A Mechanistic Analysis of Scene Text in VLM Question Answering

Anonymous ACL submission

Abstract

In-image text can help or hurt VLM predictions, but existing work treats these as separate phenomena without explaining the shared mechanism underneath. To study this mechanism, we introduce GUTIC 1.0, a controlled VQA benchmark that pairs the same image-question with four text overlays: correct answer, grounded misleading, ungrounded misleading, and irrelevant, isolating the effect of in-image text. Attribution analysis reveals that in-image text changes predictions by altering the balance of evidence between text and objects. Helpful text can act as a shortcut, leading the model to answer from the overlay rather than improving visual grounding. Harmful text can introduce competing evidence that outweighs the still-present support for the correct object. This effect is stronger when the text names a misleading object that cannot be grounded in the scene. Attention analysis localizes the image-text competition to a narrow middle-layer window, where the model's prediction aligns with whichever region (text or object) receives greater weight. All data and code will be made publicly available at `http://anonymous.for.review`

1 Introduction

Text is a common component of real-world images, appearing in forms such as street signs, food labels, screenshots and many more. Consequently, a large body of work has focused on evaluating the ability of vision-language models (VLMs) to read and reason over text (Singh et al., 2019; Fu et al., 2025) and on understanding how this ability emerges even in the absence of explicit OCR training (Park et al., 2025). In many cases, however, in-image text can also be completely irrelevant to the task, e.g. when counting the chairs in a room, describing the emotional state of a person, etc. How do VLMs handle in-image text in such cases?

Cognitive science has long shown that words can shape visual task performance in humans even when they are not part of the task itself. In tasks where people must name an image while ignoring a word shown alongside it, phonologically related distractors can facilitate naming (Diaz et al., 2014), echoing broader priming effects (Meyer and Schvaneveldt, 1971). Yet the same influence can become interference when the word competes with the intended response. Semantically related distractor words can increase object-naming errors (La Heij, 1988; Bürki and Madec, 2023), and conflicting word meanings increase errors in color-naming tasks, as in the Stroop effect (Stroop, 1935).

Figure 1: Example of the impact of in-image text on VQA. Depending on what is written in the white sheet of paper, a LLaVA model might give different answers to the question (a) correct answer (b) misleading grounded (c) misleading ungrounded (d) irrelevant text.

Analogous patterns have also been reported in VLMs, where in-image text can either aid or disrupt visual reasoning. Recent work shows that adding text or symbolic cues into images can improve the visual reasoning of VLMs even when the task does not ask about the text itself (Wang et al., 2025b; Yang et al., 2023; Rudman et al., 2025; Rahmanzadehgervi et al., 2025). Other work shows the

opposite pattern, whereby misleading in-image text can confuse VLMs and worsen performance by pulling predictions away from the relevant visual evidence (Qraitem et al., 2025; Westerhoff et al., 2025; Cao et al., 2025). These results suggest that in-image text can change how models solve visual tasks, but they do not explain whether text affects visual grounding or introduces a competing signal.

We address this question by studying helpful and harmful overlays in a controlled visual-question answering (VQA) setting where image and question remain fixed while only the overlaid text changes. Our benchmark, Grounded and Ungrounded Text In-image Conditions (GUTIC), contains 304 VQA examples, each paired with naturally integrated text-overlay variants. Figure 1 shows an example: a purely visual question, "What is on the table?", and four text overlays. Correct-answer text, shown in Figure 1(a), allows us to test whether supporting text strengthens visual grounding or instead provides a shortcut signal. Misleading grounded and ungrounded text, shown in Figure 1(b) and Figure 1(c), allow us to examine whether text behaves differently when it refers to visible versus absent objects. Irrelevant text, shown in Figure 1(d), tests whether the semantic relatedness of text matters.

We use GUTIC to study how evidence shifts between text and objects in object-based VQA, and where this competition is reflected inside VLMs. We identify two mechanisms through which in-image text changes model predictions. Helpful text can induce shortcut behavior, where the model answers from the overlay rather than improving visual grounding. Harmful text, on the other hand, introduces competing evidence that can outweigh the still-present support for the correct object. Notably, misleading text exerts a stronger effect when it is relevant and when it names a plausible but visually absent object, likely because there is no visual counter-evidence to block this option. We further identify a narrow middle-layer window where attention is most strongly associated with whether the model follows the text or the visual object, suggesting that text–object competition is reflected at an intermediate stage of multimodal reasoning.

Our findings suggest that VLMs do not simply read or ignore scene text, depending on the nature of the question/prompt; they actively weigh it against visual evidence, and this arbitration happens at an identifiable intermediate stage of processing. Understanding where and how this competition occurs is a step toward building models that can reliably distinguish task-relevant text from incidental scene content.

2 Related Work

2.1 Visual Text Benchmarks

Benchmarks such as TextVQA (Singh et al., 2019) and OCRBench (Fu et al., 2025) evaluate whether VLMs can read and reason over text embedded in images. These benchmarks typically treat scene text as task-relevant evidence, requiring models to extract textual information from signs, labels, menus, documents, or other written content in order to answer correctly. While they successfully test visual text understanding and OCR-related reasoning, they focus on settings where the task is explicitly about reading or using the text in the image. In contrast, GUTIC studies visually answerable questions where inserted text acts as an additional cue that may help, mislead, or be ignored.

2.2 Interpretability of Visual Text Processing

Interpretability studies of visual-text processing offer mechanistic insights into how VLMs process in-image text. Some studies localize visual-text processing to specific internal components, identifying attention heads associated with OCR behavior (Baek et al., 2025). Others examine how visual-text representations change across layers, suggesting that scene text may initially be processed as visual texture before its semantic content is integrated later (Wang and Ma, 2025), while Anonymous (2026) study how vision encoders develop language-relevant representations and cross-modal semantic alignment across depth. Related work on typographic robustness further shows that sensitivity to image text can also be localized to specific later-layer attention heads in CLIP vision encoders (Hufe et al., 2026). Finally, studies of training dynamics ask when text-reading ability emerges, finding that text readability appears only after substantial training, whereas broader semantic understanding develops earlier (Park et al., 2025).

In our setting, however, the central question is not only whether the model reads the text, but whether that text changes the balance of evidence used to answer. While Anonymous (2026) study reading-comprehension-like representations in standalone vision encoders using rendered-text probing and cross-modal alignment, our work studies how inserted scene text affects downstream VQA behavior in VLMs through competition be-

tween textual and visual evidence during answer generation. Existing work does not study how text embedded inside the image competes with visual object evidence when reading is not required. Our work addresses this gap using controlled VQA examples where only the textual overlay changes.

2.3 Helpful and Harmful In-Image Text

There is also increasing evidence that text in an image can either help or harm VLM performance. Work showing that in-image text can *help* has spanned embedded textual instructions (Wang et al., 2025b), symbolic region marks (Yang et al., 2023), and textual cues to improve counting (Rudman et al., 2025; Rahmanzadehgervi et al., 2025). These studies mainly show that added in-image cues can improve performance, but they do not test how the model uses those cues as evidence. In particular, they do not distinguish whether the cue strengthens grounding in the relevant visual object or provides an alternative signal that the model can rely on.

In contrast, work on *harmful* in-image text studies cases where misleading text interferes with visual reasoning. One line of work shows that VLMs can be misled by inserted text generated by the model itself, suggesting that models may help construct the textual cues that later override their visual judgments (Qraitem et al., 2025). Separate evaluations show that typographic sensitivity appears across both open and commercial models (Cheng et al., 2024). Later work extends these effects to more realistic, scene-coherent settings, including handwritten text in real-world scenes and generated text inserted into plausible image regions (Westerhoff et al., 2025; Cao et al., 2025). The same concern has been explored in safety-critical domains such as autonomous driving (Chung et al., 2024).

These in-image text studies are related to, but distinct from, work on conflicts between separate image and textual inputs (Deng et al., 2025; Pezeshkpour et al., 2025; Zhu et al., 2024). Whereas that line of work studies how VLMs reconcile visual evidence with external textual context, GUTIC setting concerns in-image text rendered within the visual scene itself. However, GUTIC differs from prior work on helpful/harmful in-image text, by treating in-image text as a controlled factor whose relation to the scene can vary. This setup allows us to ask not only whether in-image text changes predictions, but which text is most disruptive and how prediction changes reflect a shift in the balance between textual and visual evidence.

3 GUTIC Benchmark

3.1 Benchmark Overview

GUTIC (Grounded and Ungrounded Text In-image Conditions) is a controlled lightweight benchmark for studying how inserted text affects visual question answering when the task can be solved from visual evidence alone. Each GUTIC sample consists of an image, a multiple-choice question, and a fixed set of answer options. Furthermore, each sample contains four text-overlaid image variants of the original image, generated by inserting each of the four options into a chosen image region, using a generative pipeline (see Section 3.2). The four text-overlaid variant images contain:

- **Correct text**: the inserted text matches the correct answer.
- **Misleading grounded text**: the inserted text names a plausible but incorrect answer corresponding to a visible object.
- **Misleading ungrounded text**: the inserted text names a plausible answer that is absent from the image.
- **Irrelevant text**: the inserted text is unrelated to both the image and the question.

Crucially, this taxonomy disentangles semantic plausibility from visual grounding. In other words, we can compare cases where the text points to a wrong but visible object (misleading grounded) with cases where it points to a plausible answer that is not actually in the image (misleading ungrounded). We can also compare these with cases where the overlaid text is the correct answer itself (correct text), or an irrelevant answer (irrelevant text). In the "misleading grounded" variant, we refer to the object named by the in-image text as the "misleading object". A visual summary of these variants is provided in Appendix Figure 9.

3.2 Dataset Construction

Our dataset builds on GQA (Hudson and Manning, 2019), which provides scene-graph annotations and object-level bounding boxes. We add a text-region bounding box for each example, while using GQA annotations for the target and misleading grounded objects. Each sample therefore contains three bounding boxes, one introduced by our pipeline. The same text region is shared across all

variants of a sample. The final benchmark contains 304 evaluation samples.

Candidate answers are generated by GPT-5.2 based on the image, question, and scene-graph objects, and are subsequently verified through manual review to ensure semantic correctness. Similarly, in-image text variants are rendered using a scene-aware text overlay pipeline inspired by (Cao et al., 2025), which dynamically selects the optimal location to overlay the text into the scene. The prompts used for these generations are provided in Appendix H. The misleading grounded option is selected from objects present in the image, ensuring that it corresponds to a visually grounded alternative, while the ungrounded option is generated to be plausible for the question but absent from both the image and the scene-graph object list. For example, in a kitchen scene, a "plate" is visually plausible, but may not appear in the image. Additional details on filtering, validation, and the full construction pipeline are provided in Appendix A.

4 Methodology

We use two complementary region-level analyses to study how inserted text affects VQA decisions. Integrated Gradients measures whether image regions support or oppose an answer, while layer-wise attention analysis measures how highly those regions are attended to during answer generation.

4.1 Attribution Analysis

For our attribution analysis, we ask how strongly each image region supports or opposes an answer. A suitable tool for this is Integrated Gradients (Sundararajan et al., 2017), which attributes a target output to input components relative to a chosen baseline. We use a blank gray image as the baseline and compute attributions over visual tokens.

In the core experiments of Section 5, we set the Integrated Gradients target to the correct answer token for all samples and text variants to ensure that the same target token is used before and after adding text, allowing direct comparison across textual variants even when the model's prediction changes. This allows us to ask how strongly each region supports or opposes the correct answer, regardless of the model's final prediction. We also conduct an additional analysis in Appendix Section F, using the model's generated answer token as the Integrated Gradients target output.

To compare the contribution of different parts of the image, we compute a separate attribution score for the text region, the correct object region, and the misleading object region. These regions are initially localized using the GUTIC bounding-box annotations. However, attribution is computed over visual tokens. Thus, since many household objects are irregularly shaped, the visual tokens associated with an object may in some cases extend beyond its rectangular bounding box, or in other cases fall partially within it (see Figure 10). We therefore algorithmically compute targeted pixel regions using the top-k attributed visual features as seeds (see Algorithm 1). The full procedure is described in Appendix B.

To summarize attribution at the region level, we compute the net evidence of a region R as:

$$\mathrm{NetEvidence}(R) = \frac{1}{|T_R|} \sum_{i \in T_R} a_i, \qquad (1)$$

where a_i is the attribution score of token i, and T_R is the set of tokens associated with region R. Positive values indicate support for the target answer, while negative values indicate opposition.

4.2 Attention Analysis

Integrated Gradients measures regional evidence, but it does not show how the model allocates attention across regions during generation. We therefore analyze layer-wise attention weights from the first generated answer token to image tokens.

For each region R, we compute the attention ratio for a given region, by taking the average attention assigned to tokens in that region, normalized by the average attention over all image tokens:

$$D^{(\ell)}(R \mid S) = \frac{\frac{1}{|T_R|} \sum_{i \in T_R} A_i^{(\ell)}}{\frac{1}{|S|} \sum_{i \in S} A_i^{(\ell)}}, \qquad (2)$$

where T_R is the set of tokens in region R, S is the set of all image tokens, and $A_i^{(\ell)}$ is the attention weight assigned to token i at layer ℓ.

This ratio measures the relative prominence of each region during answer generation. Values above 1 indicate that a region receives more attention per token than the average image token, while values below 1 indicate less attention. Tracking this quantity across layers allows us to identify where text and object regions become most prominent inside the model.

5 Experimental Results & Analysis

We report results for two representative vision–language models, LLaVA-NeXT-7B (Liu et al., 2024) and Qwen-VL-7B (Bai et al., 2023).

All experiments are conducted in a multiple-choice setting. As described in Section 3.1, each input consists of an image, question, and four answer candidates corresponding to the four in-image text variants. For each sample, we compare the original no-text image against each variant.

5.1 Accuracy Gaps Across Text Variants

We observe statistically significant accuracy gaps in how models respond to different in-image text variants. We measure these gaps relative to the original no-text image and test significance using a two-sided McNemar test on paired predictions from the same samples.

Figure 2 shows the accuracy gaps for LLaVA-NeXT and Qwen-VL.

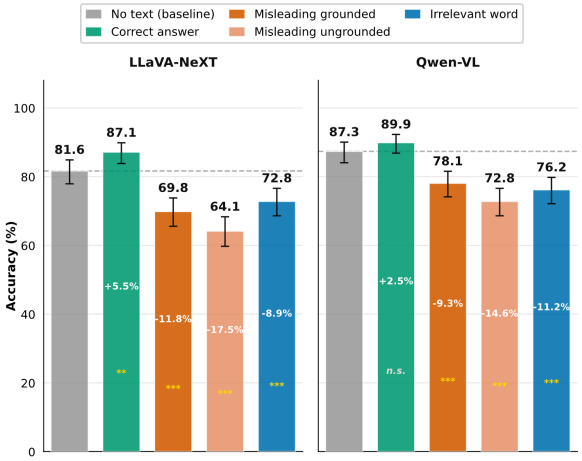


Figure 2: Accuracy under different in-image text variants for LLaVA-NeXT and Qwen-VL. Numbers inside the bars indicate the accuracy change relative to the no-text baseline. Error bars denote 95% confidence intervals. Significance is vs. no-text baseline (two-sided McNemar): *p<.05, **p<.01, *** p<.001

.

In both models, misleading text causes a larger drop when it names an ungrounded (absent) object than when it names a grounded (visible) object, with gaps of -17.5% versus -11.8% for LLaVA-NeXT and -14.6% versus -9.3% for Qwen-VL. Irrelevant text also reduces accuracy, showing that unrelated in-image text is not fully ignored. Correct text improves accuracy for LLaVA-NeXT (+5.5%) and produces a smaller, non-significant gain for Qwen-VL (+2.5%). To test whether these trends extend beyond the two main models, we also evaluate LLaVA-7B (Liu et al., 2023) and InternVL-8B (Wang et al., 2025a). These additional results are provided in Appendix C.

Our experimental observations corroborate prior work showing that in-image text can improve multimodal reasoning through visual aids (Yang et al., 2023; Rudman et al., 2025) and degrade it through typographic interference (Qraitem et al., 2025; Westerhoff et al., 2025). However, accuracy gaps alone do not explain what changes inside the model, including whether helpful and harmful text follow different internal patterns or how grounded and ungrounded misleading variants are processed differently. We address these questions next using attribution and attention analyses.

5.2 Qualitative Attribution Analysis

To visualize how inserted text changes the model's evidence, we apply the attribution analysis from Section 4.1 to individual examples. We first consider helpful text, where a correct-answer overlay moves the model toward the correct answer, then harmful text, where misleading overlays pull it away.

Figure 3 shows one case contributing to the +5.5% accuracy gap from correct-answer overlays, discussed in Section 5.1. Attribution maps are computed with respect to the correct answer token, and blue regions indicate positive support for the correct answer, while red regions indicate opposition.

Q: What is on the bed to the left of the bottle?

(a) No-text
: "Clock"

(b) Correct Answer
: "Book"

Figure 3: In (a), the correct object is the book on the bed (green bounding box), while the misleading object is the clock (orange bounding box). In (b), the helpful text overlay is the *word* "book" (blue dashed bounding box) on the wall. Moving from (a) to (b), the inclusion of the helpful text changes the model's prediction from wrong (i.e., "clock") to correct (i.e., "book"), reflecting a shortcut behavior.

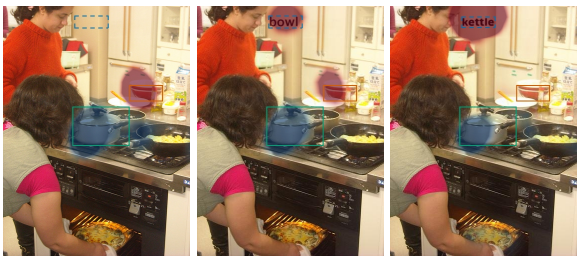
The attribution map in Figure 3(a) reveals that, for the question "What is on the bed to the left of

the bottle?", the model's incorrect answer "clock" arises from competition between the book object and the clock object. Blue attribution over the book supports the correct answer "book", while red attribution over the clock opposes that target and aligns with the model's incorrect prediction, "clock". Once the word "book" is overlaid on the wall in (b), the model changes its prediction to "book". The inserted word becomes the strongest source of positive attribution, while the clock continues to provide opposing evidence.

This suggests a shortcut behavior, where the model shifts from the competing visual object toward the overlaid textual cue. The improvement therefore does not appear to come from stronger grounding in the actual book region itself. Instead, the overlaid word provides a direct answer cue that outweighs the competing support from the clock.

Conversely, Figure 4 shows two harmful cases contributing to the -11.8% and -17.5% drops in grounded and ungrounded variants respectively.

Q: What is on the oven?



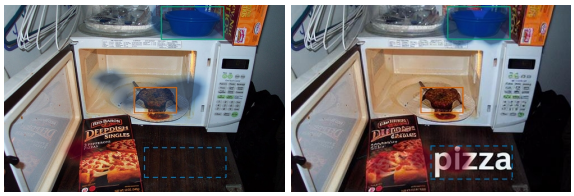
(a) No Text (b) Misleading Grounded (c) Misleading Ungrounded

: "Pot" : "Bowl" : "Kettle"

Figure 4: In (a), the model correctly predicts "pot" from the object on the oven (green bounding box). In (b), the harmful grounded overlay "bowl" (blue dashed bounding box) changes the prediction away from the correct answer "pot", and causes the text region to have an increased attribution. In (c), the harmful ungrounded overlay "kettle" (blue dashed bounding box) changes the prediction toward a plausible but *absent* object, as there is no kettle in this kitchen scene.

Figure 4 shows an example whereby, without added text, blue attribution over the pot indicates strong support for the correct answer "pot". Although some red attribution appears on the bowl, it does not dominate the decision. Once misleading text is introduced, however, the oppositional signal becomes strong enough to redirect the model's prediction. In (b), this oppositional signal is split between the grounded overlay "bowl" and the visible bowl itself (red attribution regions). This indicates that the change in prediction is caused because the text reinforces an existing visual competitor. In (c), however, the word "kettle" alone is enough to redirect the model's prediction, despite there being no kettle in the scene, producing a large red attribution region centered on the overlaid word.

Q: What is located on top of the microwave on the table?



(a) No-text (b) Misleading Grounded

: "Bowl" : "Bowl"

Figure 5: In (a), the model correctly predicts "bowl". In (b), the grounded overlay "pizza" introduces competing evidence associated with the visible pizza, but the bowl on top of the microwave remains the dominant source of support. Thus, the model's prediction remains correct.

In both cases, the prediction changes not because evidence for the pot disappears, but because overlaid text introduces stronger competing evidence. A complementary case where misleading text introduces competing evidence but the prediction remains correct appears in Figure 5 in Appendix D.

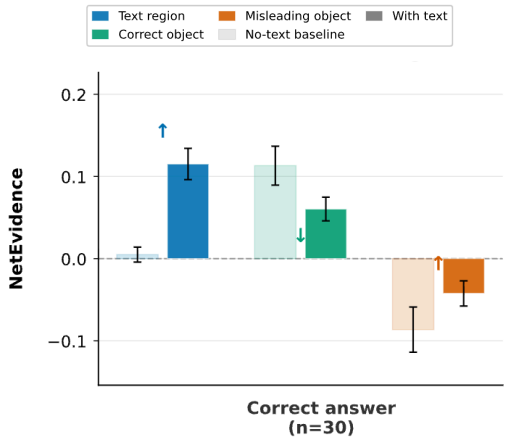
5.3 Quantitative Attribution Analysis

The qualitative examples in Section 5.2 suggest that inserted text changes predictions by shifting the balance between textual and object-based evidence. We now test whether this pattern holds systematically across the dataset.

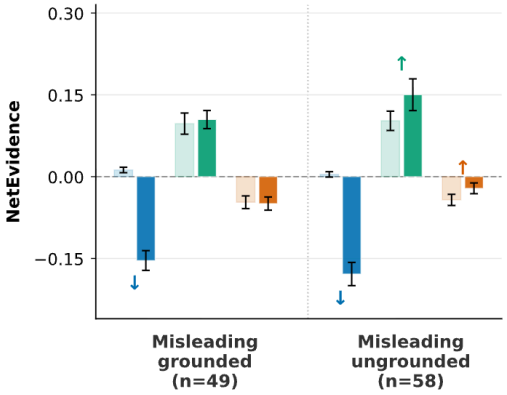
When the prediction changes (helpful and harmful text overlays). Figure 6 shows region-level attribution for LLaVA-NeXT on examples where the added text changes the prediction.

Each subfigure reports three regions: the text region, the correct object, and the misleading object. Each bar reports the *NetEvidence* score for a region, with positive values supporting the correct answer and negative values opposing it. Because the same examples are used in both the *no-text baseline* and text variants, the comparison directly reflects how adding text changes the contribution of each region.

Panel (a) shows *helpful* cases where adding the correct answer as text changes the prediction from incorrect to correct. In these examples, the NetEvi-



(a) Correct-answer text: Wrong → Correct



(b) Misleading text: Correct → Wrong

Figure 6: Region-level NetEvidence for LLaVA-NeXT when the text overlay changes the model's answer. Lighter bars show the no-text baseline while darker solid bars show the text-overlaid condition, with arrows marking the direction of change. The largest shifts occur in the text-region NetEvidence.

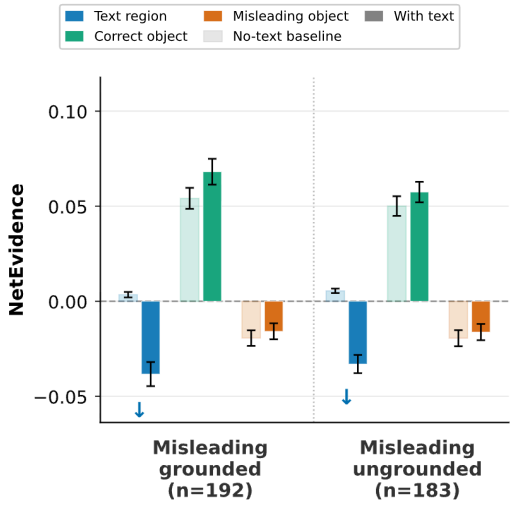
dence attribution of the text region shifts from near zero to strongly positive values, becoming the dominant source of support for the correct answer. The correct object, however, shows little or no increase relative to the baseline, indicating that the improvement comes from the added textual cue rather than stronger grounding in the visual object.

Panel (b) shows *harmful* cases where misleading text flips the prediction from correct to wrong. The correct object remains positively aligned with the correct answer and the corresponding NetEvidence changes only slightly after text is added. By contrast, NetEvidence of the text region shifts from near zero to strongly negative values, indicating that the added text opposes the correct answer.

Within the harmful cases, the grounded and ungrounded variants differ in where the competing signal comes from. Grounded misleading text names an object that is present in the image, so the overlay reinforces a visually grounded alternative that already has some support in the no-text baseline. Ungrounded misleading text, by contrast, names a plausible answer without a corresponding object in the image, so the competing signal arises primarily from the text region itself. In addition, the magnitude of the text-region attribution is consistently higher for ungrounded misleading text than for grounded misleading text, indicating that ungrounded text introduces a stronger competing signal. This mirrors the qualitative examples in which grounded misleading text amplifies an existing visually supported alternative, whereas ungrounded misleading text introduces a competing signal that is less tied to the visual scene.

Across both directions of change, the main shift comes from the text region, while object-level attributions remain comparatively stable.

When the prediction remains unchanged. Figure 7 shows cases where misleading text is added but the prediction remains correct.



(a) Misleading text: Correct Prediction Unchanged

Figure 7: Region-level NetEvidence for LLaVA-NeXT when misleading text is added but the prediction remains correct. The text region still introduces opposing evidence, but the correct object remains the dominant positive source of support.

Across all variants, the correct object remains positively aligned with the correct answer, while the text region changes its contribution depending on the overlaid word. This shows that text can still introduce competing evidence even when it does not change the final prediction. In these cases, however, the contribution of the text-region is not

large enough to outweigh that of the correct object.

The overall pattern in Figures 6 and 7 supports two conclusions. First, improvements from correct-answer text arise primarily from a strong textual cue, rather than increased grounding in the visual object. Second, ungrounded misleading text can exert a stronger influence than grounded misleading text because its competing signal is concentrated in the text region rather than distributed across a visible object and the text overlay.

To demonstrate how these quantitative behaviors generalize, we discuss additional results for Qwen-VL in Appendix E.

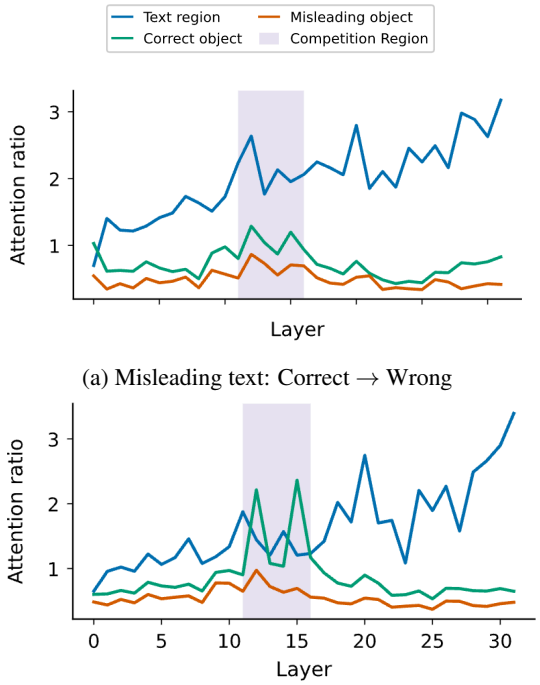
5.4 Attention Analysis Results

Having shown that inserted text shifts region-level evidence, we use the attention ratio from Equation 2 to ask where this competition appears during answer generation. This ratio compares each region's average attention to the average over all image tokens, allowing us to track text and object prominence across layers.

Figure 8 shows this ratio for the text, correct-object, and misleading-object regions in LLaVA-NeXT under the *misleading grounded* text variant.

The clearest separation appears within a middle-layer window around layers 10–15. In Figure 8(a), where the misleading text is associated with a prediction change, the text region receives higher attention than the correct-object region throughout much of this window. In contrast, in Figure 8(b), where the prediction remains correct, attention to the correct object exceeds attention to the text region. This suggests that prediction changes are associated with greater attention to the text region, whereas unchanged correct predictions are associated with greater attention to the correct object. This pattern suggests that the model's final prediction aligns with the region that dominates attention in these intermediate layers. The pattern is also repeatedly observed for several other text overlay variants, as shown in Appendix Figure 14.

Outside this middle-layer window, text attention tends to dominate, making the middle layers the main stage where correct-object attention overtakes the text region. A similar middle-layer separation, as shown in Appendix Figure 15, appears in Qwen-VL around layer 19.



(a) Misleading text: Correct → Wrong

(b) Misleading text: Correct Prediction Unchanged

Figure 8: Layer-wise attention to image regions, shown as a ratio to the average attention over all image tokens. The key difference between (a), where misleading text changes the model's prediction, and (b), where the prediction remains correct, is which curve dominates the shaded middle-layer window. Text attention is higher in this window in (a), while the correct-object curve is higher in this window in (b).

6 Conclusion

This work is a step toward a broader understanding of how VLMs arbitrate between conflicting textual and visual evidence in multimodal scenes. We study how different in-image text overlays affect visually answerable VQA tasks. By keeping the image and question fixed while varying only the inserted text, we analyze how scene text influences model predictions and examine these effects through mechanistic analyses. Using attribution analysis, we find that prediction changes are associated with shifts in the balance between textual and visual evidence, where text can either reinforce or compete with object-based evidence. Using attention analysis, we further show that this competition is reflected in a narrow intermediate-layer window where the model's prediction is associated with whether attention is allocated more to text or object regions. Overall, our results show that scene text acts as an active source of evidence competing with visual objects during question answering.

Limitations

The size of the GUTIC 1.0 dataset is a work in progress and we plan to release a much bigger version 2.0. While the controlled setup enables precise comparisons between textual conditions, we plan to expand the dataset significantly with more samples to increase the diversity of scenes and release the dataset to community for more encouragement to study the observed patterns.

References

Anonymous. 2026. Reading between the layers: Cross-modal alignment and representational hierarchy in vision transformers. *Under review*.

Ingeol Baek, Hwan Chang, Sunghyun Ryu, and Hwanhee Lee. 2025. How do large vision-language models see text in image? unveiling the distinctive role of ocr heads. *Preprint*, arXiv:2505.15865.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Audrey Bürki and Sylvain Madec. 2023. Picture-word interference in language production studies: Exploring the roles of attention and processing times. *Preprint / Conference Paper*.

Yue Cao, Yun Xing, Jie Zhang, Di Lin, Tianwei Zhang, Ivor Tsang, Yang Liu, and Qing Guo. 2025. Scenetap: Scene-coherent typographic adversarial planner against vision-language models in real-world environments. *Preprint*, arXiv:2412.00114.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. 2023. Textdiffuser-2: Unleashing the power of language models for text rendering. *Preprint*, arXiv:2311.16465.

Hao Cheng, Erjia Xiao, Jindong Gu, Le Yang, Jinhao Duan, Jize Zhang, Jiahang Cao, Kaidi Xu, and Renjing Xu. 2024. Unveiling typographic deceptions: Insights of the typographic vulnerability in large vision-language model. *Preprint*, arXiv:2402.19150.

Nhat Chung, Sensen Gao, Tuan-Anh Vu, Jie Zhang, Aishan Liu, Yun Lin, Jin Song Dong, and Qing Guo. 2024. Towards transferable attacks against vision-llms in autonomous driving with typography. *Preprint*, arXiv:2405.14169.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. 2025. Words or vision: Do vision-language models have blind faith in text? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3867–3876.

Michele T Diaz, Larson J Hogstrom, Jie Zhuang, James T Voyvodic, Micah A Johnson, and C Christine Camblin. 2014. Written distractor words influence brain activity during overt picture naming. *Frontiers in Human Neuroscience*, 8:167.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, and 5 others. 2025. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *Preprint*, arXiv:2501.00321.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Preprint*, arXiv:1902.09506.

Lorenz Hufe, Constantin Venhoff, Erblina Purelku, Maximilian Dreyer, Sebastian Lapuschkin, and Wojciech Samek. 2026. Dyslexify: A mechanistic defense against typographic attacks in clip. *Preprint*, arXiv:2508.20570.

Wido La Heij. 1988. Components of stroop-like interference in picture naming. *Memory & Cognition*, 16(5):400–410.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

David E. Meyer and Roger W. Schvaneveldt. 1971. Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90(2):227–234.

Jaeyoo Park, Sanghyuk Chun, Wonjae Kim, Sangdoo Yun, and Bohyung Han. 2025. Emergence of text readability in vision language models. *Preprint*, arXiv:2506.19389.

Pouya Pezeshkpour, Moin Aminnaseri, and Estevam Hruschka. 2025. Mixed signals: Decoding vlms' reasoning and underlying bias in vision-language conflict. *arXiv preprint arXiv:2504.08974*.

Maan Qraitem, Nazia Tasnim, Piotr Teterwak, Kate Saenko, and Bryan A. Plummer. 2025. Vision-llms can fool themselves with self-generated typographic attacks. *Preprint*, arXiv:2402.00626.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. 2025. Vision language models are blind: Failing to translate detailed visual features into words. *Preprint*, arXiv:2407.06581.

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

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. *Preprint*, arXiv:1904.08920.

John Ridley Stroop. 1935. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18(6):643–662.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. *Preprint*, arXiv:1703.01365.

Hanzhang Wang and Qingyuan Ma. 2025. Textural or textual: How vision-language models read text in images. In *Forty-second International Conference on Machine Learning*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025a. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Zhaochen Wang, Yiwei Wang, and Yujun Cai. 2025b. Cure or poison? embedding instructions visually alters hallucination in vision-language models. *arXiv preprint arXiv:2508.01678*.

Justus Westerhoff, Erblina Purelku, Jakob Hackstein, Jonas Loos, Leo Pinetzki, Erik Rodner, and Lorenz Hufe. 2025. Scam: A real-world typographic robustness evaluation for multimodal foundation models. *Preprint*, arXiv:2504.04893.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. 2021. Segformer: Simple and efficient design for semantic segmentation with transformers. *Preprint*, arXiv:2105.15203.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *Preprint*, arXiv:2310.11441.

Tinghui Zhu, Qin Liu, Fei Wang, Zhengzhong Tu, and Muhao Chen. 2024. Unraveling cross-modality knowledge conflicts in large vision-language models. *arXiv preprint arXiv:2410.03659*.

A Dataset Construction Details

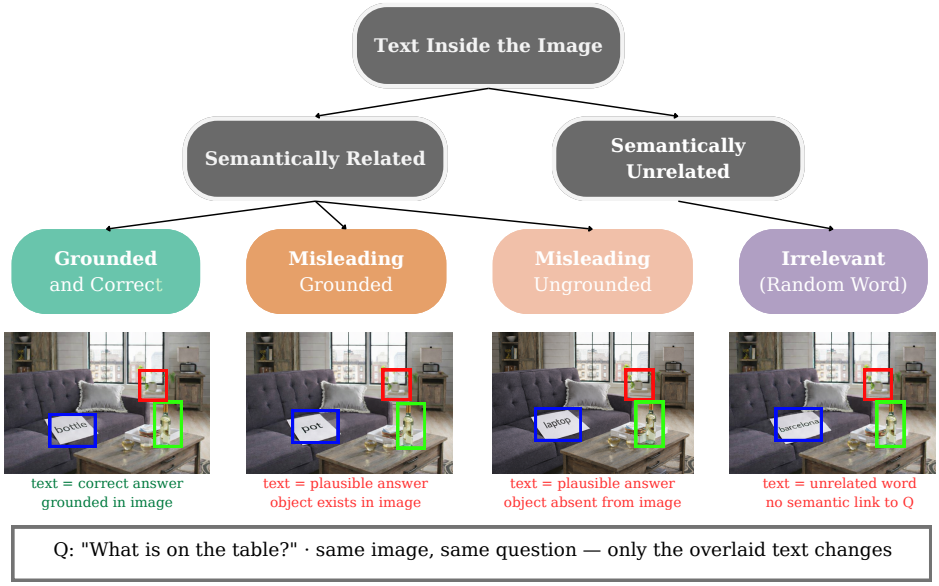


Figure 9: Overview of the four GUTIC overlay variants. The taxonomy first separates text that is related to the question from text that is unrelated, then distinguishes whether the related text names the correct object, a visible but incorrect object (i.e., a misleading grounded object), or a plausible but absent object (i.e., a misleading ungrounded object). The four image variants show how this taxonomy is instantiated using the same image-question pair, where only the word in the blue text region changes. This controlled design isolates the model's response to overlaid text that supports, competes with, or is unrelated to the visual evidence.

GUTIC is constructed from image–question pairs derived from the GQA dataset (Hudson and Manning, 2019). We retain samples where the question can be answered from visual content alone and where the correct answer corresponds to a visible object in the scene graph. We then filter out three types of samples: those whose answers cannot be grounded in the scene graph, those whose questions require reading existing image text, and those whose images lack a suitable writable region for text overlay. Writable regions are identified using SegFormer (Xie et al., 2021), followed by geometric filtering based on area, width, height, and aspect ratio to ensure legible and visually plausible text overlay. After filtering, 1,593 candidate samples remain.

For each retained sample, we generate the four textual variants described in Section 3.1 and illustrated in Figure 9. Each sample therefore consists of one no-text image together with four generated text-overlay variants, where only the inserted text differs. The selected text region is fixed across all variants, ensuring that comparisons isolate the effect of text content rather than text location.

Multiple-choice options and text overlaying positions are generated by GPT-5.2 based on the image, question, and scene-graph objects. The prompt used for this generation is provided in Appendix H. For the misleading grounded variant, the corresponding object is selected from the GQA scene graph, ensuring that it has a valid object annotation. The ungrounded option is generated to remain plausible with respect to the question while being absent from both the image and the scene-graph object list, and the irrelevant condition is constructed using a random unrelated word.

The generated text overlays are then rendered using a diffusion-based method (Chen et al., 2023), ensuring natural integration into the scene. The generated variants are verified through manual review. Samples are again rejected if the inserted text is not readable, does not align with the selected region, does not match the intended textual variant, or if the rendered image quality is degraded by the diffusion process. After manual validation, 474 high-quality examples remain.

From this set, we further select 304 examples for evaluation by removing cases where the bounding boxes corresponding to the correct object, misleading grounded object, and text region overlap, as

such overlaps can confound the mechanistic analyses in Section 5.3 and Section 5.4. The final dataset contains 304 evaluation samples.

B Attribution Analysis Post-Processing

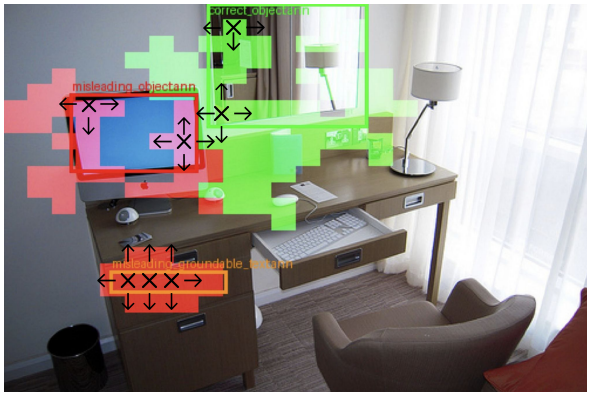


Figure 10: Example of attribution-guided region refinement. Bounding boxes define the initial regions, while the highlighted areas show the refined attribution regions obtained after expanding high-magnitude attribution tokens with consistent sign.

Algorithm 1 Attribution-Guided Region Extraction

Require: Attribution grid G, annotated box B, top-k
Ensure: Refined region R
1: Project B onto the attribution grid
2: Compute the dominant attribution sign s within B
3: Construct score map S by keeping only attribution values with sign s
4: Select the top-k highest-scoring tokens within B as seeds
5: Set threshold τ as the 90th percentile of score differences between neighboring tokens in S
6: Grow connected regions from the seeds, keeping only neighboring tokens with sign s, satisfying $|S(p) - S(n)| \leq \tau$, and $S(n) \geq \alpha \cdot S_{\text{seed}}$
7: Merge overlapping components and remove small noisy regions
8: **return** R

Attribution scores are computed at the visual-token level, while the annotated regions in the dataset are defined using rectangular bounding boxes. However, many objects in real-world scenes have irregular shapes. As a result, important attribution tokens may lie outside the annotated bounding box, while unrelated tokens may fall inside it. Directly aggregating attribution values over all tokens within the bounding box in Equation 1 can therefore introduce bias by excluding important evidence or including irrelevant regions. This can lead to unfair comparisons across regions, particularly when some objects align more closely with rectangular boundaries than others.

To reduce this sensitivity to bounding-box borders, we use the annotated bounding boxes only to identify the initial attribution seeds that define the regions used in the aggregation. Figure 10 illustrates an example of the resulting attribution regions for the correct object, the misleading grounded object, and the text region.

The process begins by identifying high-importance attribution tokens across the entire image. Specifically, we select tokens among the top-k attribution magnitudes that also lie within one of the annotated bounding boxes, and use them as seed tokens, illustrated by the "X" symbols in Figure 10. The bounding boxes are therefore used only to identify the initial seed tokens, rather than to directly define the regions used for aggregation.

Starting from the seed tokens, we iteratively expand to neighboring tokens in order to recover the full connected attribution region, including important tokens that may extend slightly beyond the original bounding-box boundaries. A neighboring token is included only if its attribution magnitude is sufficiently similar to the current region, such that the difference between neighboring attribution values remains below a threshold and the magnitude does not decrease substantially relative to the seed token. This expansion process allows the refined region to follow the actual attribution structure rather than the rectangular annotation boundaries, while avoiding low-magnitude or inconsistent tokens.

The final region is obtained by merging overlapping connected components and removing small noisy regions. The complete procedure is summarized in Algorithm 1. The region-level score is then computed as the average attribution over the resulting set of tokens.

C Preliminary Experiment Details

This section provides the full results for the experiment described in Section 5.1. For each image–question pair, we evaluate the no-text baseline and the four text variants described in Section 3.1, using the same image, question, and answer options across all variants.

We report results for four vision–language models: InternVL, LLaVA, LLaVA-NeXT, and Qwen-VL. The main paper focuses on LLaVA-NeXT and Qwen-VL, while the remaining models are included here for completeness.

Figure 11 shows consistent accuracy gaps across all evaluated models. Correct text improves ac-

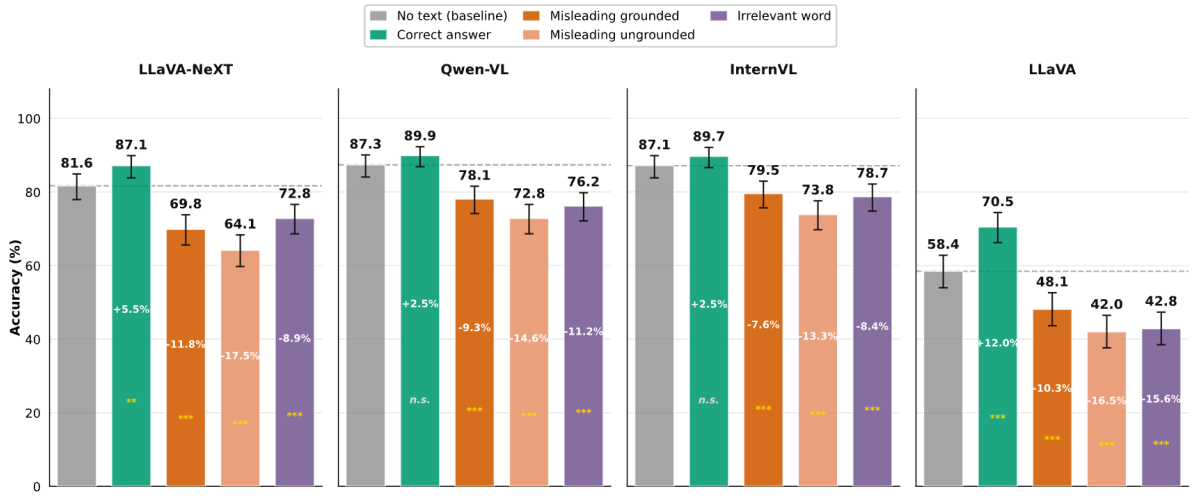


Figure 11: Full accuracy results across all evaluated models and in-image text variants. The numbers inside the bars show the accuracy change relative to the no-text baseline, while the asterisk markers indicate statistical significance under a two-sided McNemar test. Error bars show 95% confidence intervals.

curacy for LLaVA and LLaVA-NeXT, while the gains for InternVL and Qwen-VL are smaller and not statistically significant. In contrast, misleading and irrelevant text reduce accuracy across all models. We also observe a consistent difference between the misleading grounded and misleading ungrounded conditions, where misleading ungrounded text causes a larger accuracy drop in every model. These drops are statistically significant across all evaluated models.

Accuracy is computed for each variant and compared to the no-text baseline. Statistical significance is assessed using a two-sided McNemar test on paired predictions, comparing each condition to the baseline. We report significance levels as: * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$).

D Additional Qualitative Examples

This section provides an additional qualitative example complementing the attribution analysis in Section 5.2. As in the main paper, all attribution maps are computed with respect to the correct answer token. Blue regions indicate positive support for the correct answer, while red regions indicate opposition.

Figure 5 shows a complementary case in which misleading text introduces competing evidence but does not change the model's final prediction.

In the no-text baseline shown in Figure 5(a), the model correctly predicts "bowl" from the object on top of the microwave (green bounding box). Once the grounded overlay "pizza" is introduced in Figure 5(b), the text region becomes associated with additional competing evidence aligned with the visible pizza object. However, unlike the harmful cases in Figure 4, the support for the correct object remains dominant, and the prediction therefore stays correct.

This example shows that misleading text can introduce competing evidence even when it is not strong enough to overturn the model's decision. The prediction remains unchanged because the positive attribution associated with the correct object continues to outweigh the competing evidence introduced by the text overlay and the misleading grounded object.

E Qwen Attribution Analysis

Figure 12 shows the same region-level attribution analysis as Section 5.3, but for Qwen-VL. As in the main paper, NetEvidence is computed using the *correct answer* token as the Integrated Gradients (IG) target.

The same overall trends observed for LLaVA-NeXT also appear in Qwen-VL. In cases where text changes the model's prediction, which are shown in (a) and (b), the largest NetEvidence shifts occur in the text region, while object-level attributions remain comparatively stable. Helpful text shifts the text-region attribution toward positive values supporting the correct answer, whereas misleading text shifts it toward strongly negative values opposing the correct answer.

However, Qwen-VL differs from LLaVA-NeXT

in cases where misleading text is added but the prediction remains correct, shown in Figure 12 (c). Unlike LLaVA-NeXT, where the text region still introduces weak opposing evidence, Qwen-VL often assigns the text region an attribution of the opposite sign. This suggests that, when Qwen-VL rejects the misleading text, the attribution of the text region shifts in the opposite direction, indicating that the model suppresses the textual cue rather than following it.

F Attribution Analysis with Generated Answer as the Target Token

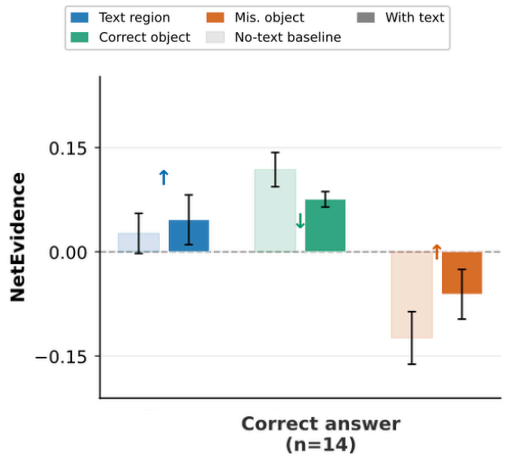
In the main paper, NetEvidence is computed using the *correct answer* token as the Integrated Gradients (IG) target. This choice ensures that the same target token is used before and after adding text, allowing direct comparison across textual variants even when the model's prediction changes.

Figure 13 shows the corresponding attribution analysis when the IG target is set to the model's generated answer token rather than the correct answer token. The figure reports results for both LLaVA-NeXT and Qwen-VL across cases where the prediction changes after adding text (rows 1 and 2) and cases where the prediction remains unchanged (row 3).

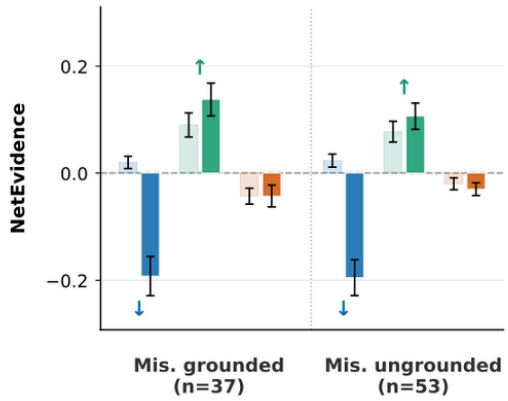
The first two rows differ from the correct-answer-target analysis because the generated answer token changes after the text overlay changes the model's prediction. In these cases, the direction of the NetEvidence reverses relative to Section 5.3. Regions that previously supported the correct answer now oppose the generated incorrect answer, while regions that previously opposed the correct answer now support the incorrect answer selected by the model.

Despite this reversal in attribution direction, the same overall attribution patterns remain. The largest NetEvidence shifts still occur in the text region. Helpful text changes are associated with increased positive evidence from the text region, whereas harmful text changes are associated with strong competing evidence introduced by the text overlay.

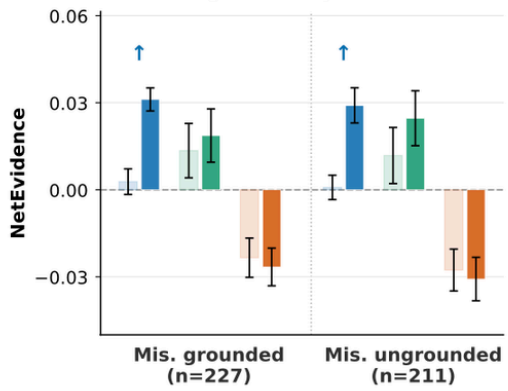
The third row, where the prediction remains correct before and after adding text, follows the same overall pattern as the correct-answer-target analysis in Section 5.3. In these cases, the generated answer token is also the correct answer token, so the attribution values are unchanged. As before, the text region introduces competing evidence while the correct object remains the dominant positive source of support.



(a) Correct-answer text: Wrong → Correct



(b) Misleading text: Correct → Wrong



(c) Misleading text: Correct Prediction Unchanged

Figure 12: Region-level NetEvidence for Qwen-VL with the correct answer as the attribution target. The figure mirrors the LLaVA-NeXT analysis from Section 5.3 across prediction-changing and prediction-unchanged cases. The largest NetEvidence shifts occur in the text region, while object-region NetEvidence remains comparatively stable. Hatched bars show the no-text baseline, while solid bars show the text-overlaid condition.

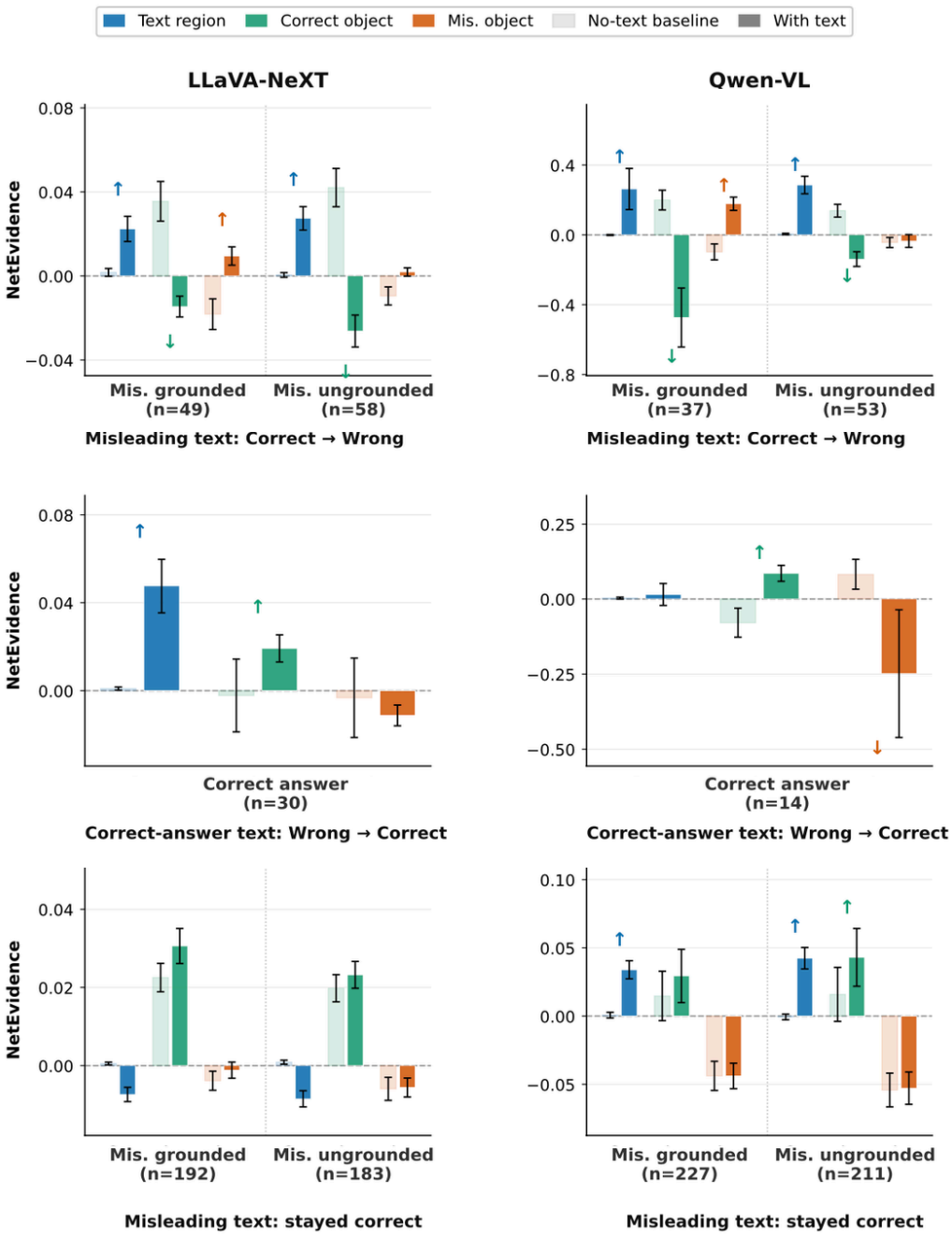


Figure 13: Region-level NetEvidence for LLaVA-NeXT and Qwen-VL when the model's generated answer is used as the attribution target. Unlike the correct-answer-target analysis, positive NetEvidence here supports the answer selected by the model. Despite the reversal in attribution direction for prediction-changing cases, the same overall patterns remain, whereas the largest NetEvidence shifts occur in the text region.

These results show that the attribution patterns observed in the main paper are not specific to the choice of the correct answer token as the IG target.

G Full Attention Analysis Across Models and Variants

This section provides the full attention results extending the analysis in Section 5.4 across models and textual variants.

Figure 14 shows the same attention analysis across all textual variants for LLaVA-NeXT. Each column corresponds to a different text variant, while the two rows separate cases in which the prediction changes from cases in which it remains unchanged.

Across variants, the same overall pattern is repeatedly observed. When the prediction changes after adding text, attention to the text region be-

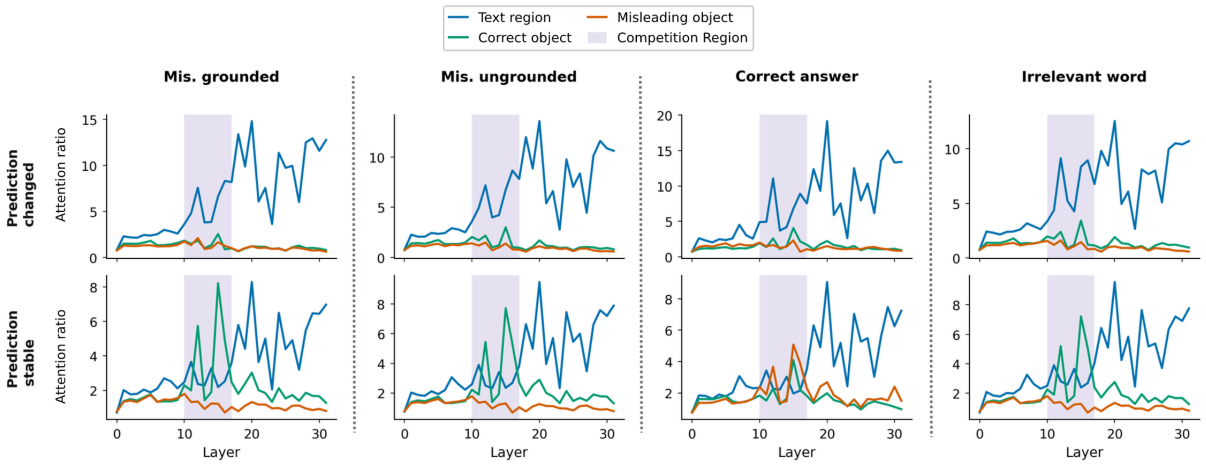
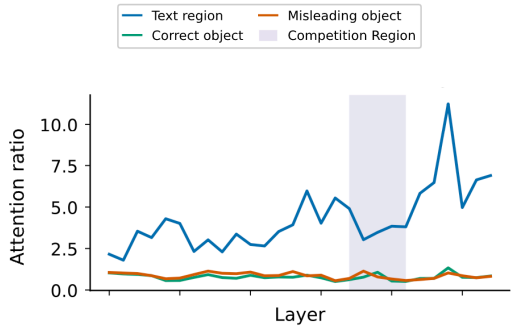
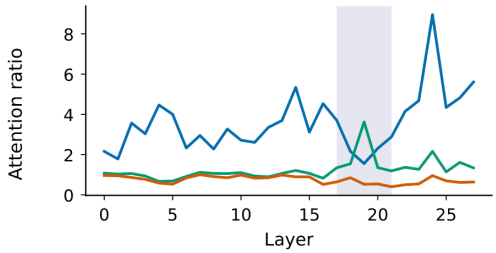


Figure 14: Layer-wise attention across all text variants for LLaVA-NeXT. The top row shows cases in which the prediction changes, while the bottom row shows cases in which the prediction remains unchanged. Across variants, the same middle-layer competition appears. Within the identified middle-layer window, cases in which the prediction changes show higher text-region attention ratios, whereas cases in which the prediction remains unchanged retain stronger object-region attention.



(a) Misleading text: Correct → Wrong



(b) Misleading text: Correct Prediction Unchanged

Figure 15: Layer-wise attention ratio for Qwen-VL under misleading grounded text. The same text-object separation observed for LLaVA-NeXT in Section 5.4 is also observed here. However, the middle-layer separation between text and object regions shifts slightly later, to around layer 19.

comes dominant within the middle-layer competition window. In contrast, when the prediction remains unchanged, attention remains higher for the corresponding object region. This corresponds to the correct object when the prediction remains correct, and to the misleading object when the model consistently follows the misleading answer.

This behavior appears not only for misleading grounded text, shown in the main paper, but also for misleading ungrounded text, correct-answer overlays, and irrelevant text. The consistency of this pattern across variants suggests that the middle-layer competition between text and object regions is a general phenomenon rather than a condition-specific effect.

Figure 15 shows the same attention analysis as in the main paper for Qwen-VL. The same middle-layer competition pattern observed for LLaVA-NeXT in the main paper is also visible here, but shifted to a later layer range.

In Qwen-VL, the clearest competition window appears around layer 19, compared to the middle-layer window around layers 10–15 observed in LLaVA-NeXT. Despite this shift, the same relationship holds. When the model's prediction follows the text, attention to the text region exceeds attention to the object regions within this window, whereas unchanged predictions are associated with greater attention to the corresponding object region.

As in the main paper, this suggests that the model's final prediction aligns with the region that dominates attention within these intermediate layers. The specific layer at which this competition appears therefore varies across architectures, while the underlying attention pattern remains consistent.

These results show that the attention–prediction relationship observed in the main paper is robust

across both models and textual variants. Although the precise layer at which this effect appears varies across architectures, the same underlying pattern remains. The model's prediction is associated with the region that dominates attention within a narrow intermediate-layer window.

H Dataset Generation Prompt

```
You are given:
Question:
{{QUESTION}}

Correct Answer:
{{CORRECT_ANSWER}}

Objects extracted from the image scene graph:
{{OBJECTS_LIST}}

You also have access to:
- the provided image
- the provided segmented image (segments have numeric IDs)

Your task consists of FIVE steps:
————————————————————
STEP 1 — Distractor From Image
————————————————————
Select the best distractor from {{OBJECTS_LIST}}.

Requirements:
- MUST be an object name that appears in {{OBJECTS_LIST}}.
- MUST be different from {{CORRECT_ANSWER}}.
- MUST match the expected answer type and semantic role implied by the question.
- MUST be a valid value for the question's answer slot.
- MUST NOT be the main referenced object mentioned in the question (e.g., do not select "bed" for "What is on the bed?" or "tray" for "What is in the tray?").
- MUST be a semantically plausible alternative answer to the question.
- Prefer the single most confusable option.

If no suitable object exists, set "distractor_in_image" to false.

————————————————————
STEP 2 — Distractor NOT In Image
————————————————————
Propose a plausible distractor object that:
- is a semantically plausible alternative answer to the question
- MUST match the expected answer type and semantic role implied by the question.
- MUST be a valid value for the question's answer slot.
- MUST NOT be the main referenced object mentioned in the question.
- MUST NOT appear in the provided image.
- MUST NOT appear in {{OBJECTS_LIST}}.

You MUST verify absence in BOTH the image and {{OBJECTS_LIST}}.
If you cannot find such an object, set "distractor_not_in_image" to false.

————————————————————
STEP 3 — Select Best Writable Segment
————————————————————
Choose the single best segment_id (from the segmented image) where text could naturally be overlaid.

Rules:
- Must be a realistically writable surface for text overlay.
- Prefer flat, solid, non-organic surfaces.
- Avoid natural elements and generally non-writable objects.

Examples of generally NON-writable objects:
{sky, tree, grass, person, earth, mountain, plant, curtain, water, sea, field, rock, lamp, cushion, sand, pillow, river, blind, flower, hill, palm, towel, light, streetlight, pole, apparel, land, fountain, swimming pool, waterfall, food, pot, animal, lake, blanket, sculpture, vase, fan, shower, flag, ceiling, road, bed, windowpane, sidewalk, sofa, mirror, house, rug, armchair, seat, signboard, skyscraper, path, stairs, runway, awning, ottoman, traffic light, tower, chandelier}

Examples of generally WRITABLE objects:
{wall, cabinet, door, table, chair, car, fence, desk, wardrobe, box, screen, book, bench, computer, bar, bus, truck, airplane, poster, van, ship, tank, microwave, plate, monitor}

If no suitable segment exists, set "segment_id" to false.

————————————————————
STEP 4 — Caption
————————————————————
Write a short caption describing the image.
The caption MUST end EXACTLY with:
"text overlaid on {segment_surface}"

If "segment_id" is false, still produce a short caption ending with:
"text overlaid on false"

————————————————————
STEP 5 — example_accepted
————————————————————
Set "example_accepted" to true ONLY if
```

```
all are satisfied:
1) "distractor_in_image" is not false
2) "segment_id" is not false
3) "distractor_not_in_image" is not
false

Otherwise set "example_accepted" to
false.

____________________
OUTPUT FORMAT (STRICT JSON)
____________________
Return ONLY valid JSON with exactly
these keys:
{
"distractor_in_image": string|false,
"distractor_not_in_image":
string|false,
"segment_id": integer|false,
"caption": string,
"example_accepted": boolean
}

No explanations. No extra text. No
markdown.
```

I Use of AI Assistants

All details in this paper were hand-written with care and attention by a human editor. However, we use AI assistants when trying to find the right words to express deep technical concepts, and for spotting typos.