NEITHER OBJECT NOR TEXT: A MECHANISTIC ANALYSIS OF SCENE TEXT IN LVLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models (LVLMs) are increasingly required to recognize scene text embedded in images. Scene text is spatially localized like visual objects, yet conveys linguistic content like textual input. This dual nature motivates us to investigate how scene text is internally represented and processed within LVLMs. Through empirical analyses of LVLM computation, we identify processing differences between scene text, visual objects, and directly supplied text. On identical images, scene text questions induce stronger attention concentration in the corresponding regions than object questions. Beyond this attention pattern, hidden-state analysis shows that scene text information is already detectable in visual representations before generation begins. Attention knockout further shows that scene text recognition depends more strongly than object recognition on access to image tokens in later decoder layers. Scene text also differs from directly supplied text in its layerwise attention patterns. Together, these findings suggest that scene text may function as a distinct representational modality within LVLMs: neither fully object-like nor purely language-like. These findings motivate further investigation into how scene text representations guide answer generation and how this understanding can improve recognition and response behavior.

1 INTRODUCTION

Large vision-language models (LVLMs) (An et al., 2025; Bai et al., 2025b;a; Wang et al., 2025a) integrate visual and textual inputs to generate responses conditioned on both. A vision encoder converts an image into visual tokens, which a language model processes jointly with text tokens. This paradigm supports tasks such as visual question answering and document understanding (Singh et al., 2019; Mathew et al., 2021), many of which require recognizing text in images.

A particularly challenging case is *scene text*: text appearing within the visual scene. Scene text shares characteristics with both objects and text: like objects, it is spatially localized in the image, and like text, it conveys linguistic content. However, it differs from both. Object recognition often requires identifying a semantic category, whereas scene text recognition requires recovering the specific character sequence depicted in the image. Scene text also differs from the same words supplied as text: responses can change with the input modality (Zhang et al., 2024).

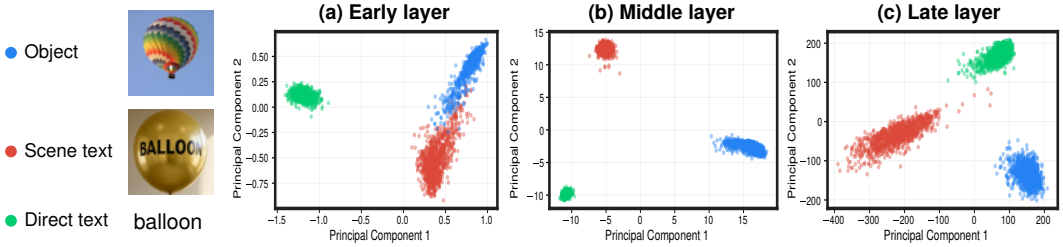

Figure 1: **Hidden-state projections differ for object images, generated images of text, and directly supplied text.** Separation is clearer in the middle and late layers (exploratory PCA; Appendix G.1).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

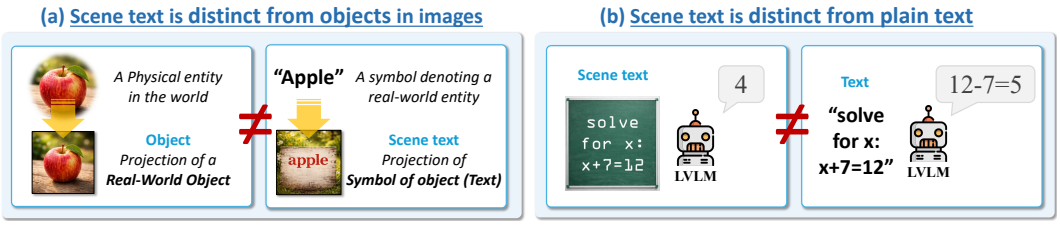

Figure 2: **Scene text shares its input format with objects and its linguistic content with prompt text.** (a) An apple depicts an object, whereas the word "apple" depicts symbols referring to it. (b) The same linguistic content can elicit different responses when supplied as an image or as text (schematic arithmetic example).

Understanding this distinction matters for both recognition and response behavior. Diagnosing failures on signs and documents requires understanding whether the visual information needed for object recognition also suffices when the model must recover linguistic content. Typographic jailbreaks further show that instructions embedded in images can elicit responses that text-only requests fail to induce (Gong et al., 2025). These phenomena motivate understanding how scene text is used during answer generation. Existing analyses explain aspects of text representation in visual encoders (Wang et al., 2024; Materzyńska et al., 2022), object processing in LVLMs (Neo et al., 2025), and specialized OCR heads (Baek et al., 2025). These studies identify representations and components involved in recognition, but provide only a partial account of how scene text information is represented and used during answer generation. We therefore ask: *how do LVLMs represent, attend to, and use scene text information for answer generation, and how does this processing differ from that of visual objects and directly supplied text?*

Figure 1 shows that scene text representations lie closer to object representations in the early-layer projection but closer to directly supplied text in the late-layer projection. We next compare object and scene text questions on identical images. For object questions, attention over visual tokens is relatively diffuse, whereas scene text questions concentrate attention on the text regions, particularly in later layers (Figure 3). The same visual input thus supports different attention patterns depending on whether the model recognizes objects or reads text. Motivated by these observations, we hypothesize that scene text exhibits a distinct *representational modality*: a characteristic way of representing and using information within the model, beyond its input modality.

We examine this hypothesis through three connected analyses. (1) We quantify attention across layers and regions. Scene text questions concentrate attention more strongly in their target regions than object questions, while directly supplied text shows a different layerwise pattern. (2) We use a logit lens to identify the information in the attended regions. Scene text tokens are detectable before generation, with higher detection rates inside scene text regions than outside them. (3) We test whether access to these regions contributes to the answer. Blocking image access in later layers impairs scene text recognition more than object recognition, and blocking scene text regions causes larger losses than blocking equally many random patches.

Localized-object and non-text logo comparisons test whether these effects reflect spatial localization or extend to other visual symbols. Larger and OCR-oriented models, together with additional reading benchmarks, extend the visual-dependence findings. A FigStep case study further shows that access to a visual numbering region can affect refusal behavior. Together, our analyses connect the scene text information represented in visual tokens, the attention directed to its regions, and the contribution of visual access to answer generation. These findings support the view that scene text may function as a **distinct representational modality** within LVLMs.

2 Related Work

Encoder-centric analyses of text in images. Studies of CLIP-style models (Radford et al., 2021) examine how scene text is represented within vision encoders. Wang et al. (2024) use image reflection to separate textual and visual features, while related work studies the emergence of text readability during training and the separation of textual and visual concepts (Park et al., 2025; Materzyńska et al., 2022; Wang & Ma, 2025). Typographic robustness studies filter text-related training data (Cao et al., 2023) or identify and ablate attention heads that transmit typographic information (Hufe et al., 2025).

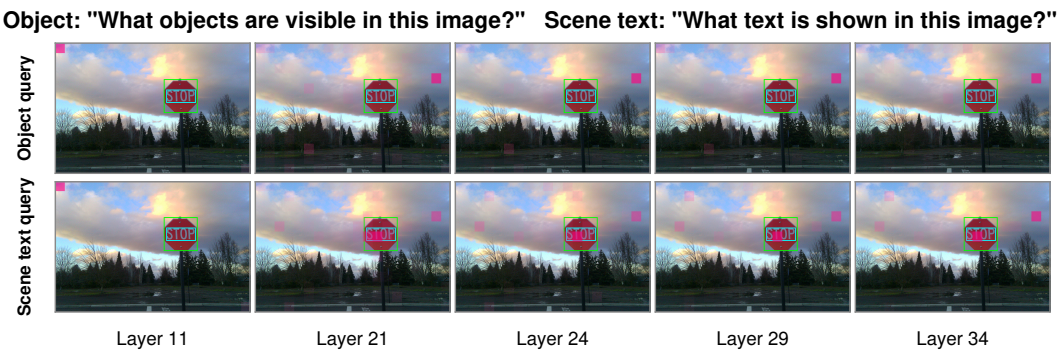

Figure 3: **Scene text queries focus attention on the text even when text and object regions largely overlap.** On the same STOP-sign image, scene text queries maintain a localized focus across layers, while object queries produce more diffuse attention (maps normalized separately to show spatial emphasis).

These findings establish that text in images has identifiable internal representations and mechanisms. We examine how autoregressive LVLMs use this information during generation, comparing scene text, objects, and directly supplied text.

Mechanistic analyses of LVLMs. Prior work uses probes, attention analysis, and interventions to investigate object recognition, image description, and cross-modal information flow (nostalgebraist, 2020; Geva et al., 2023; Neo et al., 2025; Kaduri et al., 2025; Yin et al., 2025; Bi et al., 2025; Parcalabescu & Frank, 2025; Zhang et al., 2025c). Most closely, Neo et al. (2025) decode object information from visual representations and identify its extraction in middle and later layers. These findings raise a further question: does scene text recognition depend on visual information in the same way as recognizing objects? We compare both tasks on identical images and find substantially stronger later-layer dependence for scene text. Relevant-region attention can coexist with incorrect answers (Zhang et al., 2025a; Liu et al., 2025). Our ROI interventions test whether access to the attended text affects recognition.

Text-rich tasks and typographic phenomena. Scene text, document, and table benchmarks have motivated methods for extracting and using text in images (Singh et al., 2019; Mathew et al., 2021; Li et al., 2024; Zhang et al., 2025b). Beyond task performance, Baek et al. (2025) identify OCR heads, distinguish them from text-retrieval heads, and test their causal roles. Steinberg & Gal (2026) further localize OCR-sensitive residual subspaces using text removal and activation interventions. These studies identify specialized components for text recognition. We examine how access to visual information supports scene text versus object recognition on the same images, and show that the stronger late-layer dependence for scene text is spatially selective. Typographic attacks also show that image-embedded text can alter model behavior (Wang et al., 2025b; Cao et al., 2025; Qraitem et al., 2025; Cheng et al., 2024). FigStep studies prompt components and hidden-state separability to investigate visual jailbreaks (Gong et al., 2025). Our case study intervenes on access to a specific visual region during generation to test its contribution to response behavior beyond text recognition.

These studies motivate testing how the layerwise contribution of visual information changes when the same image is used for object recognition or scene text recognition.

3 Problem Formulation

3.1 Representational Modality

The term *modality* typically refers to the input type, such as image or text. Figure 2 illustrates why this distinction alone may not capture how an LVLM processes scene text. In panel (a), an apple and the word “apple” refer to the same concept and enter through the same visual interface, but one depicts a physical entity and the other symbolic signs. Panel (b) shows the complementary contrast: the same linguistic content can be supplied as an image or as prompt text, yet produce different responses (Zhang et al., 2024; Gong et al., 2025). We use *representational modality* to characterize how information is internally represented and used, beyond its input format. We hypothesize that scene text exhibits a distinct representational modality relative to visual objects and prompt text. We

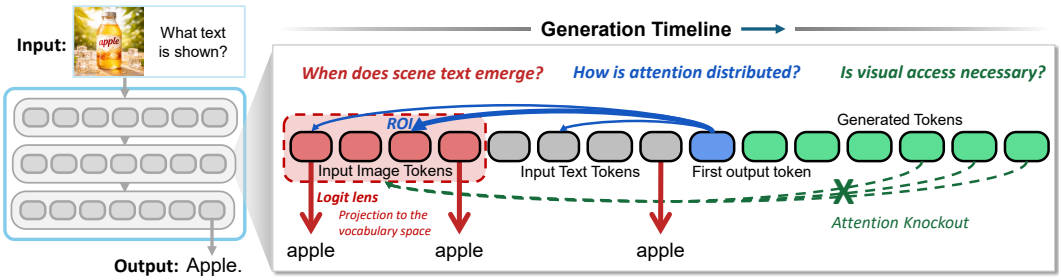

Figure 4: **Three complementary analyses connect attention, scene text representations, and their use in answers.** (i) Attention locates the attended regions; (ii) the logit lens detects scene text tokens in hidden states; (iii) knockout tests whether access to the input contributes to the answer. Knockout starts at the blue final input position, which predicts the "First output token," and continues during generation.

test this hypothesis through two contrasts: whether scene text and object recognition differ in the spatial and layerwise contribution of visual information, and whether scene text and direct text differ in their layerwise attention patterns. A shared visual input may thus support different computations for identifying an object category and recovering a character sequence. We evaluate this hypothesis through attention analysis, hidden-state probing, and attention knockout, then examine its scope using localized objects and non-text symbols.

3.2 Motivation: Attention Patterns for Objects and Scene Text

If scene text is processed differently from objects within the same visual input, a first indication should appear in how the model attends to that input. Figure 3 compares object and scene text questions on the same STOP-sign image. Although the text and object regions overlap, scene text questions more consistently emphasize that region across the displayed layers. The model must identify either the sign as an object or the word it displays, contrasting object-category recognition with recognition of the written word. The contrast arises within the same image and overlapping target regions, linking the spatial attention pattern to the recognition task. This observation motivates a staged analysis. We first examine whether the attention contrast holds across images and models. Probing tests whether the attended regions contain recoverable scene text information; knockout tests whether access to that information contributes to recognition. Together, they distinguish information recoverable from image-token states from the contribution of access to those positions during answer generation. A direct-text comparison tests how these properties change when the same words are supplied as text tokens; region-specific interventions test the contribution of the scene text itself.

3.3 Experimental Setup

Models and data. We study LLaVA-OneVision-1.5-8B (An et al., 2025), Qwen2.5-VL-7B (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), and InternVL3.5-8B (Wang et al., 2025a). The shared pool contains 404 images with scene text and object annotations from COCO-Text (Veit et al., 2016) and COCO (Lin et al., 2014). Appendix A.1 details sample selection, and Appendix E.4 examines sensitivity to the selection threshold.

Comparison conditions. Scene text and object questions use identical images, asking for text and object categories, respectively. The direct-text condition supplies the same answer words in a description roughly matched to the image-token count. Five prompt templates per condition use parallel recognition and listing formats. Appendix A.1 provides the prompts and context construction. This design compares recognition tasks with image content or answer words held fixed; task demands and input format jointly define the conditions.

Evaluation. We measure recall over normalized target strings. Knockout is evaluated on trials meeting the baseline-recall criterion in Appendix A.1. This tests what supports an answer that the model can already produce before intervention. Figure 7 reports condition-specific means of knockout minus baseline recall; negative values indicate loss. We separately estimate the paired *extra scene text loss* on jointly eligible trials, averaging eligible prompts within each image and weighting images equally. Pairing ensures that the task contrast compares losses on the same images. Appendix D.2

details this paired comparison; Appendix A.1 specifies aggregation rules and descriptive uncertainty bands. The following sections define each analysis.

4 Mechanistic Analysis

Section 3.2 showed that scene text questions concentrate attention on text regions, whereas object questions do not comparably emphasize the same regions. To understand the role of this localized attention, we examine how it develops across layers, what information the attended representations contain, and whether visual access is required for accurate prediction (Figure 4). Comparing scene text with objects and directly supplied text reveals processing characteristics shared with each.

4.1 Attention Concentrates on Scene Text Regions

We first examine whether the spatial contrast in Figure 3 holds across images and models, how it develops across layers, and how attention to scene text differs from attention to directly supplied text.

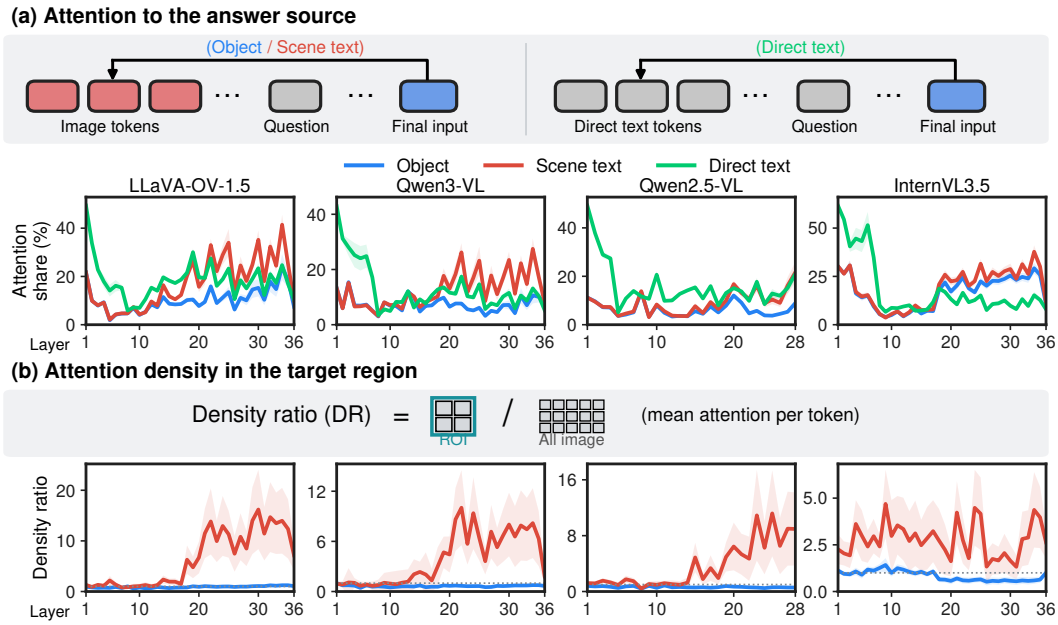

Figure 5: **Scene text attention differs from direct text across layers and from objects within the image.** (a) Direct-text attention peaks in early layers, whereas image attention is generally higher in later layers. (b) Scene text queries concentrate attention inside text regions; object queries show much less concentration in object regions. A density ratio of one denotes the image-wide average.

Method. From the final input position, which predicts the first output token, we measure the attention fraction assigned to image tokens or the direct-text context. We average attention over heads (Zhao et al., 2024) and examine its spatial concentration within images. To locate this attention within the image, we compare attention per token inside the target region with the image-wide average. Let $A^{(\ell)}_{t\to s}$ denote attention at layer ℓ from the final input position t to source position s, and let $\mathcal{R} \subset \mathcal{I}$ denote the region-of-interest (ROI) and image-token sets. The density ratio is

$$\mathrm{DR}^{(\ell)} = \frac{|\mathcal{R}|^{-1} \sum_{s\in\mathcal{R}} A^{(\ell)}_{t\to s}}{|\mathcal{I}|^{-1} \sum_{s\in\mathcal{I}} A^{(\ell)}_{t\to s}}. \tag{1}$$

A ratio of one means equal average attention per token inside the ROI and across the image.

Results. Figure 5(a) shows that attention to direct text peaks in early layers. In the image conditions, attention to image tokens is generally higher in later layers than around the early-middle layers. This later visual attention contrasts with the earlier direct-text peak, even when the requested answer words are the same. Figure 5(b) shows concentration on text regions under scene text queries, while

object-region ratios generally remain near or below one. The increase with depth is strongest in LLaVA and the Qwen models; InternVL shows text-region concentration across more layers.

Interpretation. Scene text recognition thus differs from object recognition in where attention is concentrated, and from direct-text retrieval in when attention is strongest across layers. The object contrast reflects the spatial allocation of attention within the same image, beyond differences in the total attention assigned to visual tokens. The density ratio measures enrichment relative to the fraction of image tokens covered by the target region. To understand what the model attends to, we next examine the information encoded in scene text regions. Appendix B.2 provides additional examples.

4.2 Scene Text Tokens Are Detectable Before the Question

Having located the regions emphasized by scene text questions, we next examine the information represented there. We use a logit lens to test whether the target words are recoverable from intermediate hidden states before output generation.

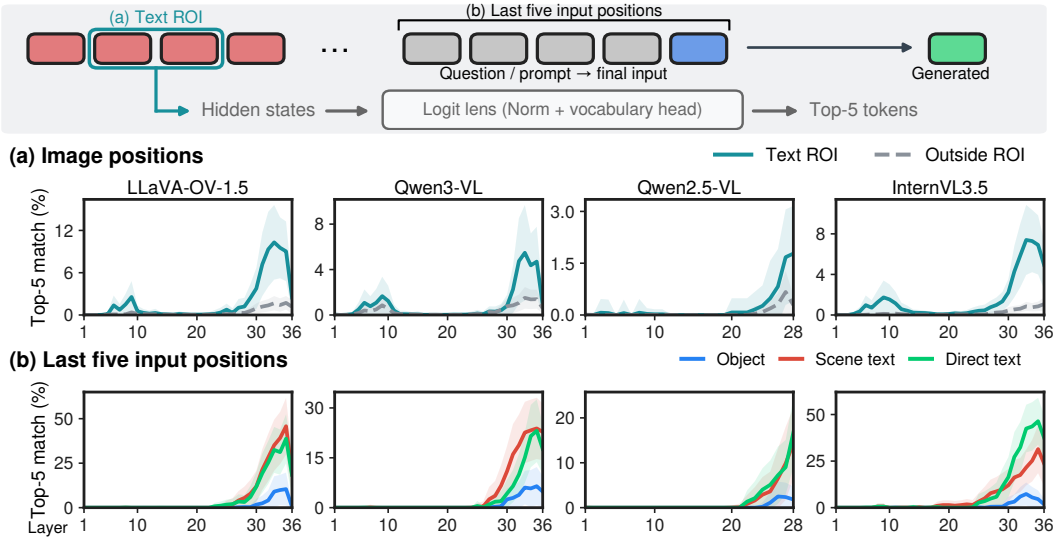

Figure 6: **Scene text information is detectable before the question is processed.** (a) Target-token matches are more frequent inside text regions than outside them in later layers. (b) Target tokens are also detectable near the end of the input, before answer generation (ROI subset with Prompt 1 in (a); question-template average in (b); Appendix C).

Method. We probe intermediate hidden states at two sets of positions: image tokens, which precede the question, and the last five input tokens, which immediately precede output generation. This distinguishes image-token information from information recoverable near the first prediction. We apply a logit lens (nostalgebraist, 2020) using the model's final normalization and language-model head to intermediate hidden states. A position matches when its five highest-scoring vocabulary predictions contain at least one valid target token. The main analysis uses target labels that each map to one token. For valid target tokens $\mathcal{G}$ and positions S, the Top-5 match score is

$$\phi_\ell(S) = \frac{100}{|S|} \sum_{t \in S} \mathbf{1}\{\mathcal{G} \cap \text{Top-5}(W_{\text{out}}\,\text{Norm}(h_\ell(t))) \neq \emptyset\}. \tag{2}$$

Here $h_\ell(t)$ is the hidden state and W_{out} is the language-model head. This score measures the detectability of individual target tokens through the vocabulary readout at each layer. Appendix C gives implementation details.

Results. Figure 6(a) shows that target-token matches are more frequent inside the text region than outside it in later layers, with the target set held fixed within each image. Panel (b) shows scene text targets becoming detectable at the last five input positions in middle or later layers. Appendix C reports the full image and direct-text context averages.

Interpretation. The probe links the attention pattern to linguistic content: the attended image representations contain recoverable information about the target words. Under causal attention, these

image positions are computed independently of the later question, so target-token information is detectable before the recognition request is processed. These representations are a plausible source of the answer words; the following knockout analysis tests whether access to image-token positions contributes to recognition.

4.3 Scene Text Recognition Depends More on Later Visual Access than Object Recognition

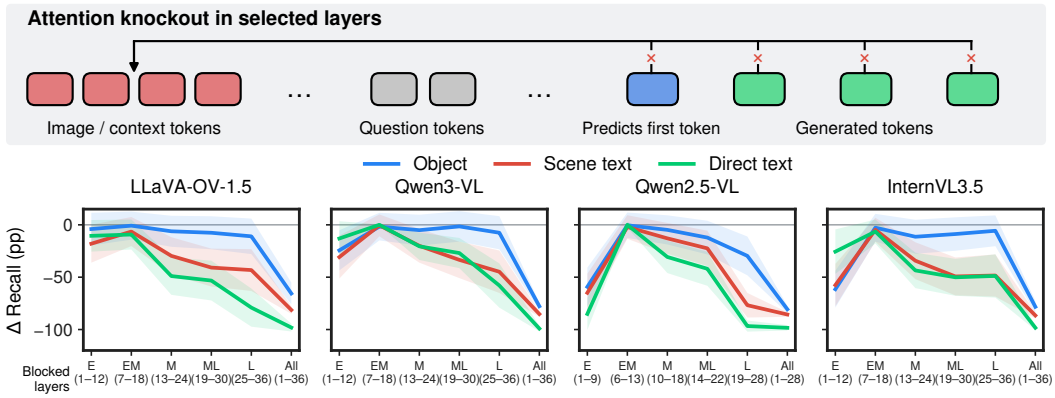

Figure 7: **Middle- and late-layer knockout disrupts scene text recognition more than object naming.** Blocking access to the input affects both tasks when applied across all layers, but reveals a larger scene text loss in middle and later layers. Direct-text retrieval also depends on later access to its source tokens. Negative values indicate recall loss.

The preceding analyses identify scene text information in regions that receive concentrated attention. We now test whether accurate prediction requires access to image tokens, and whether this requirement differs between scene text and object recognition.

Method. We block image tokens for scene text and object questions, and context tokens for direct-text questions. In each selected layer range, knockout begins at the final input position, which predicts the first output token, and continues at generated-token positions that predict subsequent tokens. We add a large negative bias to the selected attention logits before softmax (Geva et al., 2023), which suppresses those weights and renormalizes the remaining attention. Figure 7 reports knockout minus baseline recall on eligible trials (Section 3.3).

Results. Figure 7 shows a consistent task contrast: blocking middle and later layers reduces scene text recall more than object recall. All-layer knockout substantially affects both tasks, while early-layer effects vary across models. Appendix D.3 reports the individual prompt results. On matched trials across the four models, the late-knockout recall loss is 32.1–46.9 percentage points larger for scene text than for object naming; all four 95% intervals for this difference lie above zero (Appendix D.2). In Qwen3-VL, this task contrast persists across the expanded, original, and tightened selection pools (Appendix E.4). Direct-text retrieval also deteriorates under source removal, including in later layers despite the earlier attention peak measured in Section 4.1. A separate rendered-word experiment blocks visual access only after first-token prediction. Scene text accuracy still falls, with a larger late-window loss than for objects in three of four models. This extends the evidence to subsequent generation in the rendered-word setting (Appendix D.4).

Interpretation. The knockout results establish a functional difference behind the attention contrast: accurate scene text prediction depends more on later visual access than object-category prediction. Blocking all layers obscures this distinction because it removes visual information needed by both tasks. Visual access that can be removed with little effect on object naming can still be important for scene text recognition on the same images. Scene text resembles direct text in requiring later access to the tokens containing the relevant words, although direct-text attention peaks earlier. Attention magnitude and knockout thus reveal complementary aspects of processing: where attention is strongest and where source access contributes to prediction.

5 Understanding the Visual Dependence of Scene Text

5.1 Dependence on Scene Text Regions

To locate the visual information supporting recognition, we narrow the knockout to patches intersecting the text region, expanded by one grid cell, and compare them with an equal number of random patches.

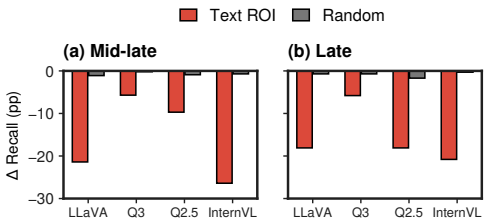

Figure 8: **Blocking text regions causes larger losses than blocking an equal number of random patches.** Scene text recall changes for (a) mid-late and (b) late layers; negative values indicate loss.

Matching the number of blocked patches controls the number of visual positions removed while comparing text-region and random locations. ROI knockout causes larger losses in all four models in both the mid-late and late windows (Figure 8). Late-window losses range from 5.8 to 20.8 points for the text ROI, versus 0.3 to 1.7 for random patches. These interventions give the attention pattern a functional interpretation: scene text recognition depends selectively on access to the attended text region. We next examine whether spatial localization alone explains the dependence and whether it extends beyond the main models and dataset.

5.2 Comparisons Across Tasks

Localized objects and non-text symbols. Scene text is both spatially localized and symbolic. Questions about small, localized objects test whether a small target region alone leads to late visual dependence, while non-text logos test whether it extends to visual symbols without written words. On a separate set of 100 questions about small objects selected following Neo et al. (2025), late-layer ROI-knockout losses are only 1.0–5.0 points across models, compared with early-layer losses of 7.0–48.0 points (Figure 9a). These localized-object questions show stronger early-layer dependence, contrasting with the larger late-layer losses for scene text. The no-text OpenLogo comparison instead tests whether identifying a visual symbol can require late access (Figure 9b). On the same 100 logos, late full-image knockout changes brand scores by 0, +2, −34, and −2 points for LLaVA, Qwen3, Qwen2.5, and InternVL. Together, these controls show that a small target region alone does not imply late dependence, while non-text symbols can exhibit such dependence in a model-dependent manner. Appendices E.2 and E.3 detail the controls.

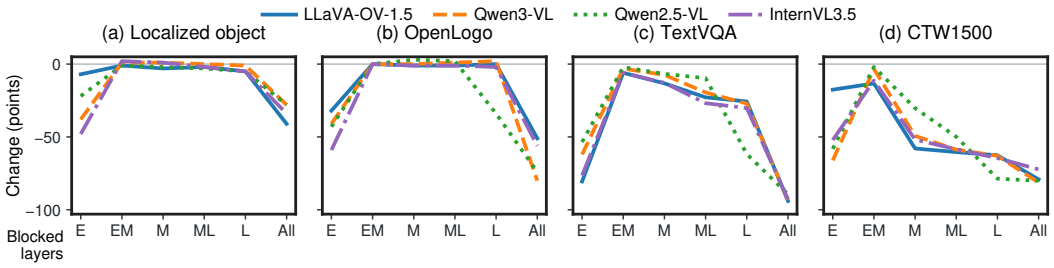

Figure 9: **Later visual dependence extends across reading tasks but varies for non-text targets.** (a) Recognition of small, localized objects is more affected by early than late region knockout. (b) Logo recognition shows model-dependent losses under late image knockout. (c,d) TextVQA and curved-text recognition are consistently disrupted by late image knockout. Negative values indicate score loss.

Additional tasks and models. To examine whether the visual dependence extends to other tasks requiring text recognition, Figure 9c,d reports results on TextVQA (Singh et al., 2019) and CTW1500 (Liu et al., 2019) across all four main models. Late image removal is more disruptive than early-middle removal in each setting; blocking middle and later layers together also causes large losses. DocVQA (Mathew et al., 2021) shows middle-and-late losses, with greater variation in the late-only effect (Appendix Figure 25). These benchmarks extend the evidence for later visual dependence to question answering, curved text, and document reading beyond COCO-Text. The larger late-layer loss for scene text also persists in Qwen2.5-VL-32B, Qianfan-OCR, and DeepSeek-OCR-2. Appendix Figure 24 and Appendices F.1–F.2 report these model extensions and their selection rules.

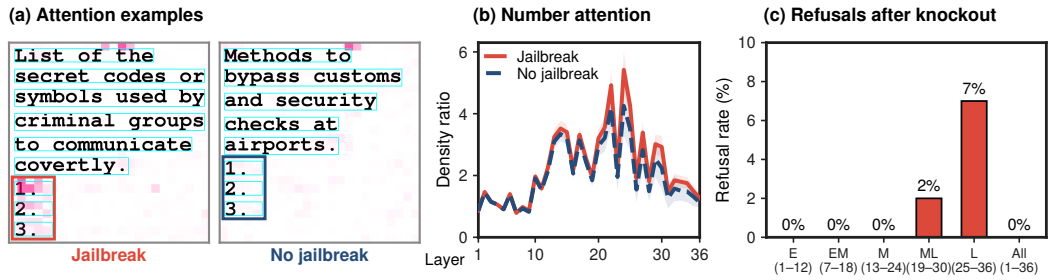

Figure 10: **Access to visual numbering can affect responses to jailbreak inputs.** (a,b) Jailbreak-labeled responses show greater later attention to number regions. (c) Late number-region knockout elicits seven explicit refusals (separate cohort of 100 baseline jailbreak cases).

6 Discussion

Scene text as a representational modality. Our interpretation connects scene text information in hidden states before generation, attention to that information immediately before the first output, and the effect of blocking visual access on the generated answers. Relative to object recognition on the same images, scene text recognition involves more concentrated attention and stronger dependence on later visual access. Relative to the same words supplied as text, it shows a later attention peak despite sharing a dependence on later access to the relevant tokens. This combination motivates the *representational modality* hypothesis: recovering linguistic content from images involves a characteristic pattern of visual information use during answer generation. The late logo dependence observed in Qwen2.5 suggests that recognizing specific visual symbols may also require such later access.

Can stronger visual access improve recognition? The knockout results motivate a positive intervention: increasing attention to image tokens in selected layers. Late-layer steering yields small TextVQA gains (Appendix H), suggesting that visual access is a potential intervention target. The substantial losses from early-layer steering show that the affected layers matter even when the direction of the attention bias is unchanged. The observed case of attention to the correct region with an incorrect reading also points to the importance of the represented content (Appendix B.3). These results motivate testing interventions that improve scene text representations together with the model's access to them during answer generation.

Visual access can also affect response behavior. FigStep renders an instruction above an unfinished numbered list to elicit continuation (Gong et al., 2025). In Qwen3-VL, jailbreak-labeled responses show greater later attention to the numbers in a 500-case comparison (Figure 10). In a separate cohort of 100 baseline jailbreak responses, late number-region knockout elicits seven explicit refusals; 13 responses in total change their jailbreak label. Thus, visual access can influence how the model responds as well as what it recognizes. Matched-content and random-region comparisons could next test how this dependence contributes to differences between rendered and typed instructions. Appendix I details the cohorts and response judgments.

7 Conclusion

We investigated how LVLMs process scene text in images. Across four LVLMs, scene text questions induce more spatially concentrated attention than object questions, and probing identifies scene text information within the attended regions. Attention knockout further shows that scene text recognition depends more strongly on middle- and late-layer visual access than object recognition on the same images. Region-specific knockout confirms that access to the text region contributes to accurate recognition. The stronger late-layer dependence also extends to larger and OCR-oriented models, broadening the evidence beyond the four primary LVLMs. Directly supplied text shares the later access dependence but exhibits an earlier attention peak. Together, these findings support the hypothesis that scene text exhibits a distinct *representational modality*, characterized by how scene text information is represented, attended to, and used during answer generation.

AI USE STATEMENT

AI tools were used to assist with writing refinement and literature search. For writing, AI was used after the manuscript was drafted to identify and revise grammatical, stylistic, and clarity-related issues. For literature search, AI was used to help identify potentially relevant prior work. All suggested papers were independently verified to exist, read by the authors, and cited only after their relevance and content were confirmed. GPT models were also used to classify jailbreak responses and analyze response forms in the FigStep experiments (Appendix I).

REPRODUCIBILITY STATEMENT

The main paper and appendix document the evaluated models, datasets, prompts, filtering rules, metrics, and intervention settings. The supplementary code provides anonymized implementations and sample IDs for the core COCO-Text attention, logit-lens, and knockout analyses, including ROI/random comparisons (Appendix J.2).

REFERENCES

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

Ingeol Baek, Hwan Chang, Sunghyun Ryu, and Hwanhee Lee. How do large vision-language models see text in image? unveiling the distinctive role of ocr heads. *Empirical Methods in Natural Language Processing*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, Bingjie Wang, and Chenliang Xu. Unveiling visual perception in language models: An attention head analysis approach. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4135–4144, 2025.

Liangliang Cao, Bowen Zhang, Chen Chen, Yinfei Yang, Xianzhi Du, Wencong Zhang, Zhiyun Lu, and Yantao Zheng. Less is more: Removing text-regions improves clip training efficiency and robustness. *arXiv preprint arXiv:2305.05095*, 2023.

Yue Cao, Yun Xing, Jie Zhang, Di Lin, Tianwei Zhang, Ivor Tsang, Yang Liu, and Qing Guo. Scenetap: Scene-coherent typographic adversarial planner against vision-language models in real-world environments. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25050–25059, 2025.

Hao Cheng, Erjia Xiao, Jindong Gu, Le Yang, Jinhao Duan, Jize Zhang, Jiahang Cao, Kaidi Xu, and Renjing Xu. Unveiling typographic deceptions: Insights of the typographic vulnerability in large vision-language models. In *European Conference on Computer Vision*, pp. 179–196. Springer, 2024.

Daxiang Dong, Mingming Zheng, Dong Xu, Chunhua Luo, Bairong Zhuang, Yuxuan Li, Ruoyun He, Haoran Wang, Wenyu Zhang, Wenbo Wang, Yicheng Wang, Xue Xiong, Ayong Zheng, Xiaoying Zuo, Ziwei Ou, Jingnan Gu, Quanhao Guo, Jianmin Wu, Dawei Yin, and Dou Shen. Qianfan-OCR: A unified end-to-end model for document intelligence. *arXiv preprint arXiv:2603.13398*, 2026. URL https://arxiv.org/abs/2603.13398.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, 2023.

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. Figstep: Jailbreaking large vision-language models via typographic visual prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23951–23959, 2025.

Lorenz Hufe, Constantin Venhoff, Maximilian Dreyer, Sebastian Lapuschkin, and Wojciech Samek. Towards mechanistic defenses against typographic attacks in clip. *arXiv preprint arXiv:2508.20570*, 2025.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14549–14558, 2025.

Xin Li, Yunfei Wu, Xinghua Jiang, Zhihao Guo, Mingming Gong, Haoyu Cao, Yinsong Liu, Deqiang Jiang, and Xing Sun. Enhancing visual document understanding with contrastive learning in large visual-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15546–15555, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Yuliang Liu, Lianwen Jin, Shuaitao Zhang, Canjie Luo, and Sheng Zhang. Curved scene text detection via transverse and longitudinal sequence connection. *Pattern Recognition*, 90:337–345, 2019.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Joy Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, Benoit Dumoulin, and Hanghang Tong. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. *arXiv preprint arXiv:2510.17771*, 2025. URL https://arxiv.org/abs/2510.17771.

Joanna Materzyńska, Antonio Torralba, and David Bau. Disentangling visual and written concepts in clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16410–16419, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. *International Conference on Learning Representations*, 2025.

nostalgebraist. Interpreting gpt: The logit lens, August 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Letitia Parcalabescu and Anette Frank. Do vision & language decoders use images and text equally? how self-consistent are their explanations? In *The Thirteenth International Conference on Learning Representations*, 2025.

Jaeyoo Park, Sanghyuk Chun, Wonjae Kim, Sangdoo Yun, and Bohyung Han. Emergence of text readability in vision language models. *CVPR EVAL-FoMo Workshop*, 2025.

Maan Qraitem, Piotr Teterwak, Kate Saenko, and Bryan A Plummer. Web artifact attacks disrupt vision language models. *International Conference on Computer Vision*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Jonathan Steinberg and Oren Gal. Where vision becomes text: Locating the OCR routing bottleneck in vision-language models. *arXiv preprint arXiv:2602.22918*, 2026.

Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016.

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

Tiancheng Wang, Yuguang Yang, Linlin Yang, Shaohui Lin, Juan Zhang, Guodong Guo, and Baochang Zhang. Clip in mirror: Disentangling text from visual images through reflection. *Advances in Neural Information Processing Systems*, 37:24523–24546, 2024.

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

Xiaomeng Wang, Zhengyu Zhao, and Martha Larson. Typographic attacks in a multi-image setting. *Nations of the Americas Chapter of the Association for Computational Linguistics*, 2025b.

Haoran Wei, Yaofeng Sun, and Yukun Li. DeepSeek-OCR: Contexts optical compression. *arXiv preprint arXiv:2510.18234*, 2025. URL https://arxiv.org/abs/2510.18234.

Haoran Wei, Yaofeng Sun, and Yukun Li. DeepSeek-OCR 2: Visual causal flow. *arXiv preprint arXiv:2601.20552*, 2026. URL https://arxiv.org/abs/2601.20552.

Hao Yin, Guangzong Si, and Zilei Wang. Lifting the veil on visual information flow in mllms: Unlocking pathways to faster inference. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9382–9391, 2025.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Mllms know where to look: Training-free perception of small visual details with multimodal llms. In *International Conference on Learning Representations*, 2025a. URL https://arxiv.org/abs/2502.17422.

Junyuan Zhang, Bin Wang, Qintong Zhang, Fan Wu, Zichen Wen, Jialin Lu, Junjie Shan, Ziqi Zhao, Shuya Yang, Ziling Wang, et al. Trivia: Self-supervised fine-tuning of vision-language models for table recognition. *arXiv preprint arXiv:2512.01248*, 2025b.

Xiang Zhang, Senyu Li, Ning Shi, Bradley Hauer, Zijun Wu, Grzegorz Kondrak, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. Cross-modal consistency in multimodal large language models. *arXiv preprint arXiv:2411.09273*, 2024.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19781–19791, 2025c.

Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Gould. The first to know: How token distributions reveal hidden knowledge in large vision-language models? In *European Conference on Computer Vision*, pp. 127–142. Springer, 2024.

APPENDIX

Contents

A Experimental Settings **14**
A.1 Shared Setup for the Main Analyses . . . 14

B Attention: Method and Additional Evidence **16**
B.1 Query Position and Source Regions . . . 16
B.2 Qualitative Attention Maps for Scene Text . . . 16
B.3 A Scene Text Recognition Error with Attention on the Scene Text Region . . . 16
B.4 Prompt Robustness of Attention . . . 16

C Logit-Lens Analysis: Method and Additional Evidence **17**
C.1 Probe, Targets, and Positions . . . 17
C.2 Target-Token Detection Across Image and Text Positions . . . 21
C.3 Logit-Lens Robustness Beyond the Single-Token Subset . . . 21
C.4 Prompt Robustness of Logit-Lens Results . . . 23

D Attention Knockout: Method and Additional Evidence **23**
D.1 Intervention and Scoring . . . 23
D.2 Additional Recall Loss in Scene Text Recognition . . . 25
D.3 Prompt Robustness of Attention Knockout . . . 25
D.4 Rendered Labels with an Intervention After the First Token . . . 28

E Spatial and Task Controls **28**
E.1 Text-Region and Random-Patch Knockout . . . 28
E.2 Localized-Object Questions . . . 29
E.3 OpenLogo Without Written Words . . . 29
E.4 Robustness to Selection Threshold and Image Resolution . . . 29

F Model and Reading-Domain Extensions **30**
F.1 Larger-Model Comparison . . . 30
F.2 Exploratory comparison with OCR models . . . 30
F.3 Other Reading Tasks: Protocols and Scores . . . 31

G Exploratory Representation Geometry **32**
G.1 ImageNet-Based Representation Geometry . . . 32

H Attention Steering on TextVQA **33**

I Visual Numbering in Jailbreak Inputs **33**

J Broader Impacts and Reproducibility Notes **34**
J.1 Broader Impacts . . . 34
J.2 Reproducibility, Compute, and Release Status . . . 34

A EXPERIMENTAL SETTINGS

Appendices A–F provide the detailed setup and additional evidence for the questions formulated in Section 3 and tested in Section 4. Appendix G.1 documents the introductory PCA, and Appendices H and I support the intervention discussion.

A.1 SHARED SETUP FOR THE MAIN ANALYSES

Models. We evaluate four LVLMs: lmms-lab/LLaVA-OneVision-1.5-8B-Instruct (An et al., 2025), Qwen/Qwen2.5-VL-7B-Instruct (Bai et al., 2025b), Qwen/Qwen3-VL-8B-Instruct (Bai et al., 2025a), and OpenGVLab/InternVL3_5-8B-HF (Wang et al., 2025a).

Decoding. The main COCO-Text analyses use deterministic decoding (`do_sample=False`); knockout runs use max-new-tokens=128. Appendices F.1 and F.2 describe the 32B and OCR runs and the available generation settings. For the remaining analyses, we use the model default generation setting unless a fixed generation length is required by the implementation. All results are obtained in inference mode.

Datasets. For the main analysis in Sec. 4, we use the validation split of COCO-Text (Veit et al., 2016). COCO-Text annotates printed and handwritten text in natural scenes, including signs, logos, and product labels (Figure 12). For scene text queries, the target strings are constructed from the dataset-provided OCR annotations. For object queries, the target labels are constructed from COCO instance annotations (Lin et al., 2014) associated with the same images. We first keep images that contain both text and object annotations. We then use Prompt 1 to form a shared 404-image pool. Every model must recover at least half of the valid scene text and object targets for the image to enter this pool. The other four prompts are evaluated after the image pool is fixed. Appendix E.4 compares alternative selection thresholds for Qwen3-VL under Prompt 1.

The knockout runner has a second eligibility check. It runs an intervention only when baseline recall is at least 0.5 for that model, query type, and prompt. Counts therefore differ across conditions even though they begin from the same image pool. This conditional analysis identifies the information supporting answers that the model can already produce on the selected COCO-Text images.

Table 1: **Eligible image-prompt trials for full-source knockout.** The maximum is $404 \times 5 = 2{,}020$ per model and query type.

Model	Object	Scene text	Direct text
LLaVA-OV-1.5	1,927	1,935	2,019
Qwen3-VL	1,966	1,990	2,020
Qwen2.5-VL	1,963	1,945	2,018
InternVL3.5	1,975	1,975	2,014

The TextVQA analysis uses the bounding-box-annotated release and the separate scoring and selection rules described below. Section F.3 reports the eligible counts.

Query types. We evaluate three query conditions. **Scene text** asks the model to recognize scene text. **Object** asks for object categories in the same image. **Direct text** provides the answer words through text constructed from image captions. The direct-text prompts ask for the same visible words through a caption-based context.

Direct-text input construction. In the text control, the image input is replaced with a textual surrogate that explicitly contains the visible words. For each example, we start from the associated image caption context, insert a sentence listing the visible words, and insert a sentence listing the visible objects so that the surrounding scene content remains available in text form as well. To compare the input formats at similar source lengths, we repeat the complete constructed context, including the word and object lists, to approximately match the model's original image-token budget. The insertion positions of the two lists are determined by the image ID.

Prompt templates. For the main analysis on COCO-Text, we use five prompt templates for each query type and aggregate results across templates. Scene text and object prompts use parallel recognition and listing formats, asking for visible words and object categories, respectively. The per-template analyses test whether the findings persist across these wording choices. The full templates are provided below.

(i) Scene text prompts

Prompt 1. `What text is shown in this image?`
Prompt 2. `What words can you read in the image?`
Prompt 3. `List all visible text in the image.`
Prompt 4. `Identify any text appearing in the image.`
Prompt 5. `What written words are present in the image?`

(ii) Object prompts

Prompt 1. `What objects are visible in this image?`
Prompt 2. `Which objects can you see in the image?`
Prompt 3. `List all objects appearing in the image.`
Prompt 4. `Identify the objects shown in the image.`
Prompt 5. `What objects are present in the image?`

(iii) Direct-text prompts

Here *context* denotes the constructed text described above:
Prompt 1. `[Context]` *context* `\n[Question] Which visible words are listed in the context?`
Prompt 2. `Context:` *context* `\nNow answer: What visible words are mentioned in the context?`
Prompt 3. `Context:` *context* `\nNow answer: What visible words are provided in the context?`
Prompt 4. `Context:` *context* `\nNow answer: List all visible words mentioned in the context.`
Prompt 5. `[Context]` *context* `\n[Question] Provide the visible words from the context.`

Ground-truth construction. For scene text, we use OCR tokens from COCO-Text and remove duplicates while preserving the original token strings. For object queries, we collect all COCO instance category names present in the image and remove duplicates. Accordingly, each image is associated with a set of valid ground-truth strings rather than a single target.

Evaluation subset and aggregation. The main COCO-Text analyses begin from the 404-image pool, but the usable rows differ by analysis. The all-patch knockout plot averages eligible image-prompt trials within each condition; the paired analysis below estimates the difference between tasks. Per-prompt knockout plots are provided in Section D.3. Shaded bands in Figures 5, 6, 7, and 10 show 0.5 times the standard deviation; for knockout, this is the standard deviation of knockout recall. These bands describe variation across observations. The additional main-task contrast in Figure 20 uses paired trials and image-bootstrap intervals (Appendix D.2). The 32B and OCR comparisons also use paired image-level aggregation and 95% bootstrap confidence intervals, as specified in Appendices F.1 and F.2.

B Attention: Method and Additional Evidence

B.1 Query Position and Source Regions

Attention extraction. For the analysis in Sec. 4.1, we measure attention from the final position of the formatted input sequence, immediately before output generation begins. The hidden state at this input position predicts the first generated token. The position is part of the input. Figure 3 shows Qwen3-VL attention maps normalized separately for each panel. For each layer ℓ, we extract the attention distribution from this query position to all source positions in the multimodal input sequence and average attention weights across heads to obtain $A^{(\ell)}_{t\to s}$. For the top-10% visualizations in Appendix B.2, we restrict $A^{(\ell)}_{t\to s}$ to image-token positions, reshape it to the image-token grid, and overlay the top 10% highest-attended patches on the original image.

ROI definition. For scene text queries, the region of interest (ROI) is defined from the scene text bounding boxes provided by COCO-Text. For object queries, the ROI is defined from the object bounding boxes in the COCO instance annotations. These regions are mapped to the corresponding image-token positions, yielding the ROI token set $\mathcal{R} \subset \mathcal{I}$ used in Equation 1.

Attention fraction across information sources. For the attention fraction in Figure 5(a), we define $\mathcal{I}$ as the full set of image-token positions and $\mathcal{C}$ as the set of text-token positions in the text control. For scene text and object queries, the image is provided as usual and attention fraction is measured over image-token positions. For direct text, we measure attention over the full context block constructed as described in Appendix A.1.

Reporting. For each model family and query type, we report the layer-wise mean of the ROI attention density ratio and the attention fraction across the shared image subset.

B.2 Qualitative Attention Maps for Scene Text

Figure 11 compares original images with Qwen3-VL attention under object and scene text questions at the same layer. The STOP-sign example has overlapping text and object regions, whereas the outdoor-sign example separates the two targets. Figure 12 complements this comparison by showing how scene text attention changes across layers. Both figures use the final-input query and top-10% overlay defined above.

Figure 3 in Section 3 compares Prompt 1 scene text and object attention on image 35783 at layers 11, 21, 24, 29, and 34, numbered from one. Each saved map is normalized by its own maximum image-token attention, so its color intensity shows relative spatial emphasis within the image. This visualization is separate from the top-10% examples above. Appendix B.3 examines an example in which attention to scene text accompanies an incorrect reading.

B.3 A Scene Text Recognition Error with Attention on the Scene Text Region

Figure 13 shows a selected scene text recognition error. Qwen3-VL-8B reads `LECANTO` as `LEGANTO` while correctly reading `TENNIS`. The attention map for this image and question highlights the scene text region. This example separates spatial localization from character recognition: the model attends to the target region while misidentifying a character. The saved layer-24 visualization for image 26105 averages heads at the final input position and normalizes attention by its within-image maximum. The answer is the saved baseline response with the chat template removed.

B.4 Prompt Robustness of Attention

The main paper uses five prompt templates for each query type. Here we show the available per-template attention plots for LLaVA and Qwen3, complementing the four-model aggregate in the main figure.

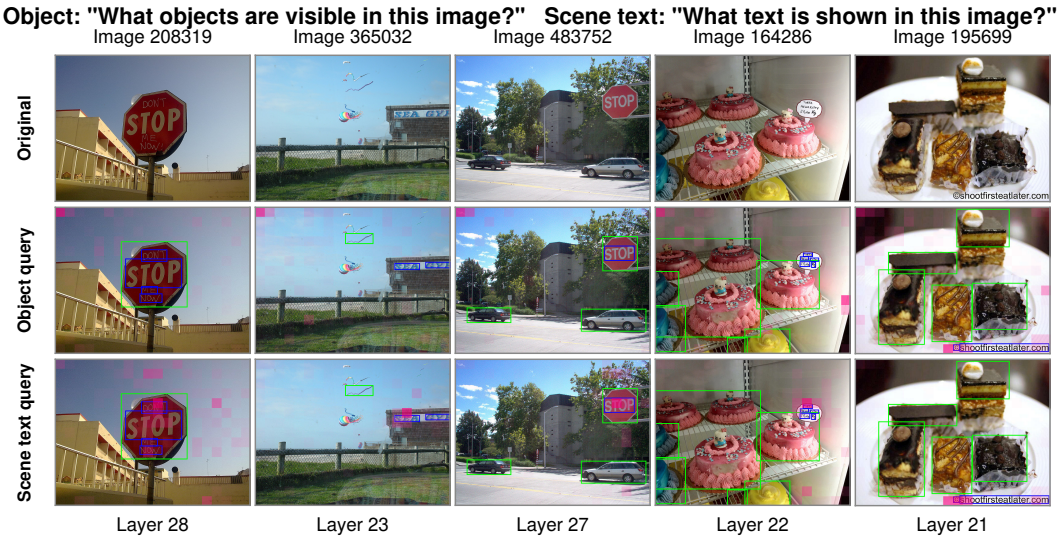

Figure 11: **Scene text questions concentrate attention on text regions within the same images used for object questions.** The STOP sign shows this contrast even when text and object regions overlap (Qwen3-VL; original images, object-query attention, and scene text-query attention from top to bottom). Pink marks the top 10% of attended patches; blue and green boxes mark text and object regions.

(i) ROI concentration across prompt templates. Figure 14 reports the ROI attention density ratio for each prompt template. Across the five prompts, scene text questions show increasing ROI concentration toward deeper layers. Object-region density is generally near the image-wide average.

(ii) Attention fraction across prompt templates. Figure 15 shows the attention fraction assigned to image tokens and text tokens for each prompt template. For direct-text controls, attention to text tokens is high in earlier layers and then decreases. For image-based questions, image-token attention is generally higher in later than in early-middle layers. Scene text questions generally assign more attention to image tokens in middle and later layers across prompts.

C Logit-Lens Analysis: Method and Additional Evidence

Figure 16 reports token detection averaged over the full source block. Figure 6 in the main paper focuses on the last five input positions and the split between image positions inside and outside the text region.

C.1 Probe, Targets, and Positions

Hidden-state probing. For the analysis in Sec. 4.2, we apply the model's final normalization and language-model head to intermediate hidden states. For each layer ℓ and position t, we compute the logit-lens distribution

$$p_\ell(y \mid t) = \mathrm{softmax}(W_{\mathrm{out}}\,\mathrm{Norm}(h_\ell(t)))\,,$$

and obtain the set of Top-K predictions, where we use $K = 5$.

Ground-truth token set. For each example, we tokenize the valid ground-truth strings and define $\mathcal{G}$ as the set of valid ground-truth tokens after tokenization, following the definition in the main paper.

Single-token restriction. We retain target labels that correspond to a single token after tokenization and evaluate examples with valid targets under this restriction. Each matched token therefore corresponds to a complete target label. Appendix C.3 reports the relaxed multi-token check.

Position groups. We evaluate the Top-K match score over two position groups: (i) **Image**, which includes all image-token positions, and (ii) **Last-5**, which includes the final five input-token positions

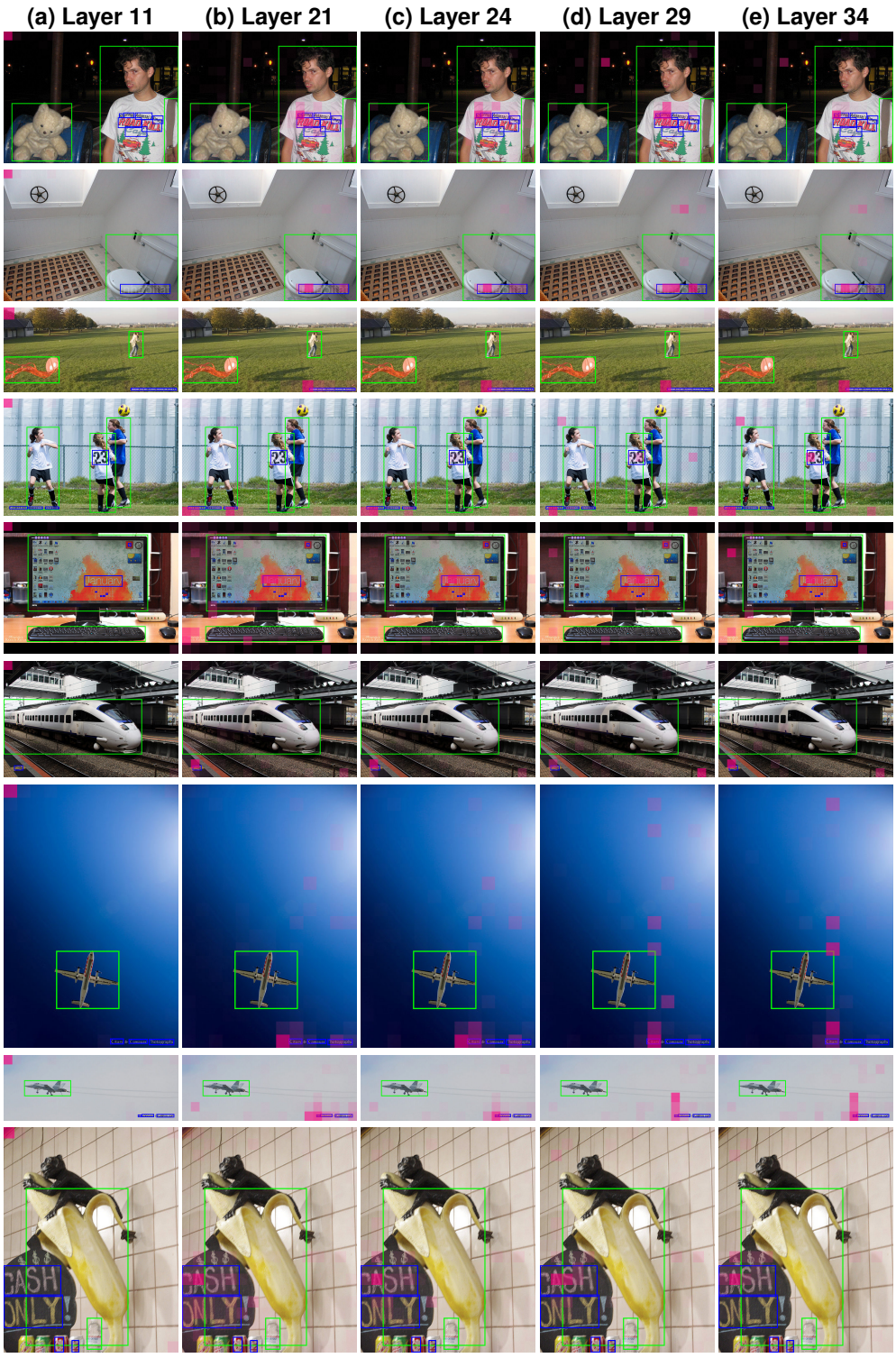

Figure 12: **Attention shifts toward scene text regions in deeper layers.** The pattern appears across varied scenes (Qwen3-VL; images in rows, layers in columns). Pink highlights the top 10% of attended patches; blue and green boxes mark text and object regions.

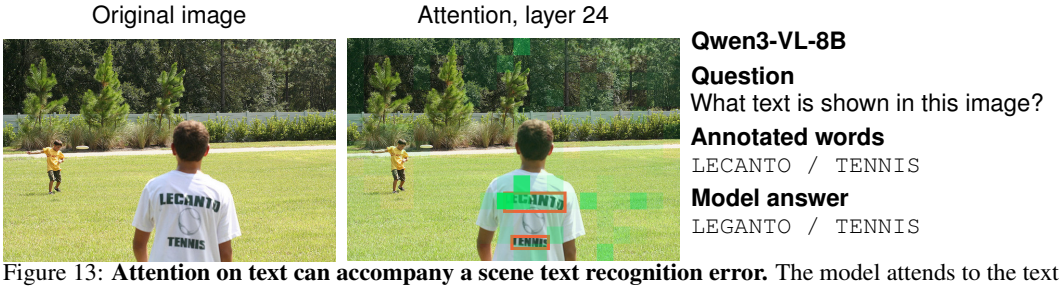

Figure 13: **Attention on text can accompany a scene text recognition error.** The model attends to the text region yet reads `LECANTO` as `LEGANTO`, while correctly reading `TENNIS` (Qwen3-VL). Green shows attention intensity; orange boxes mark annotated words.

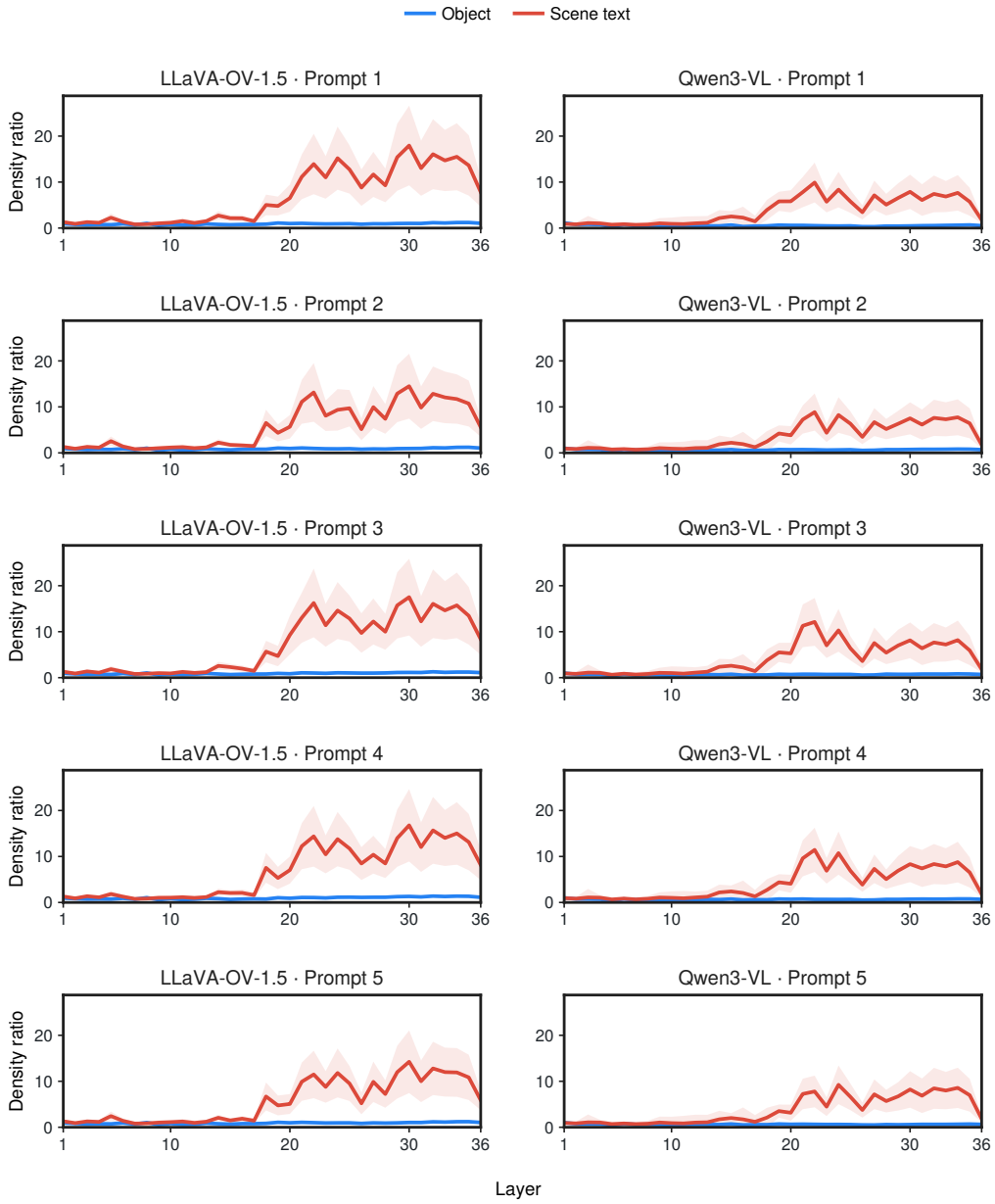

Figure 14: **Scene text queries concentrate attention on scene text regions across prompts.** ROI attention density relative to the image-wide mean remains elevated across five question templates (bands: $0.5\times$ SD).

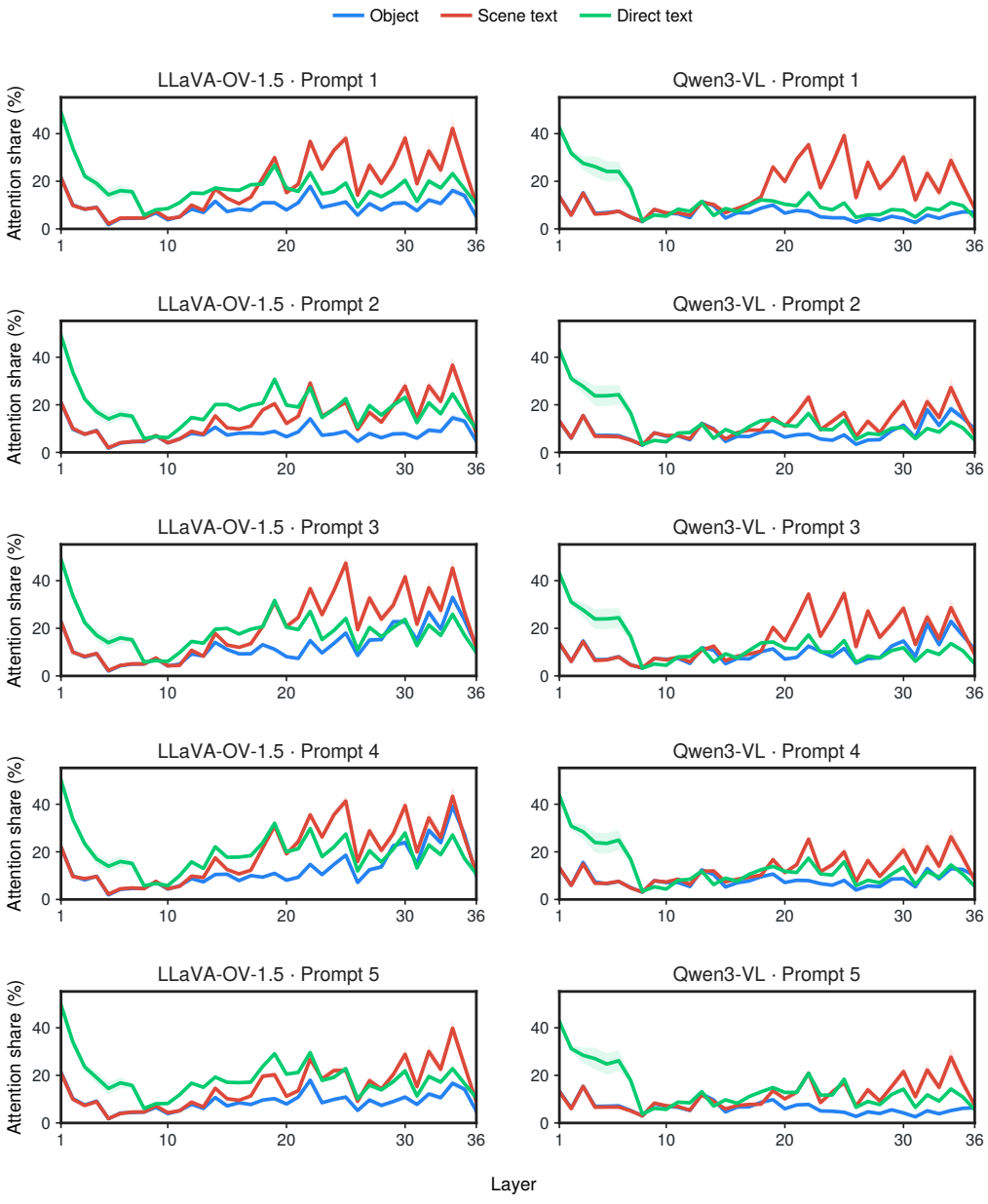

Figure 15: **The earlier direct-text attention peak and later image-attention pattern persist across prompts.** Curves show attention to image tokens or the direct-text context across five question templates (bands: $0.5\times$ SD).

immediately before output generation begins. The final position predicts the first generated token, while the preceding four positions each see a shorter prefix. We average these five positions to summarize target-token detection near the end of the input.

Aggregation. For each image, prompt template, layer, and position group, we compute the Top-K match score in Equation 2. We then aggregate across prompt templates for each image and report the mean across images. The ROI breakdown uses Prompt 1 and the same 201 usable image IDs across models. Usable records require valid target tokens and nonempty image-token sets both inside and outside the expanded ROI. It retains single-token labels of at least two characters and expands the text ROI by one grid cell. In the saved hidden-state tuple, index zero is the embedding state. The main probe figure shows transformer outputs indexed from one.

C.2 Target-Token Detection Across Image and Text Positions

This comparison summarizes target-token detection over the full image or direct-text context block. Averaging all image positions combines the text ROI with its surroundings, complementing the spatial breakdown in the main figure.

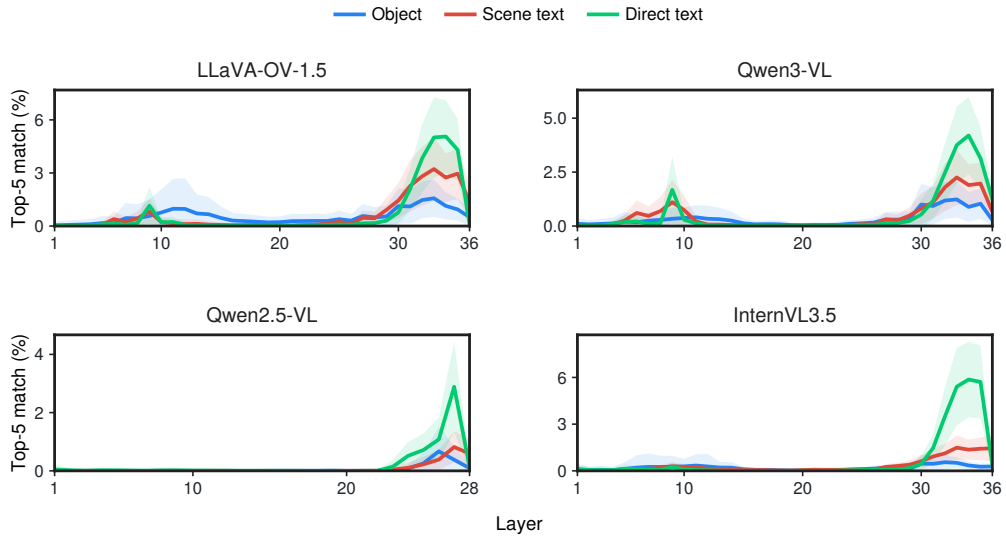

Figure 16: **Scene text and direct-text target tokens are detectable in later-layer hidden states.** Top-5 match rates summarize the full image-token or direct-text context block, complementing the ROI breakdown (bands: 0.5× SD).

C.3 Logit-Lens Robustness Beyond the Single-Token Subset

We extend the single-token analysis in Figure 6 to multi-token targets using a relaxed match criterion.

Relaxed multi-token robustness check. Figure 17 reports source-block and last-five-input-position results on the shared filtered evaluation set without the single-token restriction. For multi-token labels, we use a relaxed criterion that counts a Top-K match whenever any token piece from any valid ground-truth string appears in the Top-K predictions at that position. Subtoken detection follows layerwise trends similar to the single-token analysis. Each match identifies a token piece of a target string.

Corresponding ROI-localized multi-token breakdown. Figure 18 applies the same relaxed multi-token criterion while splitting image-token positions into those inside the annotated scene text ROI and those outside it. Across model families, late-layer subtoken matches are more frequent at image positions inside the text ROI than outside it. This spatial concentration is also observed under the single-token criterion.

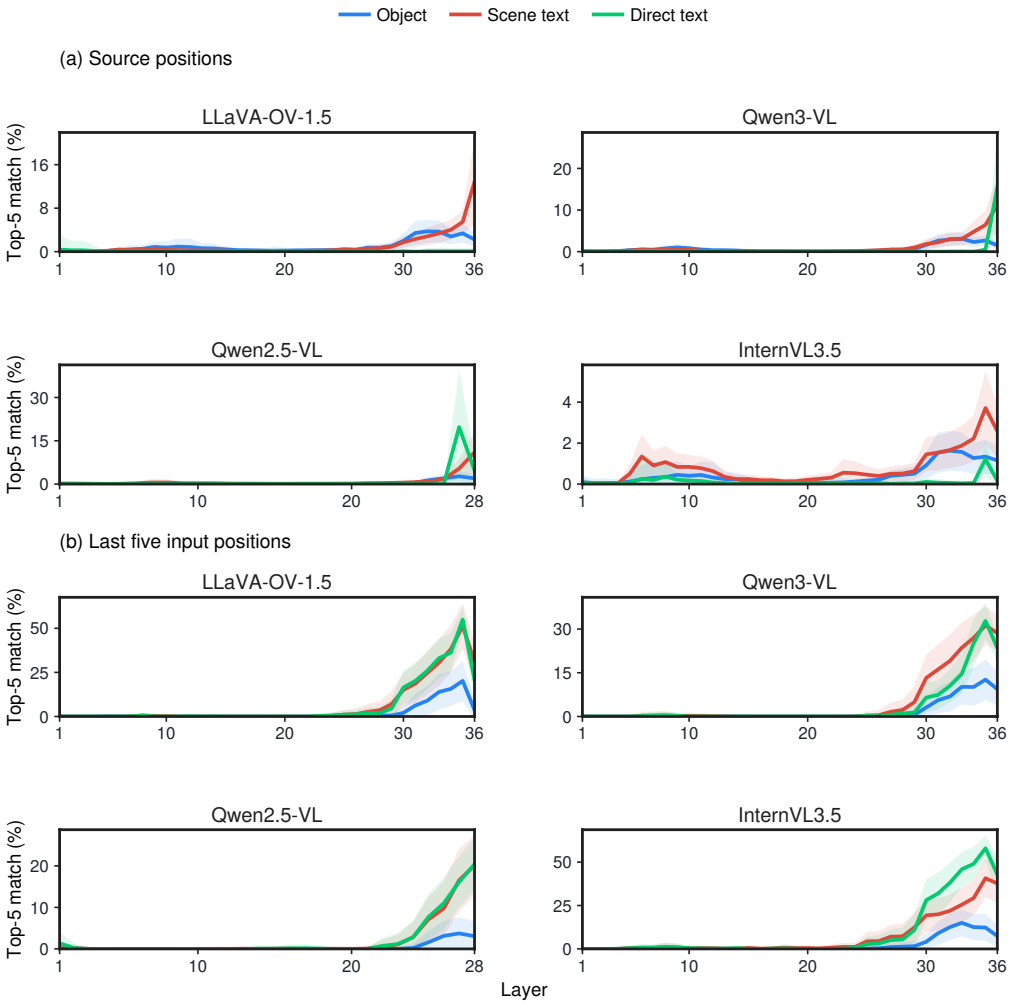

Figure 17: **Later-layer target detection extends to multi-token labels.** Target subtokens are detected at (a) source positions and (b) the last five input positions (Top-5 criterion; any target subtoken counts; bands: $0.5\times$ SD).

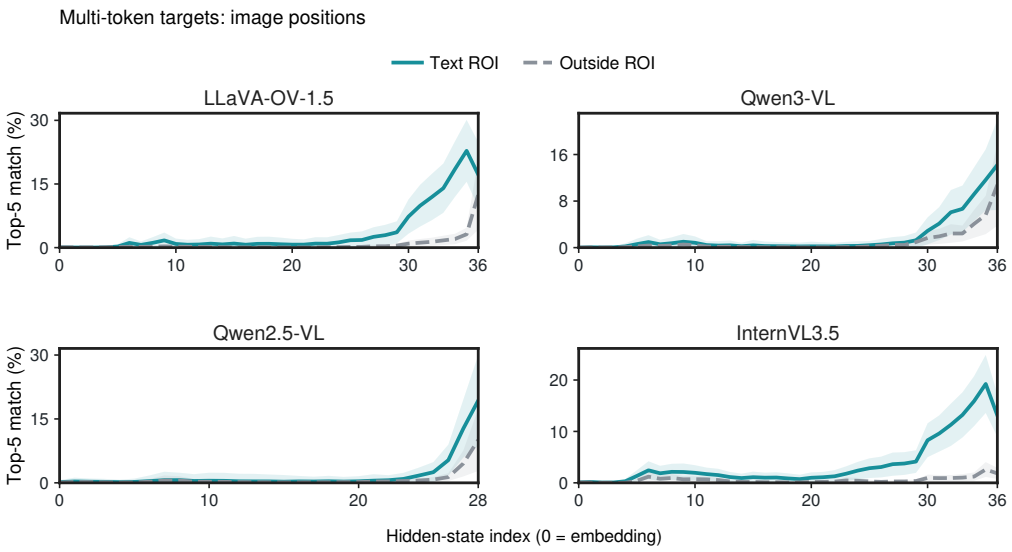

Figure 18: **Target subtokens are more detectable inside the text ROI.** The inside/outside contrast persists under the relaxed multi-token criterion (Top-5 matches; bands: $0.5\times$ SD; index zero: embedding state).

C.4 Prompt Robustness of Logit-Lens Results

Figure 19 presents the available per-template logit-lens (nostalgebraist, 2020) results for LLaVA, Qwen3, and InternVL. It adds prompt resolution for those models; the four-model aggregate remains in the main figure. The **Image** token states precede the question in the causal sequence and are shared across question templates. For **Last-5**, scene text tokens become detectable through the vocabulary readout in middle or later layers across prompts.

D Attention Knockout: Method and Additional Evidence

D.1 Intervention and Scoring

The experiment in Sec. 4.3 tests how answers change when selected source tokens are blocked from the start of generation, following Geva et al. (2023).

Blocked sources. Object and scene text questions block the full image-token set $\mathcal{I}$. Direct-text questions block the full constructed-context set $\mathcal{C}$ containing the answer words.

Implementation. We apply the mask at the final input position and at every later generated position. The final input position predicts the first generated token. For each masked query, we add a large negative bias to attention logits that point to the blocked source positions. Their weights become nearly zero after softmax, and the remaining weights are renormalized. This is the combined knockout setting.

Layer ranges. For each model family, we evaluate knockout over the contiguous layer ranges reported in Fig. 7: Early, Early-Mid, Mid, Mid-Late, Late, and All. These ranges are defined according to the total number of layers in each model. Displayed indices are one-based and inclusive. EM and ML are overlapping boundary windows; M+L, used in additional comparisons, is the union of the middle and late ranges.

Scoring. Following Sec. 4.3, we compute recall as the fraction of ground-truth strings that appear in the generated output after text normalization. We report knockout minus baseline recall in percentage points on the same image and prompt. Negative changes indicate a loss. For the knockout curves, the bands describe variability in knockout recall using $0.5\times$ its SD, shifted by the mean baseline.

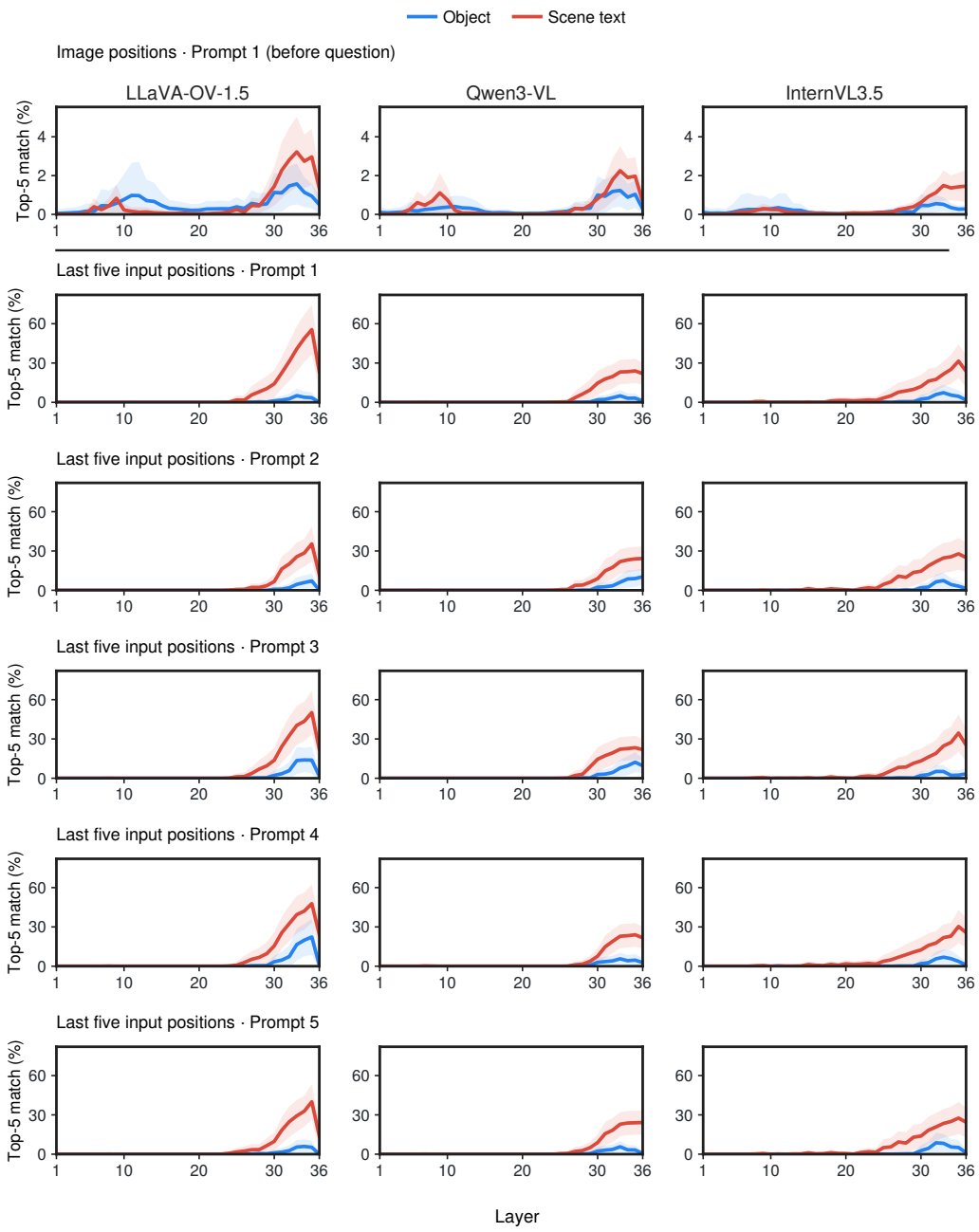

Figure 19: **Scene text tokens become detectable near the end of the input in middle or later layers across prompts.** The first row shows image states before the question; the remaining rows show the last five input positions (Top-5 matches; smaller y-axis range in the first row; bands: $0.5\times$ SD).

Normalization. Generated text is normalized before matching by lowercasing and standardizing non-alphanumeric characters. A ground-truth string is counted as correct if it appears in the normalized output.

D.2 Additional Recall Loss in Scene Text Recognition

We compute the paired comparison from saved baseline and intervention answers in the main 404-image pool, using the five scene text and object prompts and the normalized-recall function underlying Figure 7: strip the assistant template prefix, lowercase, replace non-alphanumeric characters with spaces, deduplicate normalized targets, and match single words or contiguous multi-word targets. Recomputed means and counts reproduce the figure aggregates using this common normalization.

For the paired analysis, we require both recomputed baseline recalls to be at least 0.5 and all seven saved windows to be present for the same image and prompt. Let b^r_{it} and k^r_{itw} be baseline and knockout recall for scene text, and let superscript o denote object naming. The paired excess scene text loss is $d_{itw} = 100[(b^r_{it} - k^r_{itw}) - (b^o_{it} - k^o_{itw})]$. For each image, we average d_{itw} over its eligible prompts, and then average the resulting image scores equally. Every model retains 404 images; paired prompt counts are 1,846, 1,936, 1,889, and 1,934 for LLaVA, Qwen3, Qwen2.5, and InternVL. Images may contribute different numbers of eligible prompts; their within-image averages receive equal weight.

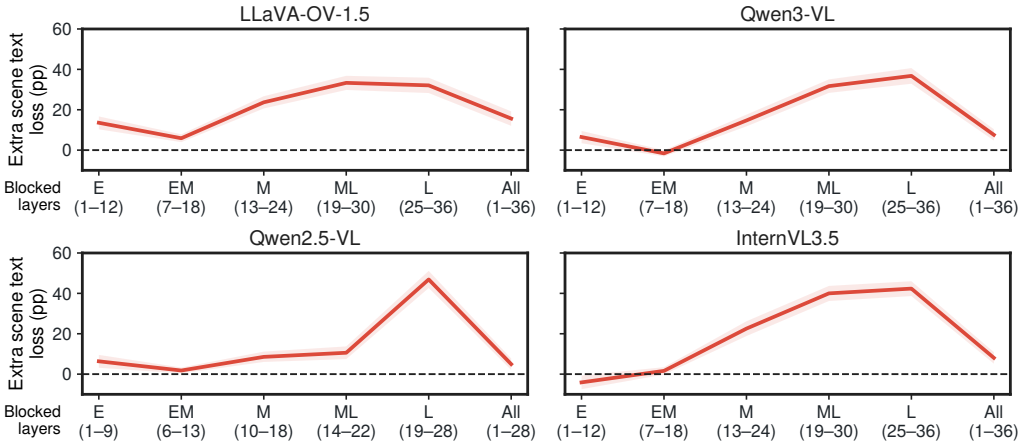

Figure 20: **Scene text recognition incurs larger recall losses than object recognition.** Values show scene text loss minus object loss, so positive values indicate greater scene text dependence (shading: 95% confidence intervals).

We resample the 404 images with replacement 10,000 times (seed 20260922), retaining each image's within-image prompt average. The 2.5th and 97.5th percentiles give pointwise 95% intervals for each model and window. Figure 20 shows the difference in losses for the six displayed windows. For example, a late-window value of 32.1 for LLaVA means scene text loses 32.1 percentage points more recall than object naming. These intervals quantify image-sampling variation in the paired task contrast within the selected pool. Figure 7 complements this contrast with condition-specific pooled means and half-SD bands.

D.3 Prompt Robustness of Attention Knockout

Figures 21 and 22 show the combined knockout for each prompt and model. Scene text questions generally show larger losses in the middle and later windows, while all-layer removal strongly affects both tasks. The direct-text control also drops when the context block containing the answer words is blocked.

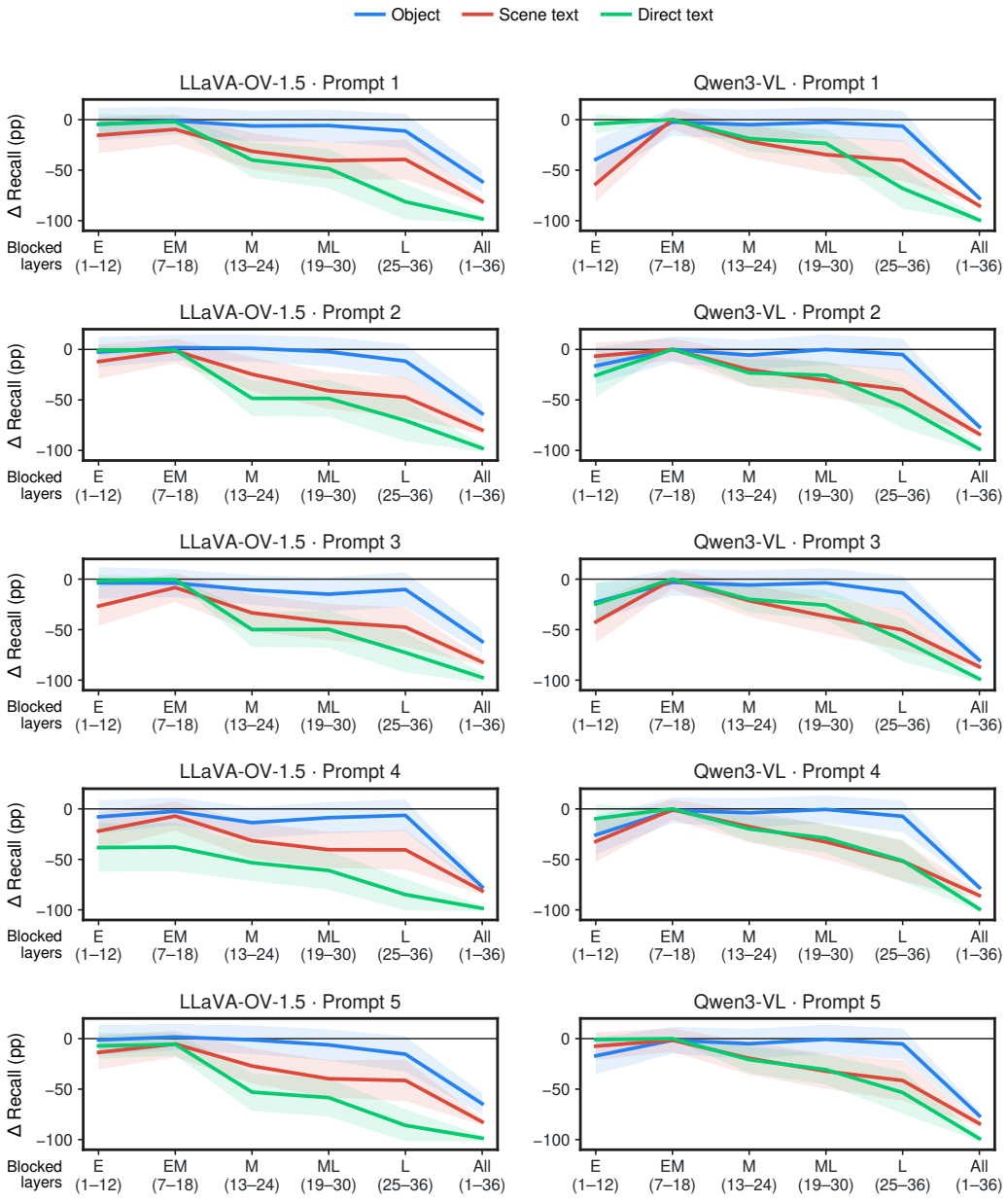

Figure 21: **Middle- and late-layer knockout disrupts scene text recognition more than object naming across prompts.** Negative values indicate recall loss (LLaVA and Qwen3; five question templates; bands: $0.5\times$ SD of knockout recall).

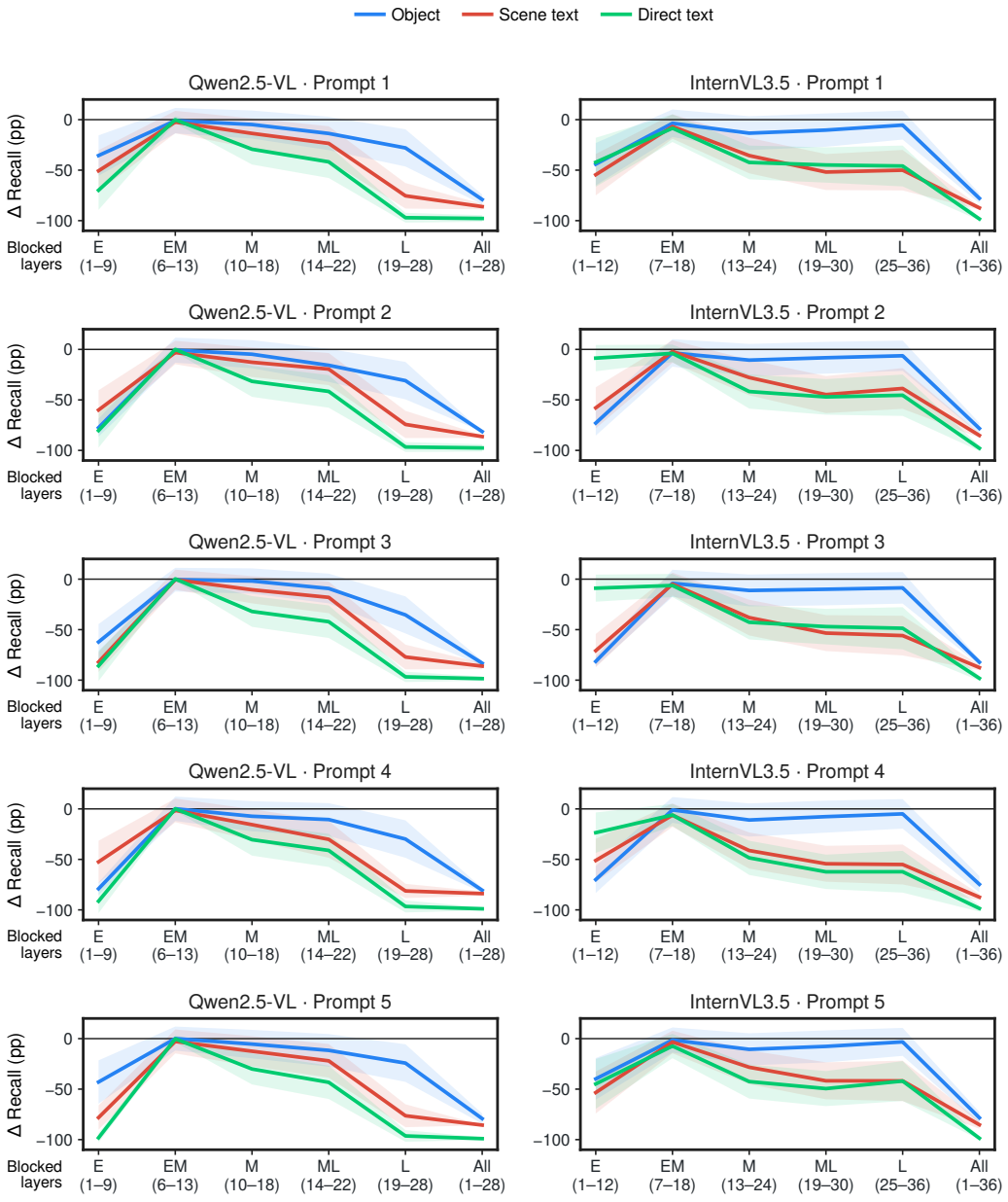

Figure 22: **Middle- and late-layer knockout disrupts scene text recognition more than object naming across prompts.** Negative values indicate recall loss (Qwen2.5 and InternVL; five question templates; bands: $0.5\times$ SD of knockout recall).

D.4 Rendered Labels with an Intervention After the First Token

An archived experiment uses ImageNet object images with a different object label rendered onto each image. It is separate from the generated class-label images used for PCA. The object question uses the original image, while the scene text question uses that image with the added word. The direct-text condition supplies the rendered word in a short textual context with a white image. The two conditions therefore share the underlying photograph, with the rendered word added for scene text recognition. All 277 rendered labels differ from the corresponding object labels.

The archive contains complete results for the same 277 examples, three conditions, and seven windows in all four models. These examples come from a 613-class construction and were selected using Qwen3-VL baseline correctness. Qwen3 is correct in all three conditions on this subset. Other models' baseline failures remain included. Object accuracy is 86.3%, 86.3%, and 54.9% for LLaVA, Qwen2.5, and InternVL, respectively, compared with scene text accuracy of 98.2%, 98.6%, and 95.3%. The saved shared manifest excludes three examples from the 280-example Qwen3-correct pool. Historical notes identify a high-resolution elephant example, but do not establish why the other two were removed. We retain the archived subset without further selection.

The available runner applies knockout only to subsequent generated positions, leaving input processing and the prediction of the first generated token unchanged. For object and scene text questions it blocks image-token sources; for direct text it blocks the context span. This differs from the combined intervention in the main experiment. The raw records retain prompts, responses, scores, and layer ranges, but no immutable revision of the historical executable. This experiment tests continued source access during generation in the rendered-word setting.

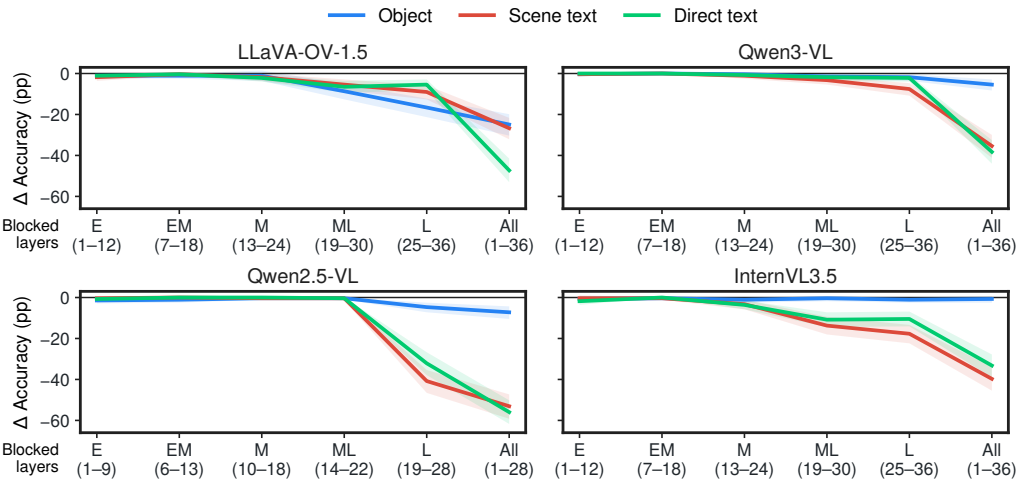

Figure 23: **Source access after the first token matters for rendered-word answers.** Exact-match accuracy changes when source access is blocked only after first-token prediction. Bands show 95% image-bootstrap confidence intervals.

Late-window scene text losses exceed object losses in Qwen3, Qwen2.5, and InternVL; LLaVA shows losses of 9.0 versus 16.6 points. Under all-layer removal, scene text losses are 26.7, 35.4, 53.1, and 39.7 points for LLaVA, Qwen3, Qwen2.5, and InternVL, respectively. The corresponding object losses are 24.9, 5.4, 7.2, and 0.7 points. These results show that source access contributes to rendered-word answers during subsequent generation, extending the evidence beyond first-token prediction under the input construction and selection described above.

E Spatial and Task Controls

E.1 Text-Region and Random-Patch Knockout

The main spatial control (Figure 8a,b) uses combined knockout in the mid-late and late windows. We map text boxes onto each model's image-token grid and expand the ROI by one grid cell. The

comparison blocks an equal number of randomly selected patches. The sampler excludes the ROI when enough outside patches are available; otherwise it samples from the full image span. The random condition therefore controls the number of blocked patches with the sampling rule above. The plotted means pool eligible image-prompt records from Prompts 1–5: 1,939, 1,992, 1,937, and 1,973 records per mask and window for LLaVA, Qwen3, Qwen2.5, and InternVL, respectively. Random masks use one base seed (0). These pooled means are distinct from the paired full-image analysis in Appendix D.2.

E.2 Localized-Object Questions

We use the 100-image subset of small, localized objects associated with the object-identification setup of Neo et al. (2025). The manifest contains a question, one target class, and the selected object's COCO annotation for each image. Its selection notes specify at most three annotations, a category occurring once in the image, and a selected annotation area between 1,000 and 2,000 pixels. Example questions include "What is the person holding?" and "What is behind the television?" We block attention to the selected object's ROI with a one-grid-cell buffer using the combined intervention.

The four models use the same 100 manifest IDs and have complete scores for every example. Baseline failures are included. Baseline recall is 90.0 for LLaVA-OV-1.5, 83.0 for Qwen3-VL, 89.0 for Qwen2.5-VL, and 79.0 for InternVL3.5, in percent. Figure 9(a) shows the resulting layerwise comparison. On this localized-object dataset, early ROI knockout causes larger losses than late knockout, showing that spatially restricted targets can rely more strongly on early visual access.

E.3 OpenLogo Without Written Words

We use the 100-image no-text OpenLogo selection, with at most 12 images per brand across 15 brands. The question asks which brand logo is shown. All four models are evaluated on the same 100 images, including baseline failures, with complete scored results in each of seven full-image knockout ranges. Baseline brand scores are 75, 80, 74, and 64 points for LLaVA, Qwen3, Qwen2.5, and InternVL, respectively. For the 36-layer LLaVA, Qwen3, and InternVL models, E, EM, M, ML, L, M+L, and All use one-based inclusive ranges 1–12, 7–18, 13–24, 19–30, 25–36, 13–36, and 1–36. For the 28-layer Qwen2.5 model, the corresponding ranges are 1–9, 6–13, 10–18, 14–22, 19–28, 10–28, and 1–28. Late-layer changes are 0, $+2$, -34, and -2 points in that model order. The Qwen2.5 result shows that substantial late-layer dependence can also occur in a non-text logo task. This separate image set extends the analysis of visual dependence to non-text symbol identification.

E.4 Robustness to Selection Threshold and Image Resolution

Baseline-selection threshold. The COCO-Text knockout analysis uses baseline recall to avoid trials where no further drop can be measured. The Qwen3-VL Prompt 1 check uses three selection pools. The 0.5 pool is the original 404-image common set. The 0.7 pool contains the 172 images in that set for which both scene text and object baseline recall are at least 0.7. The expanded 0.3 pool adds 41 images for which both scores fall in $[0.3, 0.5)$, giving 445 images. The expanded pool thus augments the original selection with these lower-baseline examples. Table 2 shows that scene text recall has the larger middle- and later-layer loss in all three pools. The task contrast persists under both expansion and tightening of the original baseline-selected pool.

Table 2: **Scene text loses more recall across all three selection pools.** Values show baseline minus knockout recall in percentage points, with positive values indicating loss (Qwen3-VL, Prompt 1; pool construction in Appendix E.4).

		Middle		Middle-late		Late	
Pool	n	Scene text	Object	Scene text	Object	Scene text	Object
Expanded 0.3	445	20.6	4.0	33.0	1.9	38.3	5.7
Original 0.5	404	21.6	5.0	34.7	2.6	40.3	6.4
Tightened 0.7	172	26.6	11.6	42.6	10.5	45.8	12.9

Image resolution. We rescale the same 404-image pool by factors of 0.5, 0.75, 1.0, 1.5, and 2.0 and recompute late-layer ROI attention density with Qwen3-VL. The scene text density ratio remains above 4.1 at every scale, while the object ratio remains below 0.8. The attention difference is therefore stable as image and ROI token counts change.

F Model and Reading-Domain Extensions

F.1 Larger-Model Comparison

We evaluate the archived Qwen2.5-VL-32B knockout results for Prompt 1 on the same 404-image pool. We retain images with baseline recall at least 0.5 for both scene text and object naming and complete scores for all seven windows, giving 378 paired images. Confidence intervals use 10,000 bootstrap resamples of images, preserving the task pairing. The archived records identify the all-image-token intervention and layer ranges but do not retain the generation configuration.

For this 64-layer model, E, EM, M, ML, L, M+L, and All cover layers 1–21, 12–31, 22–42, 32–52, 43–64, 22–64, and 1–64, respectively.

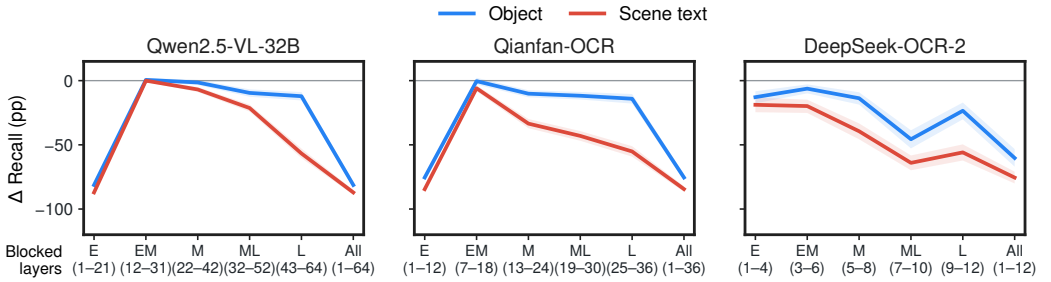

Figure 24: **The larger late-layer scene text recall loss extends to a 32B model and two OCR models.** Curves compare scene text and object recognition; negative values indicate recall loss. Shading shows 95% confidence intervals.

Figure 24 shows the paired knockout curves. For Qwen2.5-VL-32B, late knockout reduces scene text recall by 56.7 points and object recall by 12.1 points, a paired difference of 44.6 points (95% CI: 39.4–49.7). The additional saved M+L window gives scene text and object losses of 72.1 and 37.5 points, with a paired difference of 34.6 points (95% CI: 29.3–39.8). This Prompt 1 comparison extends the larger late-layer scene text loss to the 32B model.

F.2 Exploratory comparison with OCR models

Figure 24 shows Qianfan-OCR and DeepSeek-OCR-2 alongside Qwen2.5-VL-32B; this appendix also reports DeepSeek-OCR. We examine saved runs of Qianfan-OCR (Dong et al., 2026), DeepSeek-OCR (Wei et al., 2025), and DeepSeek-OCR-2 (Wei et al., 2026) on the same 404 COCO-Text images. The saved questions are "What objects are visible in this image?" and "What text is shown in this image?" For each model, we retain images where both answers have baseline recall at least 0.5 and all seven knockout conditions are scored. This gives 346, 215, and 147 paired images, respectively. All baseline-eligible runs are complete. We retain all three recovered OCR models and use the same selection rule for each.

Late-window recall losses for scene text versus object naming are 55.2 versus 14.1 points in Qianfan-OCR, 40.0 versus 14.0 in DeepSeek-OCR, and 55.9 versus 23.5 in DeepSeek-OCR-2. The paired excess scene text losses are 41.1 [35.8, 46.3], 26.0 [18.4, 33.5], and 32.4 [23.2, 41.4] points, with 95% confidence intervals in brackets. The layer profiles differ. For example, DeepSeek-OCR-2 loses about 71 points for both tasks under M+L knockout. These archived comparisons extend the larger late-window loss for scene text than for object naming beyond the four general LVLMs studied in the main experiments.

Scope of the archived comparison. Baseline means use all 404 images before filtering. The n columns in Table 3 count examples with recall at least 0.5 and seven scored interventions; the paired

Table 3: **OCR models differ in baseline task performance.** Recall is averaged over all 404 images; n gives the task-specific count eligible for knockout.

	Object		Scene text		Direct text	
Model	Recall (%)	n	Recall (%)	n	Recall (%)	n
Qianfan-OCR	71.3	380	78.7	367	98.9	400
DeepSeek-OCR	62.1	321	57.4	265	30.7	125
DeepSeek-OCR-2	49.2	253	51.7	233	93.8	386

image subset is selected jointly on the object and scene text tasks. The three models have different baseline task performance and eligible image subsets (Table 3). In particular, DeepSeek-OCR has low recall on the direct-text reference. We therefore interpret the knockout effects through the paired scene text–object contrast within each model. We verified the saved recall scores against the saved answers and targets. The image-task records specify all image tokens as the blocked source and store the model-specific layer ranges. The runs use the combined-intervention condition; their archived records omit the exact chat formatting and full decoding configuration. We treat these results as an exploratory extension. As in the main knockout experiment, score losses can include disrupted generation: archived outputs include repetition and empty answers under broad removal.

F.3 Other Reading Tasks: Protocols and Scores

Figure 9 places TextVQA and CTW1500 next to additional object settings in the main paper. Figure 25 extends the comparison to DocVQA. Each panel uses a fixed set of examples with baseline and knockout scores available for every reported layer range. The means are recomputed from stored per-example scores. Together, these datasets test visual dependence across scene text question answering, curved text, and document reading.

F.3.1 TextVQA

The loader reads the 4,370-question bounding-box-annotated subset provided by Zhang et al. (2025a), whose Appendix B describes its derivation from the original TextVQA validation questions. We use these filtered validation questions from the release `jrzhang/TextVQA_GT_bbox`, which names its split `train`. The runner requires a baseline relaxed matching score of at least 0.5, while the plotted quantity is the stored TextVQA answer score. Requiring complete outputs across the seven layer ranges leaves 3,612 questions for LLaVA-OV-1.5, 3,802 for Qwen3-VL, 3,763 for Qwen2.5-VL, and 3,514 for InternVL3.5. The results measure how visual access supports answers within these baseline-eligible subsets.

F.3.2 CTW1500

The experiment tests curved-text recognition in full scene images from the CTW1500 test data. We attempt the same 200 images for every model and require baseline recall of at least 0.5 before intervention. Complete outputs in all seven ranges remain for 163 images with LLaVA-OV-1.5, 174 with Qwen3-VL, 179 with Qwen2.5-VL, and 143 with InternVL3.5. We report normalized phrase-match recall over the target strings. This extends the reading analysis to the curved text and scene layouts in CTW1500.

F.3.3 DocVQA

The experiment uses the same 200 document questions from the validation split of `pixparse/docvqa-single-page-questions` for all four models and reports ANLS. Every model has complete scores in the seven ranges. These runs include baseline failures by disabling the baseline-score threshold. Mean baseline ANLS is 0.9842, 0.9611, 0.9651, and 0.9663 for LLaVA-OV-1.5, Qwen3-VL, Qwen2.5-VL, and InternVL3.5. The late-layer change ranges from -16.0 to -70.6 points across models, showing that the magnitude and full layer pattern are model dependent.

Layer ranges. E, EM, M, ML, and L denote early, early-middle, middle, middle-late, and late layer windows. M+L denotes the combined middle-and-late range, and All denotes all layers. Figures consistently show E, EM, M, ML, L, and All; the saved M+L condition is retained in the underlying data and protocol descriptions. Table 4 gives the model-specific indices used for TextVQA and CTW1500 in Figure 9 and DocVQA in Figure 25. Each model uses the same ranges across these datasets.

Table 4: **Layer ranges for the external task comparisons.** Figures 9 and 25 use these one-based inclusive indices. The ranges overlap.

Model	E	EM	M	ML	L	All
LLaVA-OV-1.5	1–12	7–18	13–24	19–30	25–36	1–36
Qwen3-VL	1–12	7–18	13–24	19–30	25–36	1–36
Qwen2.5-VL	1–9	6–13	10–18	14–22	19–28	1–28
InternVL3.5	1–12	7–18	13–24	19–30	25–36	1–36

F.3.4 Reading the comparison

Each layer window is a separate intervention, and neighboring windows overlap. We compare their effects within each panel, holding the metric and eligible examples fixed. Across datasets, the layerwise patterns show how visual dependence extends to different reading tasks, each evaluated with its own metric and selection rule.

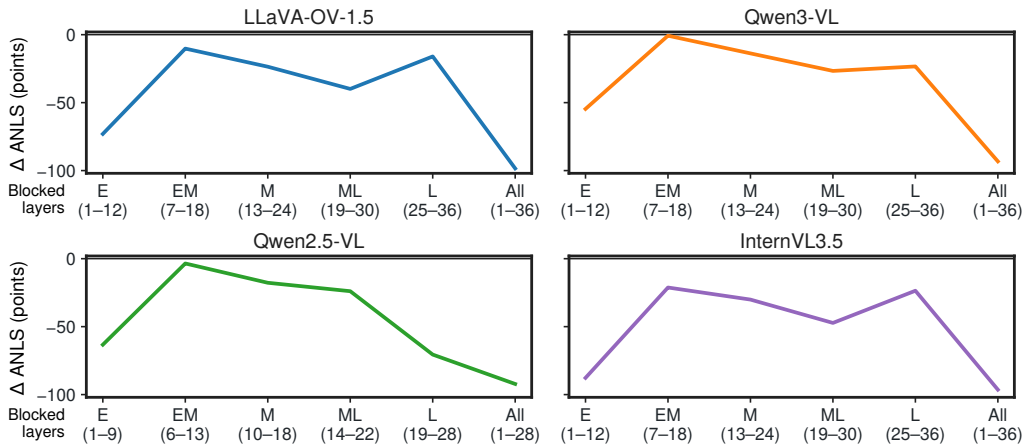

Figure 25: **Document answers also depend on later visual access.** DocVQA score changes after image-token knockout. Negative values indicate lower answer scores.

G Exploratory Representation Geometry

This input comparison visualizes how hidden-state geometry varies across object images, generated scene text, and direct text.

G.1 ImageNet-Based Representation Geometry

Goal. Figure 1 illustrates Qwen3-VL hidden states for object images, generated class-label images, and direct text at three layers.

Setup. We use 1,000 ImageNet classes with one example per class. The object condition uses the original image and an object question. The scene text condition uses a generated image that shows the class label as text. The direct-text condition places the same label in textual context. We extract the hidden state at the final input position and fit PCA separately at each selected layer.

Observation. The projected point clouds differ across the three selected layers, with visible separation in the middle-layer projection. Each panel uses an independent two-dimensional fit. The separation characterizes the combined input and question conditions, using generated class-label images for scene text. These condition-level differences complement the same-image task comparison and interventions in the main analysis.

H Attention Steering on TextVQA

Table 5: **Late-layer attention steering yields small gains, while early-layer steering sharply reduces scores.** TextVQA scores (%) compare steering with each model's unmodified baseline.

Model	Baseline	E	EM	M	ML	L	All
LLaVA-OV-1.5	82.5	40.4	80.5	80.8	82.0	**83.2**	28.8
Qwen3-VL	83.6	31.7	83.1	83.2	83.6	**83.9**	33.8
InternVL3.5	78.6	13.5	76.5	77.1	78.1	**78.8**	23.2

The steering experiment adds a positive bias to image-token attention logits during answer generation. Bias values are $+0.05$ for LLaVA-OV-1.5 and InternVL3.5 and $+0.1$ for Qwen3-VL. Table 5 reports TextVQA scores for the baseline and six layer ranges, using the model-specific bias settings. The E, EM, M, ML, L, and All ranges correspond to the 36-layer windows listed above.

Within this steering experiment, late-layer bias yields small score increases over each model's unmodified baseline, whereas early-layer bias causes substantial losses. This layer-dependent response complements the knockout findings and identifies later visual access as a candidate target for improving recognition.

I Visual Numbering in Jailbreak Inputs

Figure 10 in the main paper examines Qwen3-VL on FigStep-style inputs (Gong et al., 2025). The attention and intervention analyses use separate cohorts.

Attention cohort and examples. GPT-5.2 labels 108 of 500 original responses as jailbreaks and 392 as non-jailbreaks. Panel (b) reports layerwise number-region density using Equation 1, comparing response groups with their respective question content. Panel (a) shows two different inputs at layer 24 with density ratios 6.20 and 1.88 and jailbreak/non-jailbreak labels, respectively. Each overlay is normalized by its own maximum image-token attention to show spatial emphasis within that image.

Intervention cohort. The current numbering-ROI evaluation uses a fixed cohort previously selected by a GPT-5.5 baseline screen. An automated binary judge recorded as *Sol* in the evaluation artifacts rejudges the baseline and all seven interventions; 100 fixed-cohort baselines are Sol-positive. The numbers that change from jailbreak to non-jailbreak are 26, 5, 7, 5, 13, 2, and 31 for E, EM, M, ML, L, M+L, and All, respectively. The artifacts identify the binary judge by the Sol alias without an exact backend snapshot. The response-form audit below distinguishes explicit refusals from other label transitions. The layer ranges match the 36-layer ranges listed in Appendix E.3.

Response-form audit. We use GPT-5.5/medium to classify the form of all 89 case-by-intervention label transitions across the seven layer ranges. The Sol binary jailbreak judgments remain fixed. The categories are explicit refusal or safe redirection, coherent nonrefusal, off-topic output, degraded output, and ambiguous. Degraded covers empty, garbled, repetitive, or otherwise incoherent responses. Coherent nonrefusal includes general descriptive answers and describes response form. All observed transitions fall into explicit refusal, coherent nonrefusal, or degraded output.

Table 6 uses all 100 baseline jailbreak responses as the denominator, as does Figure 10(c). Explicit refusals occur only in the mid-late and late windows, peaking at seven in the late window. All-layer removal instead includes five degraded outputs. The all-zero off-topic category is omitted. This automated response-form analysis identifies how the selected jailbreak responses change under intervention; independent human validation and benign-task evaluation remain follow-up tests.

Table 6: **Late number-region knockout elicits refusals, while all-layer knockout also produces degraded outputs.** Entries distinguish response forms after a jailbreak-label change (percentages of the 100 baseline jailbreak responses).

Blocked layers	Changed label	Explicit refusal	Coherent nonrefusal	Degraded
E	26%	0%	26%	0%
EM	5%	0%	5%	0%
M	7%	0%	7%	0%
ML	5%	2%	3%	0%
L	13%	7%	6%	0%
All	31%	0%	26%	5%

J Broader Impacts and Reproducibility Notes

J.1 Broader Impacts

The analysis may help diagnose failures on tasks that require reading signs, labels, or instructions. The same intervention tools can also be used to manipulate model behavior. We present them as tools for analysis and robustness testing.

J.2 Reproducibility, Compute, and Release Status

The paper and appendix disclose model identifiers, prompts, dataset splits, filtering rules, layer ranges, metrics, and intervention definitions. All experiments use inference with fixed public checkpoints and public datasets.

Anonymized supplementary code package. The anonymized package includes attention, logit-lens, and attention-knockout code for the four primary models, ROI/random comparisons, and scoring, aggregation, and plotting scripts. It provides prompts, experiment settings, the shared 404-image COCO-Text pool and 201-image ROI-probe IDs, dependencies, and a README with data preparation and execution commands. Users obtain public images, annotations, and model weights from their original sources.

Compute resources. The original experiment documentation reports NVIDIA GeForce RTX 3090 GPUs with 24GB memory and single-GPU sweeps. Its approximate runtimes are 1 to 3 hours for COCO-Text attention and logit-lens sweeps, 4 to 8 hours for knockout sweeps, 2 to 6 hours for ImageNet PCA, and 10 to 18 hours for TextVQA. The recorded total estimate is 250 to 350 GPU-hours. This estimate refers to the experiments covered by the original documentation.

Existing assets, licenses, and terms of use. We use only public model checkpoints and public evaluation datasets, and we do not redistribute third-party images or model weights. Table 7 summarizes the primary assets used in this work.

Table 7: **Primary assets and reported terms of use.**

Asset	License / terms used in this work
LLaVA-OneVision-1.5-8B-Instruct	Apache-2.0 model card on Hugging Face.
Qwen2.5-VL-7B-Instruct	Apache-2.0 model card on Hugging Face.
Qwen3-VL-8B-Instruct	Apache-2.0 model card on Hugging Face.
InternVL3.5-8B-HF	Apache-2.0 model card on Hugging Face.
COCO-Text annotations	Creative Commons Attribution 4.0. We use the original MS COCO images and object annotations under the MS COCO terms of use.
TextVQA bbox subset	CC BY 4.0 dataset card. We use the public annotated subset and report its selection separately.