual single-word tokens. The explicit presence of the main subject in the caption must guide factual versus counterfactual completions, relying solely on the provided image for disambiguation.

Prompt Used to Generate Factual and Counterfactual Tokens.

You are presented with an image and an incomplete sentence describing its content. The image intentionally portrays an unusual scenario that contrasts typical or factual knowledge.

Your task is to generate two lists of tokens:

1. Factual Tokens (5 tokens): These tokens should represent words or concepts that accurately and typically complete the sentence based solely on common knowledge, without considering the unusual image.

2. Counterfactual Tokens (5 tokens): These tokens should represent words or concepts that correctly complete the sentence when explicitly considering the unusual content depicted in the image, even if it contradicts common factual knowledge.

Please format your response clearly as a JSON object as follows:

“‘json { "sentence": "INCOMPLETE_SENTENCE", "factual_tokens": ["token1", "token2", "token3", "token4", "token5"], "counterfactual_tokens": ["token1", "token2", "token3", "token4", "token5"] } “‘

Choose tokens that clearly differentiate between typical knowledge and the unusual scenario depicted by the provided image.