Read or Ignore? A Unified Benchmark for Typographic-Attack Robustness and Text Recognition in Vision-Language Models

Futa Waseda[1,2,3,†], Shojiro Yamabe[1,4], Daiki Shiono[1,5], Kento Sasaki[1], and Tsubasa Takahashi[1,‡]

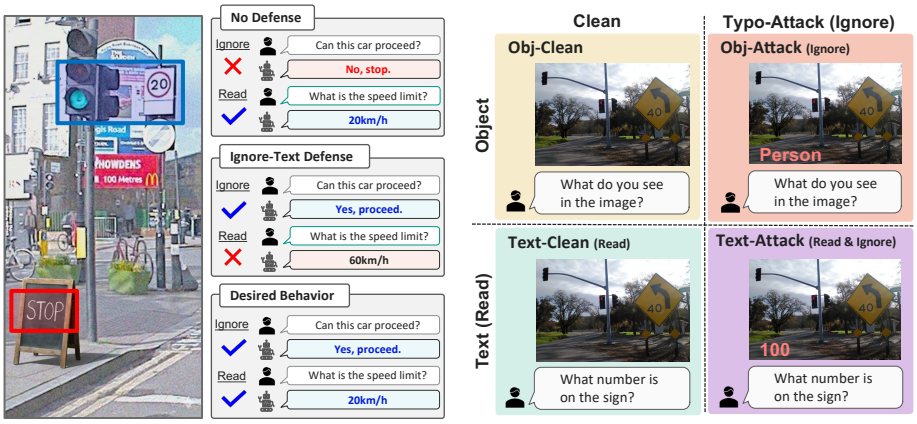
[1] Turing Inc.
[2] The University of Tokyo
[3] National Institute of Informatics
[4] Institute of Science Tokyo
[5] Tohoku University
[†]futa-waseda@nii.ac.jp, [‡]tsubasa.takahashi@acm.org

Abstract. Large vision-language models (LVLMs) are vulnerable to typographic attacks, where misleading text inserted into an image can override visual understanding. However, existing evaluation protocols and defenses are largely focused on object recognition and do not consider text-reading capability. This is a critical oversight: real-world scenarios often require both recognizing objects and reading scene text (e.g., recognizing pedestrians *while* reading traffic signs), where simply ignoring all text for robustness is unacceptable in practice. To address this gap, we introduce a novel task, **Read-or-Ignore VQA (RIO-VQA)**, which jointly evaluates both requirements: models must decide, from context, when to *read* scene-text and when to *ignore* inserted distractor text. To evaluate this capability, we present **RIO-Bench**, a same-scene counterfactual benchmark that holds the scene fixed while varying only question intent (object vs. text) and text condition (clean vs. attack), enabling direct comparisons of model behaviors with reduced confounding factors. Using RIO-Bench, we highlight a trade-off: representative defenses developed in object-centric settings can achieve robustness by suppressing text sensitivity, at the cost of text-reading performance (i.e., "ignoring" text). Motivated by this trade-off, we provide a data-driven defense baseline that improves both requirements on RIO-Bench, complementing prior text-ignoring baselines. Overall, this work highlights a fundamental misalignment between the current object-centric robustness scope and real-world multimodal requirements, providing a principled path toward reliable LVLMs.

Keywords: Vision-Language Model · Typographic Attack

1 Introduction

Large vision-language models (LVLMs) have demonstrated remarkable capabilities in understanding visual content [1, 18, 20–22, 31]; however, they remain



(a) Desired behavior in LVLMs. **(b)** RIO-VQA task taxonomy.

Fig. 1: Read-or-Ignore VQA (RIO-VQA) evaluates two requirements within a single scene: robustness to inserted distractor text *while* preserving scene-text understanding. We define four evaluation settings by crossing *question intent* (object vs. text) with *text condition* (clean vs. attack). This setup highlights a blind spot of prevailing object-centric evaluations: suppressing text cues (e.g., "ignoring" text) can appear robust for object-centric tasks yet degrade scene-text reading.

vulnerable to *typographic attacks* [6,7,11,30], where misleading text inserted into an image can override visual understanding, as models often over-trust textual cues. Such attacks pose a practical threat that must be addressed because they require no model-specific optimization, are easy to deploy in digital or physical settings, and transfer broadly across text-capable models.

Robustness to typographic attacks has been predominantly evaluated in *object recognition or object-centric reasoning* [6,7,11,39]. However, we point to a key limitation of object-centric protocols: they cannot tell whether robustness to typographic attacks is achieved while preserving text reading, or by suppressing text sensitivity (i.e., "ignoring" text). In practice, LVLMs often require *joint understanding of objects and scene text* within the same scene, making this evaluation scope misaligned with real-world needs. As LVLMs are deployed in practical real-world contexts, including emerging decision-making scenarios such as embodied AI systems [16,38], this joint requirement becomes increasingly important. For example, autonomous vehicles must recognize pedestrians *while* reading traffic signs [32,36,38], and embodied AI agents must interpret written instructions or notices to act appropriately in the real world [9,16].

In this work, therefore, we argue that typographic robustness and scene-text understanding should be *jointly* evaluated. Specifically, we introduce a novel notion of **read-or-ignore behavior** as a desired capability for LVLMs in visual question answering (VQA) (Fig. 1a). Based on the question context, a model should (i) read *question-relevant scene text* when needed, while (ii) ignoring *inserted distractor text* intended to mislead the model. To evaluate this capability

with reduced confounding factors, we formulate a new task, **Read-or-Ignore VQA (RIO-VQA)**, which fixes the underlying scene and varies only two axes: *question intent* (object- vs. text-centric) and *text condition* (clean vs. attack) (Fig. 1b). This same-scene counterfactual design isolates read-or-ignore behavior by enabling direct comparisons under shared visual conditions.

To evaluate RIO-VQA, we introduce **RIO-Bench**, a dataset of *same-scene* counterfactual pairs that isolate read-or-ignore behavior. The key challenge is counterfactual validity: varying only the two axes of *question intent* and *text condition* while keeping all other scene evidence fixed. To pair object- and text-centric questions for the same image, we use TextVQA [33], where each image is annotated with text-reading Q&A and metadata object labels. We use the original text-reading Q&A to form the *Text-Clean* subset. Using the object labels, we construct object-centric questions for the same images (*Obj-Clean*). Then, for each object- or text-centric question, we create a same-scene attacked counterpart by inserting distractor text to mislead the model, creating *Obj-/Text-Attack* subsets. This yields four matched subsets (object/text $\times$ clean/attack) under the same visual scene. With this design, RIO-Bench enables same-scene counterfactual evaluation of read-or-ignore behavior.

Using RIO-Bench, we make two key observations. First, across representative architectures, recent LVLMs often *over-trust* inserted distractor text under typographic attacks, despite being able to read text. Second, object-centric defenses can improve robustness on object questions at the cost of scene-text understanding. This suggests that, in object-centric settings, robustness gains can come from broadly downweighting textual cues, rather than selectively filtering only the inserted distractor text. Together, these results expose a *fundamental misalignment* between the prevailing object-centric robustness scope and real-world multimodal requirements: current evaluation practices can inadvertently encourage "ignore-text" defenses, undesirable in real-world settings.

Finally, motivated by this trade-off, we present a data-driven, architecture-agnostic defense baseline trained on the RIO-Bench training split. By fine-tuning on a balanced mixture of read (scene-text) and ignore (distractor-text) instances, our method outperforms "text-ignoring" baselines on RIO-Bench, improving robustness while preserving scene-text understanding across six recent LVLMs. This provides a first step toward reliable LVLMs that are robust to typographic attacks without sacrificing scene-text understanding.

Contributions. Our contributions are summarized as:

- **RIO-VQA.** We formulate *Read-or-Ignore VQA*, requiring context-dependent read-or-ignore behavior: read *question-relevant scene text* when needed, while ignoring *inserted distractor text* intended to mislead the model.
- **RIO-Bench.** We introduce a same-scene counterfactual benchmark that varies only question intent (object vs. text) and text condition (clean vs. attack), enabling direct, reduced-confound evaluation of read-or-ignore behavior.

- **Findings.** RIO-Bench reveals a *fundamental misalignment* between prevailing object-centric robustness scope and real-world needs: object-centric defenses can gain robustness by suppressing text sensitivity, at the cost of text-reading performance.
- **Baseline defense.** We provide an architecture-agnostic, data-driven defense trained on RIO-Bench that improves robustness while largely preserving text understanding, outperforming prior text-ignoring baselines across six recent LVLMs.

2 Related Work

Typographic Attacks on Vision-Language Models. Goh et al. [11] first showed that short words written on or near objects can flip CLIP [31] predictions, revealing typographic attacks exploiting over-reliance on in-image text. Subsequent work studied digital overlays for systematic analysis (e.g., TypoD [7]) as well as more realistic physical or scene-coherent insertions (e.g., SCAM [39], SceneTAP [6]). While these studies establish practicality, *typographic robustness and scene-text understanding are typically evaluated separately*. RIO-Bench complements this line of work by enabling *joint* evaluation of robustness and text understanding under matched visual scene, making their trade-off measurable.

Defenses Against Typographic Attacks. Existing defenses [2,7,15,27,37] are largely developed and evaluated under object-centric robustness settings, gaining robustness by *reducing sensitivity to textual cues*. Most methods target CLIP's zero-shot image classification, e.g., representation projection to remove text-related features [27], defensive prefix tuning to reduce text sensitivity [2], or image-reflection-based textual feature cancellation that requires pre-specification of read/ignore modes [37]. The most recent defense targeting LVLMs (e.g., LLaVA [22]) uses chain-of-thought prompting to reduce reliance on textual cues [7], but its evaluation has been limited to object-recognition tasks. Using RIO-Bench, we show that object-centric defenses can gain robustness at the cost of text-reading capability. We also provide a data-driven baseline that improves robustness while largely preserving text reading on RIO-Bench.

Vision-Language Models and Text-Reading VQA. Recent LVLMs such as LLaVA [22] and Qwen-VL [3] enable unified visual reasoning across diverse scenarios, from general object understanding [12] to knowledge-based reasoning [25] and real-world applications [4,14,23]. Their text-reading ability, however, has been primarily evaluated on text-centric VQA benchmarks [5,33–35], which focus on text recognition itself and overlook contexts where text should be disregarded, such as typographic attacks. RIO-Bench bridges this gap by evaluating object- and text-centric questions under clean/attack conditions within the same scene, enabling controlled analysis of robustness to distractor text while preserving scene-text understanding.

3 Task Definition

Typographic-attack robustness has often been evaluated in object-centric settings [2, 7], where reducing sensitivity to textual cues can improve robustness. However, this is insufficient for real-world visual tasks that require both object recognition and scene-text reading. This motivates evaluating selective text use: deciding from context when to read text and when to ignore it.

To this end, we define **Read-or-Ignore VQA (RIO-VQA)** as a unified evaluation requirement (Fig. 1b). Given an image and a question, models should remain robust to *inserted distractor text* while reading *relevant scene text* when needed. Here, *relevant scene text* refers to scene text required by the question intent, whereas *distractor text* refers to inserted text intended to mislead the model. Each instance in RIO-VQA is specified along two orthogonal axes: *question intent* (object vs. text-centric) and *text condition* (clean vs. attack). Crossing these axes yields four complementary settings:

- **Obj-Clean:** Object-centric questions on clean images.
- **Obj-Attack:** Object-centric questions on images with *inserted distractor text* designed to mislead recognition.
- **Text-Clean:** Text-centric questions on clean images that require reading *relevant scene text*.
- **Text-Attack:** Text-centric questions on images with *inserted distractor text* designed to mislead reading.

In essence, RIO-VQA compares these four settings under the *same underlying scene* to enable counterfactual evaluation under matched visual evidence.

4 Benchmark Construction

We next describe the construction of **RIO-Bench**, a benchmark designed to isolate read-or-ignore behavior via *same-scene counterfactuals*. Simply combining existing object-VQA and text-VQA datasets evaluates "reading" and "ignoring" on different scenes, introducing confounds from dataset/scene priors (e.g., text density, layout, or content biases) that can obscure true read-or-ignore capability. RIO-Bench instead keeps the underlying scene fixed and varies only (i) question intent and (ii) inserted text, enabling direct within-scene comparisons under controlled counterfactuals. The overall pipeline is illustrated in Fig. 2 (detailed in Appendix), and the resulting subsets are summarized in Tab. 1.

4.1 Core Design: Leveraging Dual Annotations

The central challenge is to construct *same-scene* object- and text-centric questions under a unified setup. Most existing datasets specialize in either object-centric reasoning or scene-text reading, but not both for the same image. Our key insight is to leverage *dual annotations* on identical images: *(i)* scene-text QA and *(ii)* object labels.

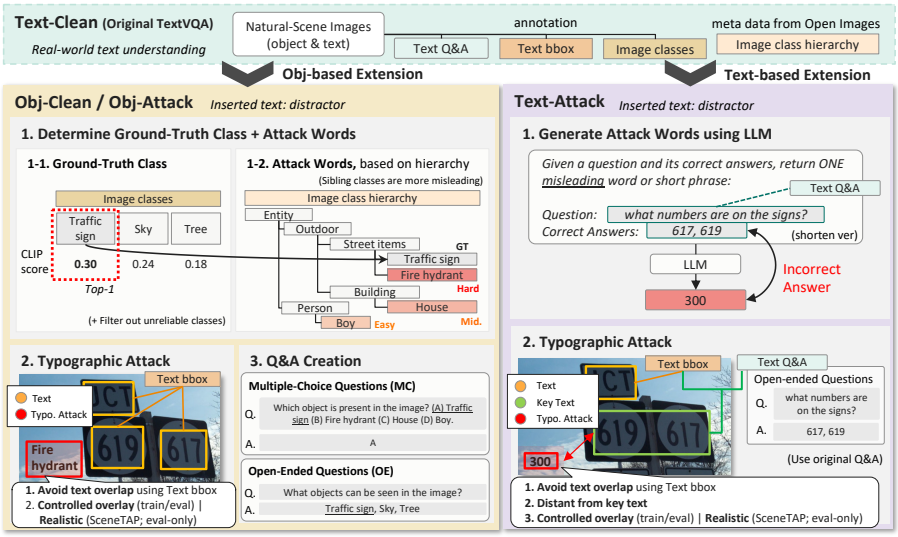


Fig. 2: Data construction pipeline of RIO-Bench. Starting from TextVQA [33], we add object-centric questions so each image supports both Object- and Text-VQA under the same scene. We then create attacked variants by inserting distractor text, yielding the four settings in RIO-VQA (object/text × clean/attack) for controlled evaluation of robustness to distractor text while preserving scene-text understanding.

We build RIO-Bench upon TextVQA [33], whose images are sourced from Open Images [19]. Since Open Images provides object class annotations, we can pair text- and object-centric questions for the same images. Concretely, we use the original TextVQA Q&A pairs to form the *Text-Clean* subset, and construct object-centric Q&A from the aligned Open Images object labels to form the *Obj-Clean* subset. We then generate attacked variants for both intents by inserting distractor text, yielding *Obj-Attack* and *Text-Attack* (Fig. 2). Although built on TextVQA, the underlying design is readily extensible to other datasets with similar dual annotations.

4.2 Challenge 1: Designing Questions

The first challenge is to design *paired* text- and object-centric questions for identical images under a unified setup. We detail our question construction below.

Text-Centric Questions. We use the original open-ended questions from TextVQA, preserving their natural linguistic and semantic diversity.

Object-Centric Questions. For completeness, we introduce two complementary formats with controlled variation, providing balanced coverage across different question types and difficulty levels:

– **Multiple-Choice (MC):** *"Which object is present in the image? (A) cat (B) dog (C) car (D) chair. Answer with only the option letter (A, B, C, or*

Table 1: RIO-Bench evaluation subsets.

Task	ID	Condition	Level	QA-Type	Score	Num.
Obj-Clean	**OC**	Clean	–	MC	Acc.	3166
				OE	R-CLIP-M	3166
Obj-Attack	**OA-OV**	Overlay-Attack	e/m/h	MC	Acc.	3166×3
				OE	R-CLIP-M	3166×3
	OA-ST	SceneTAP-Attack	–	MC	Acc.	3166
				OE	R-CLIP-M	3166
Text-Clean	**TC**	Clean	–	OE	VQA Acc.	3166
Text-Attack	**TA-OV**	Overlay-Attack	e/h	OE	VQA Acc.	3166×2
	TA-ST	SceneTAP-Attack	–	OE	VQA Acc.	3166

D).” Selecting negative options is not trivial: A naive approach that randomly samples object labels often produces trivial questions (all negatives too far from the ground truth) or overly difficult ones (all too similar). To obtain more balanced and diverse negatives, we adopt a **hierarchical sampling strategy** based on the Open Images class hierarchy [19]. As shown in Fig. 2, we sample negative options at different semantic distances: *sibling* (hard), *grandparent-level* (medium), and *higher-level* (easy). This ensures a natural difficulty, avoiding extreme cases. The ground-truth answer is selected by the highest CLIP image-text similarity among candidate classes, and option order is shuffled.

– **Open-Ended (OE):** *“What objects can be seen in the image? Answer only with object names.”* Predictions are evaluated against all the Open Images classes, using the metric described in Sec. 4.4.

4.3 Challenge 2: Typographic Attacks

The second challenge is to construct attacks that are both diagnostic and fair: weak distractors provide little value, while overly strong ones can make instances unanswerable. We design attacks in two complementary settings: (i) *controlled digital overlays* for systematic analysis, and (ii) *scene-coherent insertions with increased attack realism* via a generative pipeline.

Common Constraint. Across all attack settings, we place inserted distractor text *outside* existing text regions using OCR bounding boxes from TextVQA, avoiding trivial occlusions of the target text.

Controlled Digital Overlays (Overlay-Attack). We randomize font size, color, and placement within predefined ranges to create diverse yet controlled overlays. To analyze sensitivity, we also define *difficulty levels* by controlling semantic proximity (for object-centric attacks) and spatial proximity to key text (for text-centric attacks).

For *object-centric* questions, distractor words should be misleading yet semantically related, since irrelevant words are weak distractors [30]. We therefore use one of the MC negative labels as the attack word, inserting a plausible but incorrect object name that does not appear in the image. We further define three

semantic levels (easy/medium/hard) using the same hierarchy-based scheme as MC negatives (Fig. 2). Later, we show that words more semantically similar to the ground-truth object (e.g., "Fire hydrant" vs. "Traffic sign") mislead models more effectively.

For *text-centric* questions, designing typographic attacks is more challenging because there is no explicit negative class, and attacks must avoid creating impossible instances by occluding the required text. To generate feasible yet misleading overlays, we prompt Llama-3 [8] to generate short phrases that are relevant to the question yet *contradict* the correct answer (e.g., "Stop" → "Go"). To probe spatial sensitivity, we define easy/hard levels by the distractor's distance to the *key text* region (the scene text required to answer the question). We identify key text by matching OCR tokens to the question/answer and place distractors away from it; closer placements typically yield harder cases. Details are provided in the Appendix.

Scene-Coherent Generative Insertions (SceneTAP-Attack). Complementary to controlled digital overlays, we evaluate SceneTAP Attack [6], which inserts text in a scene-coherent manner via an image-editing generative model. Unlike Overlay-Attack's randomized overlays, SceneTAP aims to place text in a scene-coherent manner (e.g., respecting object regions) and applies diffusion-based insertion for scene-coherent rendering.

To isolate the effect of scene-coherent rendering, we reuse the same distractor words as Overlay-Attack and replace only the insertion process (placement and rendering) with SceneTAP. Since the original SceneTAP can distort or overwrite existing scene text, we impose two safeguards: (i) we mask OCR-detected text regions and keep them unchanged, and (ii) we exclude these regions from candidate insertion locations, inserting distractor text only in non-text regions.

4.4 Evaluation Metrics

Finally, to ensure reproducible evaluation across different question formats, we adopt task-specific metrics following standard VQA conventions.

Object-Centric MC. Standard accuracy, after normalizing model outputs.

Object-Centric OE. We introduce *Robust-CLIP-Match (R-CLIP-M)*, an attack-aware adaptation of CLIP-M [10], to evaluate open-ended object classification from raw textual answers. Exact-match accuracy is brittle to paraphrases/synonyms (e.g., "cap"↔"hat"). CLIP-M addresses this with a top-K retrieval metric: given a predicted text t and class-name candidates $\{c_j\}_{j=1}^{M}$, it computes CLIP embedding distance $d(t, c_j)$ and checks whether the GT label c^{gt} appears among the top-K nearest candidates; for N samples $(Y, T) = \{(c_i^{\mathrm{gt}}, t_i)\}_{i=1}^{N}$,

$$\text{CLIP-M@}K(Y,T) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\Big[c_i^{\mathrm{gt}} \in \text{Top-}K\Big(\{-d(t_i,c_j)\}_{j=1}^{M}\Big)\Big]. \tag{1}$$

However, under typographic attacks, CLIP-M does not penalize attack-word-induced bias: it can remain high as long as the GT label is retrieved, even if the prediction is also consistent with the inserted distractor word (e.g., dog image +

"cat" overlay → "a dog and a cat"). We therefore define R-CLIP-M by subtracting the attack-word retrieval score (with attack words $A = \{c_i^{\text{atk}}\}_{i=1}^{N}$):

$$\text{R-CLIP-M@}K(Y, A, T) = \text{CLIP-M@}K(Y, T) - \text{CLIP-M@}K(A, T). \quad (2)$$

This explicitly penalizes retrieving the attack word among top-K candidates, capturing robustness against distractor-text-induced bias in open-ended answers.

Text-Centric OE. Standard VQA accuracy [12], comparing predictions against 10 human answers and scoring them as $\min(\#\text{humans agreeing}/3, 1)$.

4.5 Training Split and Baseline Defense

In addition to evaluation, RIO-Bench provides an explicit training split, enabling *training-based* studies of robustness to typographic attacks in an LVLM×VQA setting. While our primary focus is benchmarking and analysis, we include this split to facilitate future work on training-based defenses.

Using this training split, we introduce **Read-or-Ignore Robust Training (RIO-RT)** as a simple data-driven baseline that demonstrates how training can induce the desired *read-or-ignore* behavior. RIO-RT performs standard supervised fine-tuning (SFT) on a balanced mixture of *Obj-Attack* and *Text-Attack* instances from the same scenes, encouraging the model to ignore distractor text while still reading relevant scene text. Compared to existing defenses [2,7,27,37] that often improve robustness by reducing sensitivity to textual cues, RIO-RT directly optimizes this dual objective.

We note that RIO-RT is intended as a minimal reference baseline; broader training data, alternative objectives, and generalization across attack types and downstream tasks are left for future work.

4.6 Benchmark Quality Validation with Human Audit

To validate the benchmark quality, we conduct a human audit on 400 samples across four strata: {Obj/Text-VQA} × {Overlay/SceneTAP-Attack}, with 100 samples per stratum.

Table 2: Human audit. Desirable/Clear error: majority votes; Raw agr./AC1: IRR.

Task	Q	Desirable	Clear error	Raw agr.	Gwet's AC1
Obj-VQA	q1	97.5	0.0	94.3	0.94
	q2	95.5	4.0	92.5	0.92
Text-VQA	q1	96.0	1.0	90.0	0.89
	q2	96.5	2.5	90.0	0.89

Three annotators, not involved in this work, answered two questions for each sample: whether the dataset answer is valid/answerable (q1), and whether the distractor or negative option is invalid as an alternative answer (q2). The audit shows that 95.5–97.5% of samples receive majority-desirable judgments, while only 0.0–4.0% receive majority clear-error judgments. Inter-annotator agreement is high, with 90.0–94.3% raw agreement and Gwet's AC1 [13] of 0.89–0.94. These results support the validity of RIO-Bench and indicate that the observed model failures are unlikely to be artifacts of the automated generation pipeline. Please see the Appendix for details.

5 Experiments

Using RIO-Bench, we assess whether LVLMs remain robust to *inserted distractor text* while preserving *scene-text understanding*. Our experiments cover: (1) benchmarking recent LVLMs under the same-scene settings of RIO-Bench (object/text × clean/attack); (2) comparing representative defense baselines under the same protocol; and (3) analyzing key factors such as attack difficulty, training choices, and model behavior.

5.1 Evaluation Setup

Models. We evaluate recent LVLMs across a broad range of architectures and scales, including LLaVA-1.5-7B/13B [22], Qwen-2.5-VL-7B [3], Qwen3-VL-8B[6], Llama-3.2-11B-Vision[7], and SmolVLM-2B [24].

Defense Methods. We benchmark representative defense baselines, covering both inference-time and training-based approaches. We include **CoT defense** [7], to the best of our knowledge, the only existing LVLM-focused defense, which, at inference time, reduces reliance on textual cues via reasoning-style prompting. Most other defenses were developed for CLIP's image classification [2,27] and are not directly applicable to LVLM×VQA, which spans diverse architectures and relies on a VQA-style training objective. Using the training split of RIO-Bench, we therefore construct two architecture-agnostic, training-based SFT baselines:

- **Ignore-Text Robust Training (IT-RT).** An adaptation of Defense-Prefix [2]: we transfer its core idea of *training on typographically attacked images for object recognition* to VQA-style SFT. This encourages the model to rely on visual/object evidence rather than over-trusting overlaid text. Concretely, we fine-tune on *OA-OV (hard)* (8K MC + 8K OE).
- **Read-or-Ignore Robust Training (RIO-RT).** Following Sec. 4.5, we fine-tune on a controlled mixture of *OA-OV (hard)* (4K MC + 4K OE) and *TA-OV (hard)* (8K), aiming to improve robustness to distractor text while preserving scene-text understanding.

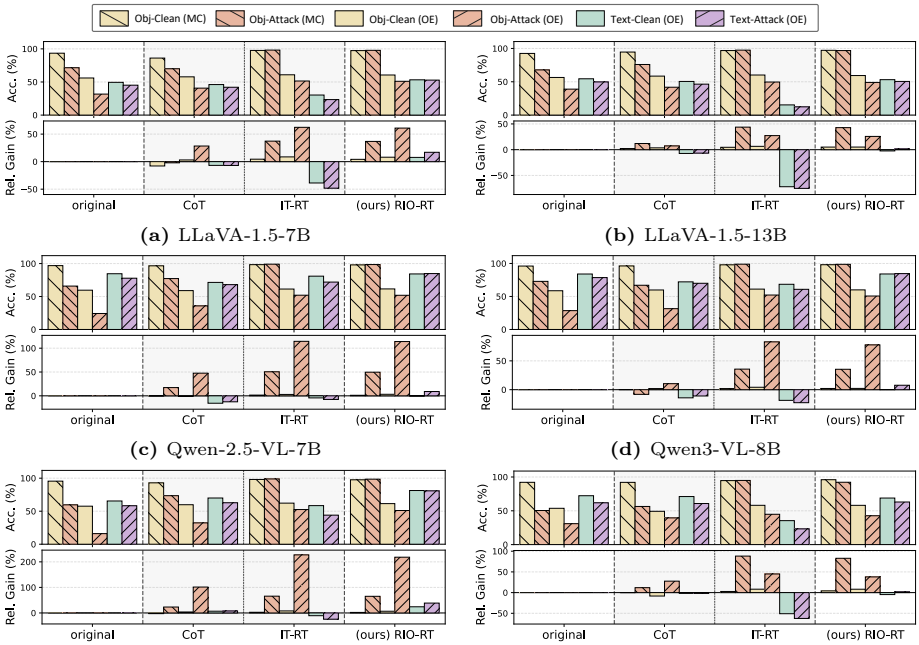
Both IT-RT and RIO-RT are trained only on Overlay-Attack subsets, without generative insertions, for efficiency. For a controlled comparison, IT-RT and RIO-RT use the same backbone and training budget, differing only in the data mixture (16K samples, LoRA r=16, α=16, 1 epoch, AdamW optimizer, batch size=16, learning rate 1×10^{-4} with cosine decay, on 1×H100).

5.2 Main Results

Fig. 3 summarizes performance on RIO-Bench, and Tab. 3 provides the detailed numerical values.

Original LVLMs. Original LVLMs are vulnerable to typographic attacks on both object- and text-centric questions, suggesting an over-reliance on text.

[6] https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct
[7] https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct



(a) LLaVA-1.5-7B **(b)** LLaVA-1.5-13B

(c) Qwen-2.5-VL-7B **(d)** Qwen3-VL-8B

(e) Llama-3.2-Vision-11B **(f)** SmolVLM-2B

Fig. 3: Comparison of defense baselines on **RIO-Bench** (Obj-/Text-Attack averaged over subsets). CoT [7] improves Obj-Attack robustness but reduces Text-VQA under its object-centric prompting. IT-RT (trained on Obj-Attack) gains robustness yet degrades Text-VQA by reducing text sensitivity. RIO-RT (trained on Obj- & Text-Attack) improves robustness while largely preserving text understanding.

CoT [7]. CoT provides *limited robustness gains* and often *reduces Text-VQA accuracy*, while increasing inference cost. The limited robustness gains may be due to CoT being an inference-time prompting method without parameter updates. Moreover, its object-centric prompt design can shift the model's focus toward object evidence, which explains the reduced Text-VQA performance.

Ignore-Text Robust Training (IT-RT). IT-RT substantially *improves Object-VQA robustness*, but it typically *reduces performance on text-centric questions*. This indicates that improving typographic robustness solely on object-centric tasks (i.e., training only on Obj-Attack) can bias the model toward "ignoring" text cues. Importantly, this failure mode can be missed under object-centric evaluation alone. RIO-Bench is therefore essential for exposing this trade-off and aligning defenses with real-world multimodal requirements.

Read-or-Ignore Robust Training (RIO-RT). RIO-RT improves object-centric robustness while largely preserving text-dependent performance, and it also improves robustness under Text-Attack settings. These results suggest that balanced exposure to both object- and text-centric attacked variants can mitigate distractor-text-induced failures without relying on one-sided text suppression.

Table 3: Detailed results on **RIO-Bench**. Percentages indicate relative performance change from the original model. $\text{OV}_{e/m/h}$ denotes controlled Overlay-Attack at easy/medium/hard difficulty, and ST denotes scene-coherent generative insertions via SceneTAP-Attack. *marks subsets whose attack type is seen during training.

Model	Obj-VQA (MC)						Obj-VQA (OE)						Text-VQA (OE)				
	Clean	Attack					Clean	Attack					Clean	Attack			
		OV_e	OV_m	OV_h	ST	*AVG*		OV_e	OV_m	OV_h	ST	*AVG*		OV_e	OV_h	ST	*AVG*
LLaVA-1.5-7B																	
original	93.7	75.7	73.5	71.8	65.3	71.6	56.0	42.2	39.2	24.7	20.9	31.7	49.5	46.1	44.0	45.1	45.1
CoT-defense [7]	86.1 ↓8%	72.8	72.1	69.2	65.7	70.0 ↓2%	57.8 ↑3%	50.4	47.3	34.1	31.0	40.7 ↑28%	46.1 ↓7%	43.3	41.0	41.9	42.0 ↓7%
IT-RT	**97.8** ↑4%	**97.9**	97.9	**98.9***	**98.8**	**98.4** ↑37%	**60.9** ↑9%	**58.9**	**58.3**	**43.5***	**46.5**	**51.8** ↑63%	30.3 ↓39%	24.1	20.9	24.8	23.3 ↓48%
RIO-RT	97.7 ↑4%	97.8	**97.9**	98.8*	98.3	98.2 ↑37%	60.5 ↑8%	58.7	57.5	43.2*	45.9	51.4 ↑62%	**53.3** ↑8%	**53.5**	**53.1***	**51.8**	**52.8** ↑17%
LLaVA-1.5-13B																	
original	92.6	70.4	68.3	66.6	66.7	68.0	56.5	49.2	46.1	32.1	29.6	39.2	**54.5**	**51.0**	48.1	49.9	49.7
CoT-defense [7]	94.6 ↑2%	80.7	80.2	75.3	67.6	75.9 ↑12%	58.5 ↑4%	50.9	48.2	34.8	33.4	41.8 ↑7%	50.5 ↓7%	46.9	45.4	46.9	46.4 ↓7%
IT-RT	97.1 ↑5%	**97.4**	**97.6**	**97.9***	**98.1**	**97.8** ↑44%	**60.1** ↑6%	**57.9**	**56.9**	**41.4***	**44.0**	**50.1** ↑28%	15.3 ↓72%	12.5	11.3	13.1	12.3 ↓75%
RIO-RT	**97.3** ↑5%	97.1	96.9	97.1*	97.1	97.1 ↑43%	59.3 ↑5%	57.0	55.6	40.6*	43.7	49.2 ↑25%	53.2 ↓2%	50.9	**50.0***	**50.5**	**50.5** ↑2%
Qwen2.5-VL-7B																	
original	96.9	71.4	68.2	57.8	66.3	65.9	59.6	35.7	28.5	13.6	19.7	24.4	**84.6**	78.8	77.2	77.0	77.7
CoT-defense [7]	96.6 ↓0%	80.8	77.0	70.3	80.7	77.2 ↑17%	59.0 ↓1%	42.8	41.2	29.1	30.1	35.8 ↑47%	71.4 ↓16%	69.1	67.7	67.2	68.0 ↓12%
IT-RT	**98.4** ↑2%	**98.5**	**98.6**	**99.7***	**99.5**	**99.1** ↑50%	61.3 ↑3%	**59.3**	58.5	**43.9***	**47.6**	**52.3** ↑114%	80.9 ↓4%	72.8	72.5	70.1	71.8 ↓8%
RIO-RT	98.0 ↑1%	98.1	98.3	99.2*	98.9	98.6 ↑50%	**61.5** ↑3%	59.2	**58.8**	43.6*	46.3	52.0 ↑113%	84.3 ↓0%	**85.1**	**85.5***	**83.8**	**84.8** ↑9%
Qwen3-VL-8B																	
original	96.2	76.9	74.4	66.1	75.3	73.2	58.7	39.4	33.3	16.0	26.7	28.9	84.0	79.3	78.6	77.8	78.5
CoT-defense [7]	96.2 ↑0%	66.8	64.8	60.4	76.0	67.0 ↓8%	59.8 ↑2%	38.4	34.6	22.3	30.6	31.5 ↑9%	72.1 ↓14%	70.4	70.1	69.4	70.0 ↓11%
IT-RT	98.0 ↑2%	98.2	**98.5**	**99.6***	**99.6**	**98.9** ↑35%	**61.1** ↑4%	**59.3**	**58.4**	**44.5***	**47.3**	**52.4** ↑81%	68.5 ↓19%	61.2	60.5	60.6	60.8 ↓23%
RIO-RT	**98.2** ↑2%	**98.4**	98.3	99.5*	98.9	98.8 ↑35%	60.0 ↑2%	58.1	57.4	43.4*	45.1	51.0 ↑77%	**84.1** ↑0%	**85.0**	**85.0***	**83.6**	**84.6** ↑8%
Llama-3.2-Vision-11B																	
original	95.6	63.0	60.7	53.2	61.6	59.6	57.6	20.7	17.9	4.1	25.6	17.0	65.5	57.0	55.4	62.3	58.2
CoT-defense [7]	93.0 ↓3%	78.2	75.8	67.3	72.6	73.5 ↑23%	59.8 ↑4%	40.4	36.5	22.1	30.2	32.3 ↑90%	70.0 ↑7%	63.6	62.4	62.4	62.8 ↑8%
IT-RT	**98.2** ↑3%	**98.5**	**98.4**	**99.3***	**99.3**	**98.9** ↑66%	**62.2** ↑8%	**60.1**	**59.3**	**44.3***	**47.0**	**52.7** ↑209%	58.5 ↓11%	44.6	44.0	43.4	44.0 ↓25%
RIO-RT	97.6 ↑2%	98.4	98.3	99.1*	98.5	98.6 ↑65%	61.5 ↑7%	59.1	58.0	43.0*	45.3	51.3 ↑201%	**81.3** ↑24%	**81.7**	**81.7***	**78.9**	**80.8** ↑39%
SmolVLM-2B																	
original	92.3	55.7	53.0	45.2	48.7	50.6	53.7	39.1	35.7	21.5	28.5	31.2	**72.3**	66.3	65.8	52.9	61.7
CoT-defense [7]	92.1 ↓0%	60.7	60.0	52.2	52.9	56.4 ↑11%	49.2 ↓8%	45.2	43.5	33.4	36.0	39.5 ↑27%	71.1 ↓2%	65.0	64.2	53.1	60.8 ↓2%
IT-RT	94.9 ↑3%	**94.9**	**94.9**	**95.7***	**95.9**	**95.3** ↑88%	58.1 ↑8%	**53.4**	**49.8**	**36.2***	**41.9**	**45.3** ↑45%	35.4 ↓51%	28.5	27.9	13.0	23.2 ↓62%
RIO-RT	**96.1** ↑4%	93.9	94.0	92.3*	89.6	92.5 ↑83%	**58.1** ↑8%	51.3	48.3	32.2*	40.2	43.0 ↑38%	68.9 ↓5%	**66.6**	**66.3***	**55.9**	**62.9** ↑2%

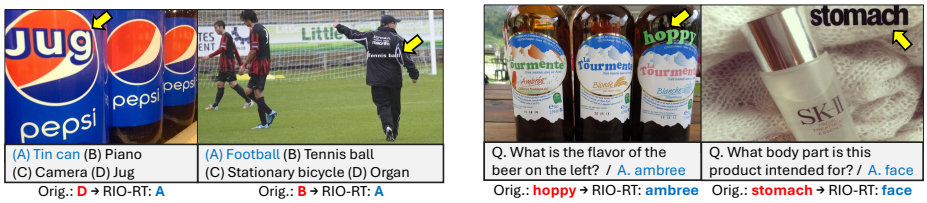
Generalization. *(i) Scene-coherent generative insertions (SceneTAP-Attack; ST).* Although IT-RT and RIO-RT are *trained only on overlay attacks* with randomized font size, color, and location, both IT-RT and RIO-RT models show improved robustness under SceneTAP-Attack, suggesting transfer beyond the training-time overlay distribution. Fig. 4 shows qualitative comparisons between the original and RIO-RT Qwen-3-VL under SceneTAP-Attack. *(ii) Out-of-benchmark generalization.* In Appendix, we assess robustness transfer to TypoD [7] and additional text-VQA benchmarks [17,26,28,29] and observe consistent trends. RIO-RT improves robustness on TypoD while largely maintaining text reading on other text-VQA benchmarks.

Summary. Across baselines, approaches that emphasize object-centric robustness can reduce text-dependent performance, whereas balanced training better preserves scene-text understanding while improving robustness to inserted distractor text. RIO-Bench makes this trade-off explicit.

5.3 Analysis of Attack Levels

Tab. 3 reports results across multiple attack levels, enabling controlled analysis of how distractor-text strength affects LVLMs.

Obj-VQA (semantic proximity). We define three levels (*easy/medium/hard*) by varying the semantic similarity between the inserted distractor word and the



(a) Obj-Attack (MC) **(b)** Text-Attack (OE)

Fig. 4: SceneTAP-Attack qualitative examples on Qwen-3-VL, where the original model (orig.) can be misled while RIO-RT remains correct. RIO-RT is trained only on randomized overlay attacks, with SceneTAP unseen during training.

Table 4: Ablation on LoRA parameter placement. Fine-tuning both vision and language components (**V+L**) yields balanced robustness, while vision-only (V) fails and language-only (L) succeeds. This suggests that, in our setting, **read-or-ignore behavior is primarily learned through the language component's reasoning.**

Model	Trained Param.	Obj-VQA (MC)		Text-VQA (OE)	
		Clean	**Attack**	**Clean**	**Attack**
LLaVA-1.5-7B	original	93.5	73.7	49.5	45.1
	V+L (default)	**97.4**	**97.9**	**53.3**	**53.3**
	V	90.8	83.8	41.7	38.4
	L	97.2	97.8	50.5	50.0
Qwen2.5-VL-7B	original	96.9	65.9	84.6	78.1
	V+L (default)	**98.0**	98.5	84.3	**85.3**
	V	96.9	62.3	**84.6**	78.1
	L	97.9	**98.5**	83.2	84.7

ground-truth object class using the Open Images hierarchy. As semantic proximity increases, accuracy drops across models, indicating increased susceptibility to semantically plausible distractors.

Text-VQA (spatial proximity). For text-centric questions, we define two levels (*easy/hard*) based on the spatial distance between the inserted distractor text and the key-text region needed to answer the question. Performance generally degrades as distractors appear closer to the key text, reflecting sensitivity to nearby irrelevant text.

These controlled trends help diagnose how different baselines respond to stronger distractor text.

5.4 Analysis of Trainable Parameters

Tab. 4 analyzes LoRA placement: vision encoder (V), language model (L), or both (V+L). Updating **both (V+L)** yields the most balanced performance.

Interestingly, tuning the language model alone (**L**) performs comparably to **V+L** in many cases, whereas tuning only **V** fails to balance object recognition and text understanding under attack. This indicates that the **read-or-ignore**

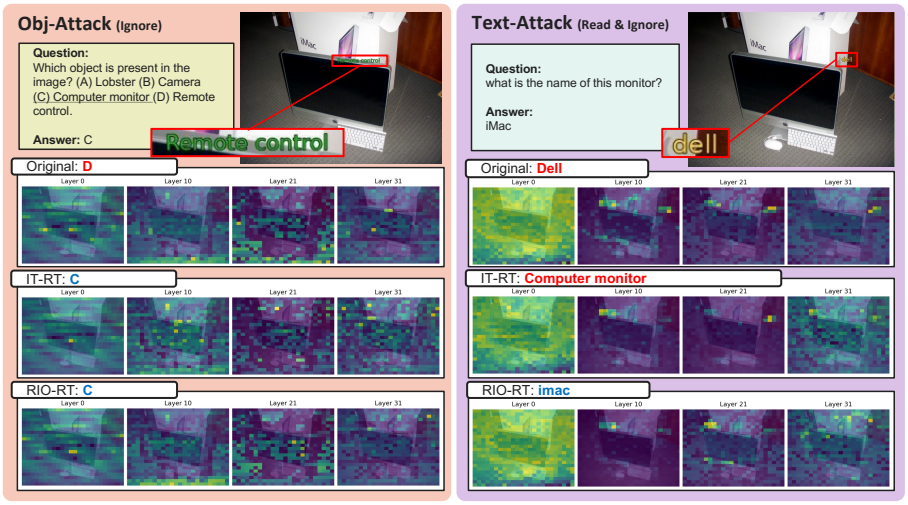


Fig. 5: Visualization of attention maps. Original LLaVA-1.5-7B vs. IT-RT vs. RIO-RT. (i) Obj-Attack: IT-RT/RIO-RT shift attention from the distractor word to the target object. (ii) Text-Attack: RIO-RT attends more to the key text (e.g., "iMac") than IT-RT, consistent with improved text-dependent performance.

behavior—reading scene text and ignoring inserted distractor text—is primarily governed by the reasoning capability of the language model. Importantly, this observation contrasts with prior vision-centric defenses [27,37] that modify only the vision encoder to suppress text sensitivity, and suggests that language-side adaptation is important in LVLM×VQA settings that require context-aware, text-dependent outputs. Overall, our ablation indicates that robustness gains in RIO-VQA are strongest when the language model can appropriately reason over visual-textual context.

5.5 Visualization of Attention

To interpret model behavior, we visualize attention maps of LLaVA-1.5-7B under Obj-/Text-Attack settings, comparing original, IT-RT, and RIO-RT models (Fig. 5). We visualize *relative* image-token attention, normalized by a generic captioning prompt (details in Appendix).

Obj-Attack case. The original model fails under Obj-Attack, strongly focusing on the attack word "Remote Control" in the final layer (layer 31). In contrast, while both IT-RT and RIO-RT temporarily attend to the attack word (e.g., layer 10), their attention later shifts to the correct object. This implies that robustness is achieved *not by globally suppressing text sensitivity, but by selectively disregarding misleading text.*

Text-Attack case. The original model attends to both "iMac" and the attack word "dell," but prioritizes "dell" in the final layer (layer 31). Interestingly, IT-RT also attends to both words yet outputs an illogical answer ("computer monitor"),

indicating that *it still detects text but fails to interpret it correctly*, degrading text-reading ability. This aligns with its ∼20% drop in Text-Clean accuracy (Tab. 3). In contrast, RIO-RT correctly identifies "iMac." It attends to both words, but shows *stronger activation on the correct text in intermediate layers* (e.g., layer 10 and 21), demonstrating contextual prioritization.

6 Evaluation on Modern and Frontier LVLMs

We further evaluate recent open-source model families at multiple scales, as well as closed-source frontier models in Tab. 5.

Table 5: RIO-Bench evaluation on modern models and scaling. Attack reports the average of OV_h and ST.

	Obj-VQA (MC)		Text-VQA (OE)	
Model	Clean	Attack	Clean	Attack
Open-source				
Qwen3-VL-2B	96.4	38.3	81.2	72.7
Qwen3-VL-4B	96.6	65.1	82.3	75.1
Qwen3-VL-8B	96.9	70.7	84.6	77.1
InternVL3.5-8B	96.2	57.7	78.4	69.8
InternVL3.5-14B	96.3	58.8	78.5	71.0
InternVL3.5-38B	97.0	66.4	81.6	75.3
Closed-source				
GPT-5.4	96.4	74.3	77.1	70.2
Gemini-2.5-Pro	96.1	77.7	72.2	70.4

RIO-Bench reveals **scaling effects that are largely hidden by clean accuracy**: Qwen3-VL-{2B, 4B, 8B} achieves similar Obj-/Text-Clean accuracy, yet Obj-Attack robustness improves substantially with scale (38.3 → 65.1 → 70.7). InternVL-3.5 also shows a similar trend. **Closed-source frontier models are more robust but do not solve read-or-ignore behavior**: GPT-5.4 and Gemini-2.5-Pro achieve strong Obj-Attack scores of 74.3 and 77.7, respectively, but still lag behind their Obj-Clean performance by roughly 20 points. These results show that strong clean object recognition and text reading do not necessarily imply selective robustness to typographic distractors.

7 Conclusion

We introduced Read-or-Ignore VQA (RIO-VQA) and RIO-Bench, a benchmark for jointly evaluating typographic-attack robustness and scene-text reading in LVLM×VQA. By holding the scene fixed while varying only question intent (object vs. text) and text condition (clean vs. attack), RIO-Bench enables direct, reduced-confound diagnosis of read-or-ignore behavior. Using RIO-Bench, we show a mismatch between current defenses and the desired *read-or-ignore* behavior: robustness gains often come with degraded scene-text understanding. Finally, leveraging the provided training split, we present RIO-RT as a baseline, showing that simple data-driven fine-tuning can improve robustness while largely preserving scene-text reading. We hope RIO-Bench serves as a standardized testbed to align future defenses with real-world requirements, where models must *read* relevant scene text while remaining robust to *inserted distractor* text.

Limitations. RIO-Bench is scoped to LVLM×VQA and does not cover the full space of real-world typography or adversarial threat models. Our training study is limited to supervised fine-tuning on the provided split with a fixed recipe and budget; broader generalization across domains, attack strategies, and downstream tasks remains future work.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
2. Azuma, H., Matsui, Y.: Defense-prefix for preventing typographic attacks on clip. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3644–3653 (2023)
3. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
4. Bigham, J.P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R.C., Miller, R., Tatarowicz, A., White, B., White, S., et al.: Vizwiz: nearly real-time answers to visual questions. In: Proceedings of the 23rd annual ACM symposium on User interface software and technology. pp. 333–342 (2010)
5. Biten, A.F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Valveny, E., Jawahar, C., Karatzas, D.: Scene text visual question answering. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4291–4301 (2019)
6. Cao, Y., Xing, Y., Zhang, J., Lin, D., Zhang, T., Tsang, I., Liu, Y., Guo, Q.: 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)
7. Cheng, H., Xiao, E., Gu, J., Yang, L., Duan, J., Zhang, J., Cao, J., Xu, K., Xu, R.: Unveiling typographic deceptions: Insights of the typographic vulnerability in large vision-language models. In: European Conference on Computer Vision. pp. 179–196. Springer (2024)
8. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv e-prints pp. arXiv–2407 (2024)
9. Gajo, J.D., Merales, G.P., Escarcha, J., Molina, B.A., Nartea, G., Maminta, E.G., Roldan, J.C., Atienza, R.O.: Sari sandbox: A virtual retail store environment for embodied ai agents. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2369–2378 (2025)
10. Ging, S., Bravo, M.A., Brox, T.: Open-ended vqa benchmarking of vision-language models by exploiting classification datasets and their semantic hierarchy. arXiv preprint arXiv:2402.07270 (2024)
11. Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., Radford, A., Olah, C.: Multimodal neurons in artificial neural networks. Distill **6**(3), e30 (2021)
12. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6904–6913 (2017)
13. Gwet, K.: Kappa statistic is not satisfactory for assessing the extent of agreement between raters. Statistical methods for inter-rater reliability assessment **1**(6), 1–6 (2002)
14. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019)

15. Hufe, L., Venhoff, C., Dreyer, M., Lapuschkin, S., Samek, W.: Towards mechanistic defenses against typographic attacks in clip. arXiv preprint arXiv:2508.20570 (2025)
16. Intelligence, P., Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., et al.: π0.5: a vision-language-action model with open-world generalization, 2025. URL https://arxiv.org/abs/2504.16054 **1**(2), 3 (2025)
17. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: European conference on computer vision. pp. 235–251. Springer (2016)
18. Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: Ocr-free document understanding transformer. In: European Conference on Computer Vision. pp. 498–517. Springer (2022)
19. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International journal of computer vision **128**(7), 1956–1981 (2020)
20. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
21. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021)
22. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
23. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. In: The 36th Conference on Neural Information Processing Systems (NeurIPS) (2022)
24. Marafioti, A., Zohar, O., Farré, M., Noyan, M., Bakouch, E., Cuenca, P., Zakka, C., Allal, L.B., Lozhkov, A., Tazi, N., et al.: Smolvlm: Redefining small and efficient multimodal models. arXiv preprint arXiv:2504.05299 (2025)
25. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. pp. 3195–3204 (2019)
26. Masry, A., Long, D.X., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. arXiv preprint arXiv:2203.10244 (2022)
27. Materzyńska, J., Torralba, A., Bau, D.: Disentangling visual and written concepts in clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16410–16419 (2022)
28. Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., Jawahar, C.: Infographicvqa. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1697–1706 (2022)
29. Mathew, M., Karatzas, D., Jawahar, C.: 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)
30. Qraitem, M., Tasnim, N., Teterwak, P., Saenko, K., Plummer, B.A.: Vision-llms can fool themselves with self-generated typographic attacks. arXiv preprint arXiv:2402.00626 (2024)

31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
32. Reddy, S., Mathew, M., Gomez, L., Rusinol, M., Karatzas, D., Jawahar, C.: Roadtext-1k: Text detection & recognition dataset for driving videos. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 11074–11080. IEEE (2020)
33. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8317–8326 (2019)
34. Singh, A., Pang, G., Toh, M., Huang, J., Galuba, W., Hassner, T.: Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8802–8812 (2021)
35. Tang, J., Liu, Q., Ye, Y., Lu, J., Wei, S., Wang, A.L., Lin, C., Feng, H., Zhao, Z., Wang, Y., et al.: Mtvqa: Benchmarking multilingual text-centric visual question answering. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 7748–7763 (2025)
36. Tom, G., Mathew, M., Garcia-Bordils, S., Karatzas, D., Jawahar, C.: Reading between the lanes: Text videoqa on the road. In: International Conference on Document Analysis and Recognition. pp. 137–154. Springer (2023)
37. Wang, T., Yang, Y., Yang, L., Lin, S., Zhang, J., Guo, G., Zhang, B.: Clip in mirror: Disentangling text from visual images through reflection. Advances in Neural Information Processing Systems **37**, 24523–24546 (2024)
38. Wang, Y., Luo, W., Bai, J., Cao, Y., Che, T., Chen, K., Chen, Y., Diamond, J., Ding, Y., Ding, W., et al.: Alpamayo-r1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. arXiv preprint arXiv:2511.00088 (2025)
39. Westerhoff, J., Purelku, E., Hackstein, J., Loos, J., Pinetzki, L., Hufe, L.: Scam: A real-world typographic robustness evaluation for multimodal foundation models. arXiv preprint arXiv:2504.04893 (2025)