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

Anonymous Author(s)
Affiliation
Address
email

Abstract

Vision-language models (VLMs) increasingly leverage diverse knowledge sources to address complex tasks, often encountering conflicts between their internal parametric knowledge and external information. Knowledge conflicts can result in hallucinations and unreliable responses, but the mechanisms governing such interactions remain unknown. To address this gap, we analyze the mechanisms that VLMs use to resolve cross-modal conflicts by introducing a dataset of multimodal counterfactual queries that deliberately contradict internal commonsense knowledge. We localize with logit inspection a small set of heads that control the conflict. Moreover, by modifying these heads, we can steer the model towards its internal knowledge or the visual inputs. Finally, we show that attention from such heads pinpoints localized image regions driving visual overrides, outperforming gradient-based attribution in precision.

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 However, these two sources can sometimes contradict each other, for example when the pretraining knowledge becomes outdated Lazaridou et al.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.

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 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 can reliably

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

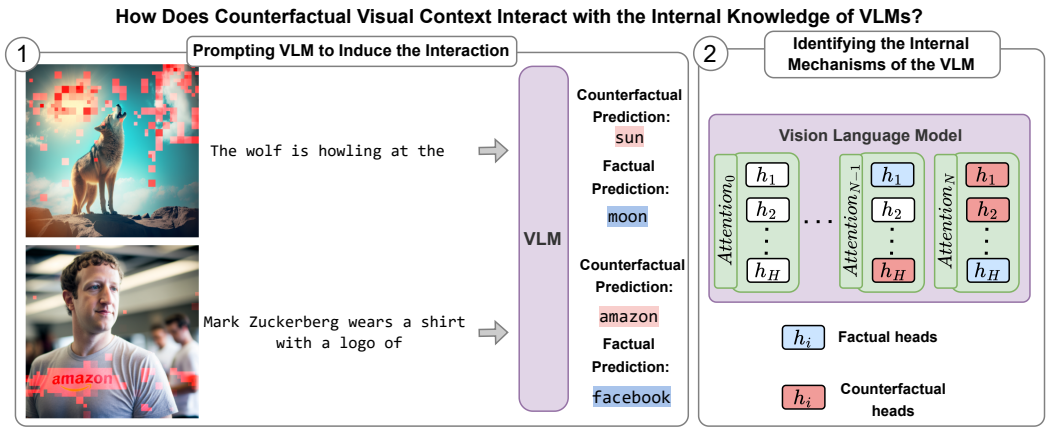

Figure 1: **Overview of Our Approach.** (**Left**) We construct prompts that induce a conflict between a vision-language model's internal factual knowledge and counterfactual visual context. (**Right**) We then analyze which components in the model mediate this tension, identifying attention heads and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

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. 4.1);
2. We identify the attention heads that promote factual and counterfactual responses, ranking their importance with logit attribution (Sec. 4.2);
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. 4.3);
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. 4.4).

2 Related Work

Most prior work on knowledge conflicts has focused on language models and unimodal tasks, leaving the multimodal domain underexplored Xu et al. [2024].

The analyses of knowledge conflicts in language models have largely been behavioral, showing that when resolving conflicts between contextual and internal knowledge, language models can overrely on their internal knowledge or contextual information, depending on factors such as model size Longpre et al. [2021] and conflicting external information Chen et al. [2022]. Wang et al. [2024] found that even SOTA language models often fail to report inconsistencies between in-context information and their internal knowledge. Few works have analyzed the internal mechanisms underlying conflict resolution. Ortu et al. [2024] identified two heads that mediate between factual and counterfactual information, while Jin et al. [2024] showed that pruning specific heads can steer the model's reliance toward internal or contextual sources.

In the multimodal domain, studies on VLMs have primarily focused on benchmark construction and black-box evaluation [Le et al., 2023, Han et al., 2024, Golovanevsky et al., 2025b, Guan et al., 2024]. For example, Han et al. [2024] introduced a dataset probing contextual knowledge conflicts introduced by deceptive visual elements in prompts, while Golovanevsky et al. [2025b] proposed NOTICE, using semantically corrupted image pairs to analyze attention heads behavior in LLaVA and BLIP. Additionally, Le et al. [2023] introduced `COCO-Counterfactuals`, a dataset comprising minimally-edited counterfactual image pairs, and Liu et al. [2024c] developed `ConflictVis` to

evaluate conflicts between visual input and parametric knowledge. However, both studies limited their analyses to evaluating model behavior and prompt structures, without investigating the internal mechanisms by which models resolve such conflicts.

During the final phase of this project (June 2025), we discovered concurrent research by Golovanevsky et al. [2025a], which introduces steering vectors to control model predictions and examines how varying visual input affects the competition between modalities. Their approach uses pairs of images, one consistent with the model's internal knowledge and one modified to introduce a counterfactual variation, focusing on simple object attributes such as color or size. Our approach differs in that we use images depicting complex scenes that contradict common sense, combined with captions specifically designed to generate commonsense responses aligned with the model's internal knowledge, thereby generating a conflict.

3 Background and Methods

3.1 Model Architectures

This study investigates how visual input interacts with the model's internal knowledge during text generation in VLMs. Given a sequence of k image-text tokens, a VLM encodes the image using a vision encoder and the text using an embedding matrix, producing the residual stream $\mathbf{x} \in \mathbb{R}^{d \times k}$, where d is the hidden dimension of the model. The residual stream is processed through L layers, each composed of an attention block $\mathbf{a}^l$ and an MLP block $\mathbf{m}^l$. After the final layer, it is projected to the vocabulary space via an unembedding matrix $W_U \in \mathbb{R}^{d \times |V|}$.

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.

3.2 Dataset Construction

To study how VLMs resolve conflicts between visual context and internal parametric knowledge, we introduce WHOOPS-AHA!, a dataset specifically designed to support mechanistic analysis of multimodal knowledge conflicts. To the best of our knowledge, this is the first dataset explicitly created for conducting mechanistic investigations in this context. Each instance in WHOOPS-AHA! is constructed to provoke a targeted semantic contradiction between these two sources of information

WHOOPS-AHA! builds 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, (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. 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, for the case of 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",..`}. All generated content is manually verified to ensure a clear distinction between factual and counterfactual continuations. Full prompt details are provided in appendix F.

3.3 Analytical Tools

Logit Inspection To identify the internal components of VLMs responsible for the competition between inner knowledge and conflicting visual context, we trace the evolution of token logits across the model's architecture. Specifically, 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} at various

layers and components 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 directly 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 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.

4 Results

4.1 Inducing the Conflict between Inner Knowledge and Visual Context

To systematically induce competition between visual input and internal knowledge, we construct the WHOOPS-AHA! dataset as described in Sec. 3.2. Each example of WHOOPS-AHA! includes a counterfactual image, a sentence describing the image, and two sets of plausible next-word candidates proposed by GPT-4o: S_{fact}, consistent with commonsense knowledge, and S_{cofa} aligned with the counterfactual visual context. We identify t_{fact} as the token in S_{fact} with the highest probability using only the textual part of the prompt. We consider only the first token if a candidate word is tokenized into multiple tokens. Then, using the full multimodal input (image and text), we select t_{cofa} as the token with the highest probability from S_{cofa}. For example, when prompted with the sentence "`The wolf is howling at the`", LLaVA-NeXT and Gemma3 predict the factual token `moon` with probabilities of 78% and 100%, respectively. However, when the corresponding image is included, both models shift to the counterfactual token `sun`, with probabilities of 26% (LLaVA-NeXT) and 44% (Gemma3), while the probability of `moon` drops to 17% and 0.02%. We filter out ambiguous cases in which S_{cofa} contains tokens with a probability higher than S_{fact} in the text-only setup, keeping 436 examples for LLaVA-NeXT and 432 for Gemma3. In the following sections, we always prompt

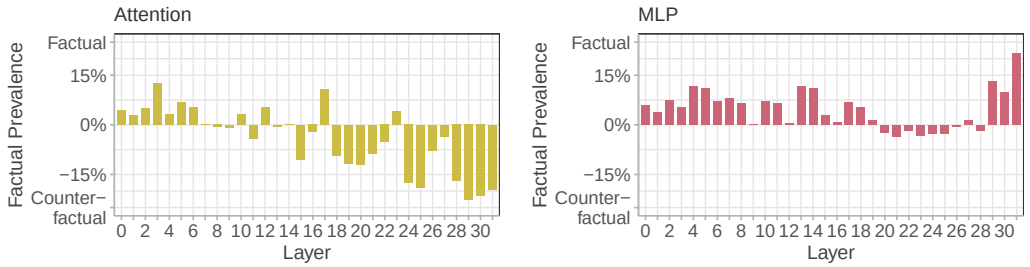

Figure 2: **Factual Prevalence in Attention and MLP Blocks.** The plot shows the factual prevalence of attention and MLP blocks in LLaVA-NeXT across layers, indicating whether each component promotes predictions aligned with factual knowledge or counterfactual visual context. Positive values correspond to blocks favoring the factual (commonsense) continuation. Negative values indicate preference for the counterfactual continuation induced by the image. The results reveal a functional distinction: attention blocks tend to support counterfactual information (**left**), whereas MLP blocks frequently promote the model's internal knowledge (**right**).

the model with image and text using t_{fact} and t_{cofa} to assess whether different model components promote internal knowledge or contextual information. Notably, introducing the image reduces the preference of the model for the commonsense token: the prediction of the factual token t_{fact} drops to 27% for LLaVA-NeXT and 24% for Gemma3. 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.

4.2 The Tension Between Inner Knowledge and Visual Context is Localized

Building on the controlled knowledge conflict induced by WHOOPS-AHA!, we study how the competition between factual and counterfactual continuations is resolved internally and which components mediate it. To do this, we use the Logit Lens technique to analyze the hidden state at the *final token position* of the prompt, after each attention block and MLP, projecting it into the vocabulary space (see Sec. 3.3). We then compute, across the dataset, how often the logit of the factual token t_{fact} is larger than that of the counterfactual token t_{cofa}. This gives an accuracy score for each component that reflects whether it tends to promote the factual or counterfactual mechanism. To measure the strength of this tendency, we compute the factual preference strength, which is defined as the difference between the fraction of examples for which $t_{\text{fact}} > t_{\text{cofa}}$ and 0.5, the random baseline. A value near zero indicates no consistent tendency to favor factual versus contextual information across the dataset, while higher values reflect stronger, more polarized behavior. This method allows us to localize the components that modulate the interaction between visual inputs and internal knowledge.

Functional Separation Between Attention and MLP Layers. We first compare the contributions of attention and MLP blocks to the prediction of t_{fact} and t_{cofa}. Figure 2 shows the results for LLaVA-NeXT (see appendix C for similar results on 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. This pattern is consistent with previous findings on the role of upper-layer MLPs in 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 Appendix, Fig. 8 for Gemma3). The distribution shows that only a small subset of heads exhibit a strong, consistent alignment with t_{fact} or t_{cofa}. Moreover, consistently 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

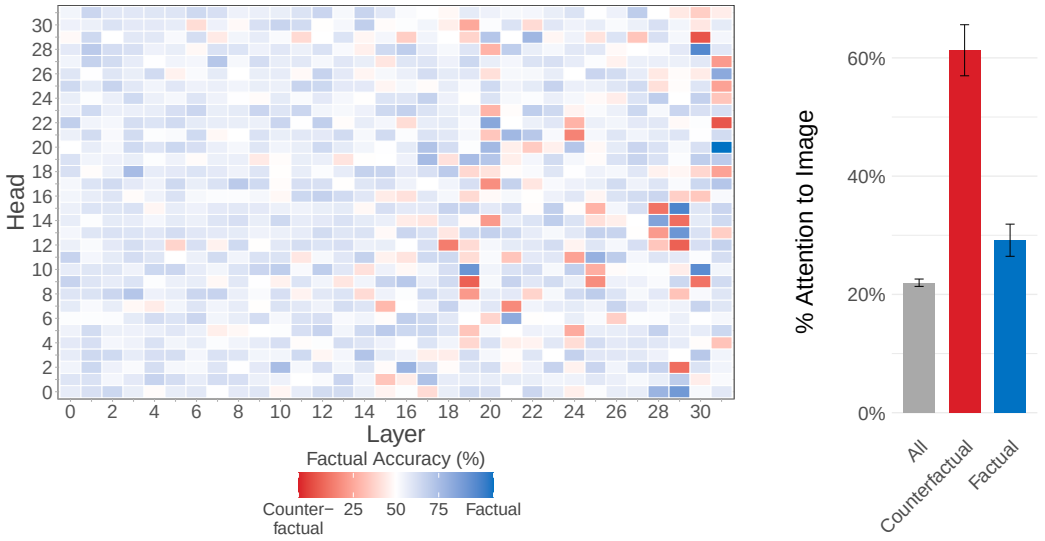

Figure 3: **Contribution of Attention Heads to Factual and Counterfactual Predictions.** (**Left**) Factual accuracy of individual attention heads in LLaVA-NeXT, based on Logit Lens projections at the final token position. Blue indicates heads that tend to favor the factual token (reflecting inner knowledge), 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.

and visual context is resolved late in the forward pass. In the analyses of the next sections, we focus on the 20 heads that promote the factual and counterfactual tokens more strongly. We specifically choose 20 heads because this represents the optimal balance between effectiveness and stability of intervention. Indeed, the incremental improvement in factual accuracy begins to diminish after this point (see Appendix, Fig. 10). 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 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 directlyin late layers of the model, with limited intermediate processing. These findings are consistent for Gemma3, and we report the analysis in appendix C.

4.3 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 predic-

tions. To this end, we apply the targeted intervention strategy introduced in section 3.3, modifying the attention weights to steer the output of the model towards one mechanism or the other.

Guided by our earlier observation that counterfactual heads attend more to visual tokens, we design a bidirectional intervention that selectively adjusts attention values based on head type and token modality. For counterfactual heads, we modify their attention to image tokens; for factual heads, we target their attention to text tokens. In both cases, we apply a multiplicative adjustment at the final token position. Each intervention simultaneously enhances the attention of one group to its relevant modality while suppressing the other group's attention, for instance increasing the attention to image tokens for counterfactual heads while reducing attention to text tokens for factual heads, and vice versa. This approach enables us to modulate the relative influence of factual and counterfactual mechanisms on the model prediction.

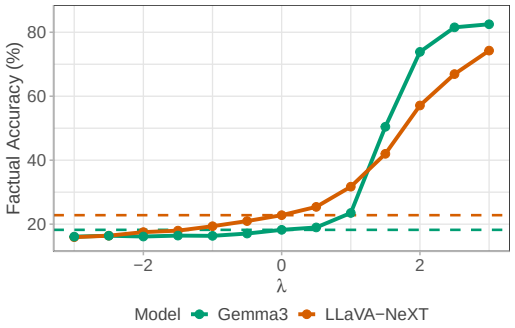

Figure 4: **Intervention on Target Attention Heads.** Change in factual accuracy under different levels of intervention strength (λ). For $\lambda < 0$, we boost the counterfactual heads (on image tokens) and weaken the factual heads (on text tokens); for $\lambda > 0$, we do the opposite. The intervention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

Figure 4 shows the results of our intervention for LLaVA-NeXT (orange profile) and Gemma3 (green profile). For LLaVA-NeXT, the baseline accuracy, defined as the proportion of examples in which the factual token t_{fact} receives a logit higher than the counterfactual token t_{cofa}, is 22%. When we increase attention from factual heads and decrease it from counterfactual heads, 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%.

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 support this choice of intervention range, we also assess its impact on generation quality; the full results, based on KL divergence between generated outputs, are reported in Appendix E.

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, confirming that the observed effects are specific to the heads identified as aligned with factual or counterfactual mechanisms. The complete results for the control experiment are reported in Appendix Fig. 9.

4.4 Counterfactual Predictions Depend on Localized Image Regions

In the previous sections, we analyzed the conflict between contextual information and internal knowledge using WHOOPS-AHA! prompts, which induce a competition between counterfactual visual cues and factual model knowledge. This analysis revealed that specific attention heads at the final token position mediate this conflict, with heads aligned with the visual context attending strongly to image tokens and thereby 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 the upper layers? To address these questions, we perform two complementary analyses: (i) we identify the image patches most responsible for driving counterfactual predictions using attention

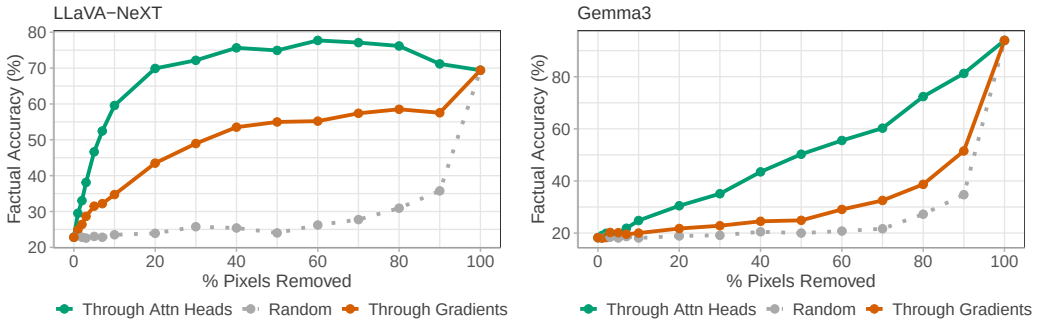

Figure 5: **Ablation of Relevant Pixels.** The plot shows the effect of ablating different percentages of image pixels in LLaVA-NeXT. The green line corresponds to pixels selected based on the highest attention from counterfactual heads, while the orange line corresponds to pixels with the highest gradient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

Figure 6: **Qualitative Examples of Visual Regions Driving Counterfactual Predictions.** Highlighted image regions correspond to visual patches identified as most responsible for counterfactual predictions using attention-based attribution. In both examples, the model generates a visually grounded but factually incorrect token (e.g., rainbow, fruit) instead of the commonsense alternative (black, tissue). The highlighted areas align with semantically meaningful and visually anomalous content, demonstrating that counterfactual outputs are grounded in localized, interpretable image features.

and gradient-based attribution methods, as described in section 3.3; and (ii) we ablate the identified patches by setting the corresponding visual token embeddings to zero at the input of the transformer, and measure the resulting change in inner knowledge accuracy. In addition to the quantitative analysis, we inspect the selected image patches to assess whether they correspond to intuitive counterfactual regions or visually salient objects contradicting the model's internal knowledge. To test the specificity of our findings, we also perform a control experiment in which we randomly sample an equivalent number of image patches for ablation. This allows us to assess whether the identified regions are uniquely responsible for triggering counterfactual predictions or whether any removal of visual input affects the model's behavior.

Quantitative Analysis of Patch Attribution and Ablation The results of the experiments are shown in Figure 5. We observe that the ablation of visual patches identified through attention-based attribution leads to a sharp and consistent increase in factual accuracy as more pixels are removed (green profiles). For instance, in the case of LLaVA-NeXT, factual accuracy improves markedly with the ablation of just 10–30% of the top-ranked patches and eventually plateaus around 80%. This suggests that counterfactual predictions are primarily driven by a small, localized subset of visually salient regions. Gradient-based attribution (shown in red) also yields a substantial increase in factual accuracy, though the effect is less pronounced and saturates earlier, 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, never approaching the improvements achieved through targeted attribution. 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 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-top). 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-bottom). These observations confirm that the model's counterfactual outputs are not arbitrary but grounded in semantically meaningful and localized image features.

5 Discussion

Our work explores how VLMs internally handle conflicts between visual input and internal factual knowledge, building upon earlier studies of similar phenomena. Specifically, we extend the interpretability framework of Ortu et al. [2024], originally developed for textual language models, into the multimodal domain. Recent related studies, such as Liu et al. [2024c] and Han et al. [2024], have constructed diagnostic benchmarks to measure model susceptibility to conflicting visual and textual cues; however, these works primarily focus on evaluating model outputs without deeply analyzing 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 is specifically designed to elicit clear, controlled knowledge conflicts by pairing visually anomalous scenes with commonsense textual prompts, deliberately provoking model-internal tensions rather than broadly diagnosing hallucinations. Thus, while all these datasets address related multimodal conflicts, our approach uniquely emphasizes controlled, interpretable conflict scenarios to facilitate detailed mechanistic analysis. Methodologically, our focus on mechanistic interpretability shares conceptual similarities with recent approaches such as Golovanevsky et al. [2025b]. However, while Golovanevsky et al. [2025b] introduce a novel method to identify attention heads with broad functional roles across tasks in VLMs, they do not examine how models resolve conflicts between visual input and internal knowledge. In contrast, we directly localize this competition to a small set of attention heads and demonstrate their causal role through targeted interventions that shift model predictions. Unlike general hallucination mitigation strategies proposed by Liu et al. [2024a] and Leng et al. [2023], which typically apply training or decoding methods at scale, our method emphasizes precise localization of conflict triggers and inference-time interventions. Although our analysis is not a ready-to-deploy solution, it offers a foundational mechanistic understanding crucial for developing targeted, interpretable interventions in multimodal models.

6 Conclusion

In this work, we investigated how counterfactual visual inputs interact with the internal knowledge representations of VLMs during generation. To this end, we introduced WHOOPS-AHA!, a dataset that pairs visually anomalous scenes with textual prompts designed to elicit either a commonsense (factual) continuation or one grounded in the visual counterfactual. This setup enables fine-grained analysis of how conflicting visual and textual cues influence model behavior. We showed that a small set of attention heads mediate this competition. These heads also exhibit distinct modality preferences and play a causal role in determining the model's output. By intervening on their attention weights, we were able to shift predictions in a controlled way, favoring either the internal knowledge or the visual context. Finally, we demonstrated that these heads provide accurate attribution of the visual regions responsible for counterfactual completions, outperforming standard gradient-based attribution techniques. These findings contribute to a deeper mechanistic understanding of multimodal reasoning in VLMs and offer a foundation for developing more interpretable and controllable systems under conflicting input conditions.

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.

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.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.146. URL https://aclanthology.org/2022.emnlp-main.146/.

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.

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.

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/.

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.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. 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*, pages 7052–7063, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.565. URL https://aclanthology.org/2021.emnlp-main.565/.

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.

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

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.

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

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. Resolving knowledge conflicts in large language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=ptvV5HGTNN.

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.

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 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. 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.

C Experimental Analysis for Gemma-12b

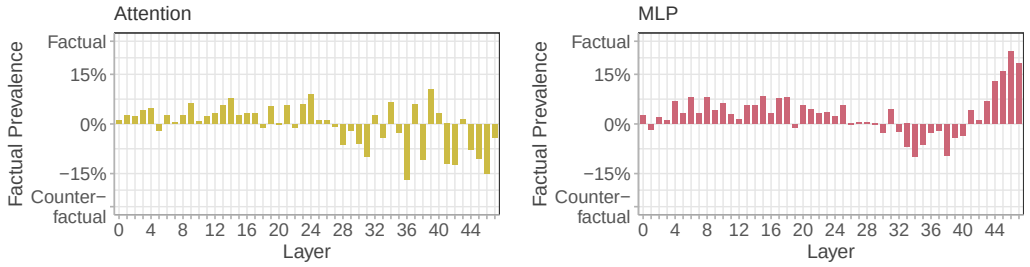

Figure 7: **Factual and Counterfactual Contributions of MLP and Attention Blocks in Gemma3.** Layer-wise deviation from 50% factual accuracy for attention and MLP blocks, as measured by the relative logits of t_{fact} and t_{cofa} via Logit Lens. Positive values indicate a bias toward the factual token, 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.

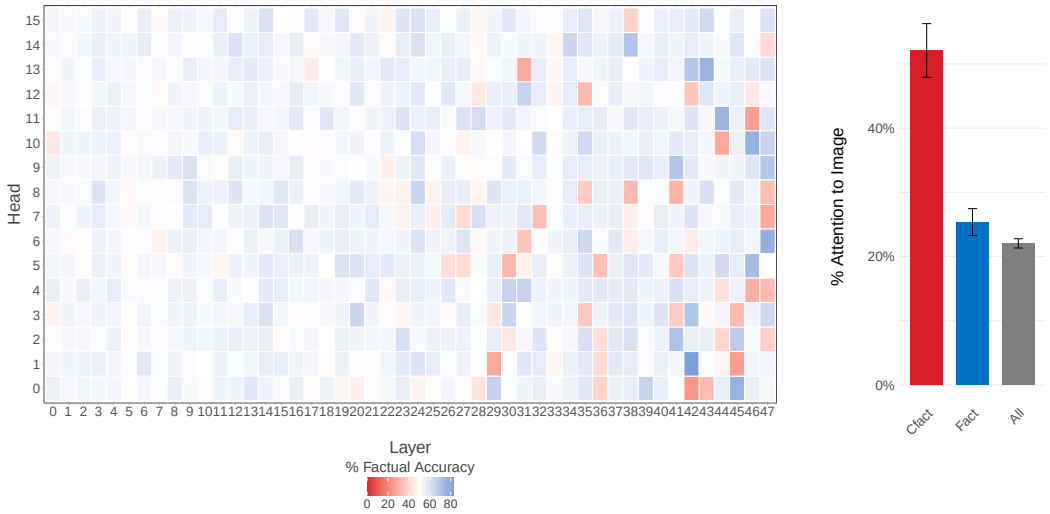

Figure 8: **Factual and Counterfactual Contributions of Attention Heads for Gemma3.** (**Left**) Factual accuracy of individual attention heads in Gemma3, computed using Logit Lens projections of the final token's hidden state. Blue indicates heads that more frequently favor the factual token (t_{fact}), 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

D Additional Results

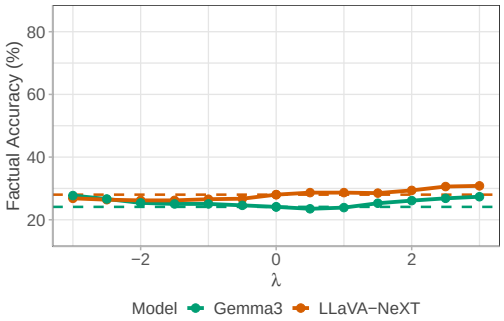

Figure 9: **Control Experiment: Intervention on Random Attention Heads.** Change in factual accuracy under varying levels of intervention strength (λ) applied to 100 randomly selected attention heads. The results show no substantial deviation from baseline, confirming the specificity of the identified target heads.

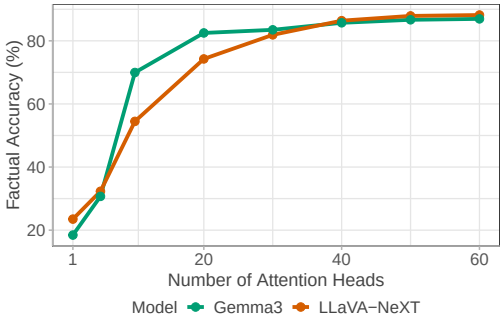

Figure 10: **Effect of Intervening on Varying Numbers of Attention Heads.** Change in factual accuracy as a function of the number of attention heads involved in the intervention. Each value x indicates that x heads are selected from both the factual and counterfactual groups. Intervention strength is 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.

E Impact of Intervention on Text Generation

To quantify the impact of our intervention on text generation quality, we 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 1).

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 11 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.

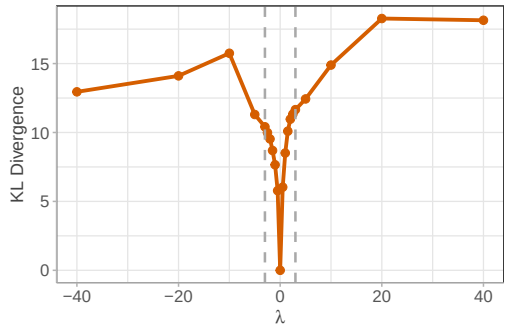

Figure 11: **KL Divergence Between Generated Captions at Different Intervention Strengths in LLaVA-NeXT.** Symmetric increase in KL divergence around $\lambda = 0$, with rapid divergence until $|\lambda| = 3$ and stabilization near $|\lambda| = 10$. Higher intervention magnitudes cause substantial shifts in the generated token distribution, indicating degradation in caption quality.

F 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.