DEFENDING VLMS AGAINST TYPOGRAPHIC ATTACKS VIA QUERY-IRRELEVANT SEMANTIC EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Typographic attacks pose a critical threat to vision-language models (VLMs) by injecting misleading text into images and causing models to rely on adversarial textual cues rather than visual evidence. Existing defenses often require model-specific modifications, additional training, or access to internal model components, limiting their applicability to modern closed-source VLMs. In this paper, we propose QuISE, a model-agnostic, training-free black-box defense based on query-irrelevant semantic editing. QuISE first identifies text regions likely to affect the current query through influence-aware text localization. QuISE then replaces these regions with two semantically distinct replacement texts that are irrelevant to both the query and the image. The final answer is determined by answer consistency across the edited images. Extensive experiments on three typographic-attack benchmarks, four attack settings, and four VLMs show that QuISE consistently improves defended accuracy. QuISE achieves a recovery rate of 67.9–75.0% with a harm rate of 0.5–1.1%.

1 INTRODUCTION

Vision-language models (VLMs) use text embedded in images to answer queries about documents and charts (Mathew et al., 2021; Masry et al., 2022), signs, and everyday scenes (Singh et al., 2019; Biten et al., 2019). This capability also creates a semantic attack surface: typographic attacks insert misleading text that redirects answers away from visual evidence without changing the queried object. Unlike pixel-level perturbations, these attacks exploit readable semantics and remain effective in synthetic, real-world, and scene-coherent settings (Westerhoff et al., 2026; Cao et al., 2025). A practical defense should therefore suppress the influence of misleading attack text without requiring access to the target VLM's internals.

Existing defenses only partly meet this requirement. Defense-Prefix (Azuma & Matsui, 2023) and Dyslexify (Hufe et al., 2026) are specifically designed for CLIP-style models. Defense-Prefix learns model-specific prefixes that are tightly coupled with the CLIP representation space, while Dyslexify requires identifying and ablating internal attention heads in the CLIP vision encoder. Although effective for their target architectures, these approaches are difficult to generalize to closed-source VLMs, where model internals are inaccessible and the task paradigm extends beyond CLIP-style classification. Figure 1 contrasts existing defenses with QuISE's black-box design across VLMs.

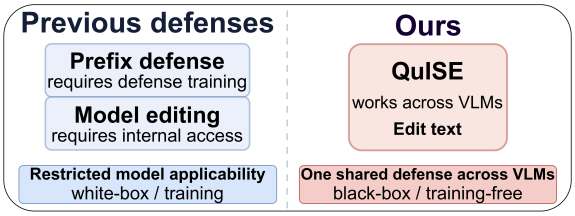


Figure 1: Comparison of existing defenses and QuISE.

To explore a black-box defense against typographic attacks on VLMs, we modify the inserted attack text to mitigate its effect. We first investigate which properties of attack text should be modified. Specifically, attack text has two fundamental aspects: appearance and semantics. Controlled modifications show that appearance modifications provide limited defense, whereas semantic modifications restore answer accuracy close to that on clean images.

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.

We further investigate what semantics the attack text should be transformed into. Simply removing visible text is undesirable, because direct deletion may damage local visual evidence and disrupt the surrounding visual structure. Meanwhile, arbitrary semantic replacement may introduce new answer-related cues and remain influential to the model. Therefore, we analyze textual influence in the full multimodal input, where image text's effect is determined by its interaction with both the query and the visual content. Specifically, we characterize replacement semantics from two perspectives: (1) query relevance, which measures whether the text directly provides information needed to answer the query; and (2) image relevance, which measures whether the text is supported by or consistent with the visual content. We compare replacements under the four semantic conditions formed by these two properties. The comparison shows that query relevance largely determines textual influence. Image relevance substantially affects the answer when the text is query-relevant, but has limited impact once the text becomes query-irrelevant. Consequently, query-irrelevant text is less likely to steer the model and is better suited for defensive replacement. Among the four conditions, Q0-I0, which is irrelevant to both the query and the image, provides the strongest defense.

Based on these observations, we propose **QuISE**, a model-agnostic, training-free black-box defense based on *query-irrelevant semantic editing*. Without first determining whether an image is clean or attacked, QuISE applies two stages to each image–query pair. *Influence-Aware Text Localization* combines target-VLM semantic judgment with auxiliary spatial grounding to identify text regions that may affect the current query. *Query-Irrelevant Semantic Editing-based Defense* then creates two edited images using semantically distinct replacements verified to be irrelevant to both the query and the image. QuISE adopts an edited answer only when both edited images yield the same valid answer. Otherwise, it uses the target VLM's answer on the input image. This consensus-based selection supports answer recovery under attack.

Our main contributions are: (1) To our knowledge, we present the first black-box defense against typographic attacks on VLMs via semantic editing. Our design stems from two key observations: semantic modifications outperform appearance modifications against the evaluated attacks, and query relevance largely determines textual influence. (2) We propose QuISE, a model-agnostic, training-free black-box defense combining influence-aware localization, query-irrelevant semantic editing and consensus-based answer selection. (3) We evaluate QuISE on three attack benchmarks and four open-weight and closed-source VLMs, and show that it outperforms three representative defenses.

2 Related Work

2.1 Visual Text Understanding in VLMs

VLMs have progressed from contrastive image–text representation learning to general-purpose visual assistants. LLaVA (Liu et al., 2023) introduced visual instruction tuning for open-ended question answering and multimodal dialogue. Subsequent models have further strengthened fine-grained perception and text-rich image understanding. The Qwen-VL family (Bai et al., 2023; 2025) supported visual grounding, text reading, high-resolution perception, and document parsing, while LLaVA-OneVision (Li et al., 2025) and InternVL (Chen et al., 2024) extended multimodal reasoning across diverse visual tasks and input settings. Within this broader progression, interpreting image text has become an important component of VLM reasoning. TextVQA (Singh et al., 2019) and ST-VQA (Biten et al., 2019) evaluated whether models could answer queries by jointly reasoning over scene text and visual context, whereas OCRBench (Liu et al., 2024) broadened this evaluation to text recognition, scene-text VQA, and document understanding.

2.2 Typographic Attacks on VLMs

Typographic attacks exploit the tendency of VLMs to treat image text as semantic evidence. Multimodal Neurons (Goh et al., 2021) demonstrated that misleading labels could override depicted content in CLIP classification. Disentangling Visual and Written Concepts in CLIP (Materzyńska et al., 2022) related this behavior to the entanglement of written and visual concepts in the image encoder. Recent studies have extended the threat to generative VLMs and more realistic settings. Self-generated Typographic Attacks (Qraitem et al., 2024) used an LVLM to produce context-dependent misleading text, while SCAM (Westerhoff et al., 2026) evaluated both synthetic and physically captured attacks. SceneTAP (Cao et al., 2025) jointly planned adversarial text and its placement to produce

scene-coherent attacks. Typographic Attacks in a Multi-Image Setting (Wang et al., 2025b) improved stealth by selecting diverse, non-repeating attack words across an image set. Image text can also redirect model behavior as an instruction. FigStep (Gong et al., 2025) converted harmful instructions into typographic visual prompts to bypass text-side safety alignment, whereas Goal Hijacking via Visual Prompt Injection (Kimura et al., 2024) embedded an alternative task in the image to replace the model's objective. These studies show that typographic attacks affect not only classification and query answering, but also instruction following and multimodal safety.

2.3 Defense against Typographic Attacks

Only a limited number of defense approaches have been proposed for typographic attacks. Defense-Prefix (Azuma & Matsui, 2023) learned a special prefix token placed before class names, making CLIP-based classifiers less sensitive to misleading words in the image. PAINT (Ilharco et al., 2022) fine-tuned an open-vocabulary model on a target patching task and interpolated the original and fine-tuned weights to improve the target capability while preserving general performance. Dyslexify (Hufe et al., 2026) identified attention heads that causally transmitted typographic information through the CLIP vision encoder and selectively ablated them without retraining. Defense-Prefix depends on a CLIP-style class-name interface, whereas PAINT and Dyslexify require access to model weights or internal components. These requirements limit their direct application to open-ended and closed-source VLMs. QuISE instead localizes query-conditioned influential text, applies semantically distinct replacements verified to be irrelevant to both the query and the image, and changes the input-image answer only when the two edited answers form a valid consensus. This design enables a training-free input-level defense without access to the target model's internal components.

3 Motivation

To guide the design of a black-box defense against typographic attacks, we examine how modifications to inserted attack text affect VLM answers. We compare appearance and semantic modifications, then identify which replacement semantics most effectively weaken attack-text influence.

3.1 Change What: Appearance or Semantics?

Setup. We sample 600 attacked image–query pairs, with 200 each from SCAM, SceneTAP, and SELF. Each modification is applied only within the inserted-text region. Qwen2.5-VL-7B-Instruct and LLaVA-OneVision-8B answer the same query on each resulting image.

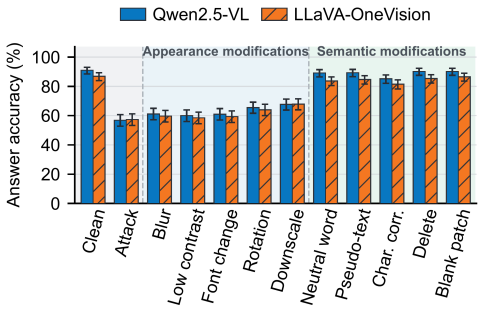


Figure 2: Accuracy by modification type.

Clean denotes the image before attack text is added, whereas *Attack* denotes the image with attack text. The five appearance modifications preserve the semantics of attack text: *Blur*, *Low contrast*, *Font change*, *Rotation*, and *Downscale*. The five semantic modifications alter or remove textual meaning: *Neutral word* replaces attack text with a broader concept (e.g., "pen" → "object"); *Pseudo-text* replaces it with meaningless characters; *Character corruption* randomly replaces characters with symbols; *Delete* removes the text region; and *Blank patch* fills the region with background color.

Observation 1: Semantic modifications defend better than appearance modifications. Figure 2 compares the two modification families across both VLMs. Attacked images achieve an average accuracy of 57.08%. Appearance modifications improve the accuracy to 64.64%. Semantic modifications substantially raise it to 86.57%, approaching the 88.92% accuracy on clean images. This indicates that the semantic content of inserted text is the primary factor driving typographic attacks.

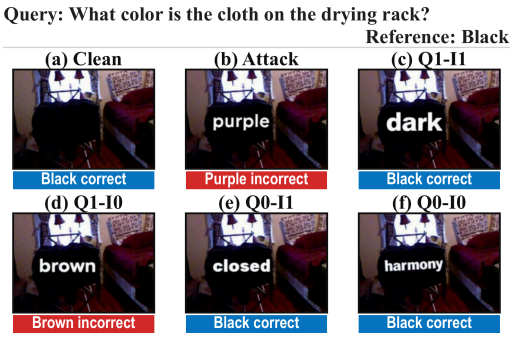


Figure 3: Four semantic edits and VLM answers.

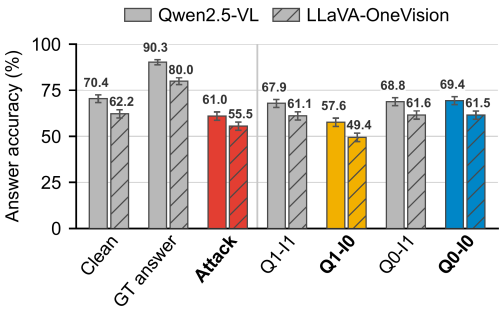


Figure 4: Acc under four semantic conditions.

3.2 Which Semantics for Defense?

Semantic properties. Observation 1 shows that semantic modification is more effective than appearance modification. We therefore study which replacement semantics can minimize the influence of inserted text. Specifically, we characterize semantics along two dimensions: query relevance (**Q**) and image relevance (**I**), where 1 and 0 denote relevance and irrelevance, yielding four conditions: **Q1-I1** (relevant to both), **Q1-I0** (query-relevant only), **Q0-I1** (image-relevant only), and **Q0-I0** (irrelevant to both). Figure 3 illustrates these conditions with an example.

Setup. We generate all four verified semantic variants for 1,810 image–query pairs: 1,010 from SCAM and 800 from SceneTAP. For each pair, we evaluate the *Clean* and *Attack* versions, a version with the ground-truth answer inserted as image text, and the four semantic variants. The *Clean* and *Attack* versions serve as the pre-attack and attacked references, respectively. The four semantic variants are evaluated as controlled defense modifications.

Observation 2: Query relevance largely determines textual influence. Figure 4 compares the four semantic conditions. For query-relevant replacements, Q1-I1 consistently outperforms Q1-I0 by 10.28 and 11.66 points on Qwen2.5-VL and LLaVA-OneVision, respectively. In contrast, Q0-I1 and Q0-I0 achieve comparable performance, indicating that image relevance has limited impact once the text is query-irrelevant. These results show that query relevance is the primary factor determining whether inserted text influences VLM decisions. Among the four conditions, Q0-I0 provides the strongest defense. It improves accuracy over the *Attack* condition by 8.40 and 6.02 points on Qwen2.5-VL and LLaVA-OneVision, respectively, while approaching *Clean* performance. We therefore select Q0-I0 as the target replacement semantics for QuISE.

4 Method

4.1 Problem Formulation

Given an image–query pair (I, q), let $\mathcal{M}$ denote the target VLM, which produces:

$$y_0 = \mathcal{M}(I, q). \tag{1}$$

The input image may contain visible text. It may contain attack text inserted by a typographic attack. We call I a clean image if it contains no attack text, and an attacked image if it contains attack text. At inference, the defense only observes (I, q) and does not know whether I is clean or attacked. We aim to design a defense $\mathcal{D}$ that outputs a correct answer:

$$\hat{y} = \mathcal{D}(\mathcal{M}; I, q), \tag{2}$$

where the ground-truth answer y^* is evaluation-only. An effective defense should suppress attack-induced textual influence without indiscriminately editing all visible text regions.

We consider a black-box setting: $\mathcal{D}$ interacts with $\mathcal{M}$ only via input manipulation and output analysis, with no access to model parameters, gradients, probabilities, or intermediate representations.

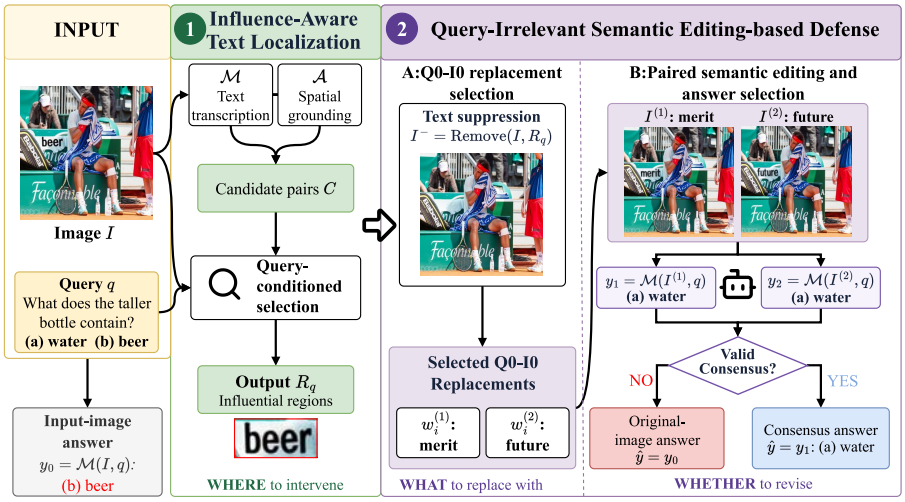


Figure 5: Overview of QuISE.

4.2 Overview

Figure 5 shows the two stages of QuISE. First, *Influence-Aware Text Localization* identifies text regions that may affect the answer to the query. Second, *Query-Irrelevant Semantic Editing-based Defense* replaces each localized region with two verified Q0-I0 replacements to generate two edited images. QuISE accepts an edited answer only if both edited images yield consistent responses.

4.3 Influence-Aware Text Localization

Not every visible text region affects the answer to the current query, so it is unnecessary to edit all detected text. QuISE identifies text regions that may influence the target model's answer and edits only the selected regions. Specifically, QuISE first extracts candidate text regions from the image. The target model $\mathcal{M}$ is used to recognize visible text, while a fixed auxiliary component $\mathcal{A}$ provides the corresponding spatial locations. The resulting candidate set is represented as:

$$C = \text{Associate}_{\mathcal{M},\mathcal{A}}(I) = \{(l_j, t_j)\}_{j=1}^{N}, \tag{3}$$

where C contains N candidate location–text pairs. For the j-th candidate, l_j denotes the text location and t_j denotes its content. The candidate set only describes visible text and does not indicate whether a region affects the query. Given (I, q, C), QuISE uses the target model to select influential regions:

$$R_q = \text{Select}_{\mathcal{M}}(I, q, C) = \{(l_i, t_i)\}_{i=1}^{K} \subseteq C, \tag{4}$$

where R_q contains K selected location–text pairs for subsequent semantic editing. Given the image and query, the target model selects all text regions that may inform the answer.

4.4 Query-Irrelevant Semantic Editing-based Defense

Q0-I0 replacement selection. After locating potentially influential text regions, QuISE edits their semantics. Direct removal may discard useful visual information and disrupt the local structure. Arbitrary replacement may introduce new answer-related cues. QuISE therefore replaces the text in each localized region with content irrelevant to both the current query and the image. This operation reduces the influence of potentially misleading semantics while preserving the visual form and layout of the selected regions as much as possible.

To obtain suitable replacements, QuISE first removes the localized text regions R_q:

$$I^- = \text{Remove}(I, R_q), \tag{5}$$

I^- is the text-removed image for evaluating image relevance. For each localized region, QuISE uses the target model $\mathcal{M}$ to generate candidate replacements conditioned on the query q and I^-. It retains

only candidates that are irrelevant to both the query and the image, referred to as Q0-I0 replacements. For the i-th localized region, QuISE selects two replacements with different semantics:

$$P_i = (w_i^{(1)}, w_i^{(2)}), \qquad i = 1, \ldots, K, \tag{6}$$

where K counts localized regions and $(w_i^{(1)}, w_i^{(2)})$ is the replacement pair selected for region i.

Paired semantic editing and answer selection. Using the replacement pairs, QuISE generates two edited images:

$$(I^{(1)}, I^{(2)}) = \mathrm{Edit}\big(I, R_q, \{P_i\}_{i=1}^{K}\big), \tag{7}$$

where each edited image applies one replacement from every pair. The target model produces:

$$y_1 = \mathcal{M}(I^{(1)}, q), \qquad y_2 = \mathcal{M}(I^{(2)}, q). \tag{8}$$

QuISE accepts the edited answer only when both answers are valid and consistent:

$$\hat{y} = \begin{cases} y_1, & y_1, y_2 \text{ are valid and } y_1 = y_2, \\ y_0, & \text{otherwise}, \end{cases} \tag{9}$$

where $y_0 = \mathcal{M}(I, q)$ is the target model's answer on the input image. An answer is valid if it can be parsed into the task-specific answer space and is neither empty nor a refusal.

Preserving clean-image predictions. QuISE is designed to limit unnecessary answer changes on clean images. If no candidate text region is identified or no region is selected, QuISE skips editing and retains the input-image answer y_0. On a clean image, a selected region can contain text needed to answer the query, such as a sign label or chart value. The two edited images replace this text with different Q0-I0 replacements. If the answer depends on the selected text, the edited answers may differ or become invalid. No valid consensus is then formed, and QuISE retains y_0. This fallback can limit performance loss on clean images, including queries that require reading visible text.

5 Experiments

We evaluate QuISE's defense against typographic attacks. We also assess whether it preserves performance on clean images. A clean image contains no attack text. It may contain no visible text or text unrelated to the query. It may also contain text that must be read to answer the query.

5.1 Experimental Settings

Target models and datasets. We evaluate QuISE's defense effectiveness using Qwen2.5-VL-7B-Instruct (Bai et al., 2025), LLaVA-OneVision-8B (Li et al., 2025), InternVL3.5-8B (Wang et al., 2025a), and GPT-4.1-mini (OpenAI, 2025) on SCAM (Westerhoff et al., 2026), SceneTAP (Cao et al., 2025), and the self-generated typographic attack benchmark (SELF) (Qraitem et al., 2024). SCAM contains 1,162 image–query pairs with both synthetic and real-world attacked variants, reported separately; SceneTAP and SELF contain 1,572 and 4,662 attacked pairs, respectively. Each target model is thus evaluated on 8,558 attacked instances and their clean counterparts. To evaluate clean text-reading performance, we use fixed subsets of 500 image–query pairs each from TextVQA (Singh et al., 2019) and ST-VQA (Biten et al., 2019). These images contain no attack text, and each query requires reading visible text to obtain the answer.

Baselines. We compare QuISE with representative defenses. Artifact-aware Prompting (AAP) (Qraitem et al., 2025) is a prompt-based mitigation that first asks the VLM to inspect suspicious text or logos and then answers the query using the inspection context. Defense-Prefix (Azuma & Matsui, 2023) learns defensive prefixes to reduce CLIP sensitivity to typographic text, while Dyslexify (Hufe et al., 2026) removes CLIP vision-encoder attention heads that transmit textual information. We also include Localized Deletion, which shares QuISE's localization stage but removes text regions instead of editing their semantics, to evaluate the contribution of semantic editing. AAP is evaluated on the same target models and instances as QuISE. Since Defense-Prefix and Dyslexify rely on CLIP-family backbones, we compare their defense gains, RR, and HR on the same benchmark inputs while noting the different target models. CLIP-based defenses are excluded from TextVQA and ST-VQA because they do not support open-ended scene-text question answering.

Metrics. Following prior work (Westerhoff et al., 2026; Cao et al., 2025; Qraitem et al., 2024), we use answer accuracy as the primary metric. We report attacked accuracy $\mathrm{Acc}_{\mathrm{att}}$, defended accuracy $\mathrm{Acc}_{\mathrm{def}}$, and improvement $\Delta = \mathrm{Acc}_{\mathrm{def}} - \mathrm{Acc}_{\mathrm{att}}$. For sample i, $c_i, a_i, d_i \in \{0, 1\}$ indicate clean, attacked, and defended answer correctness. Recovery Rate (RR) measures correction of attack-induced errors; Harm Rate (HR) measures corruption of correct attacked answers:

$$\mathrm{RR} = 100\,\frac{|\{i : c_i = 1, a_i = 0, d_i = 1\}|}{|\{i : c_i = 1, a_i = 0\}|}, \qquad \mathrm{HR} = 100\,\frac{|\{i : a_i = 1, d_i = 0\}|}{|\{i : a_i = 1\}|}. \tag{10}$$

Higher accuracy and RR, together with lower HR, indicate better defense performance. We report overall results across 8,558 attacked instances and per-setting results.

For these clean text-reading tasks, we use VQA soft accuracy for TextVQA and average normalized Levenshtein similarity (ANLS; threshold 0.5) for ST-VQA. For each defense policy r, we define the macro-average score $M_r = (S_r^{\mathrm{TextVQA}} + S_r^{\mathrm{ST\text{-}VQA}})/2$ and performance retention $\mathrm{Ret}_r = 100 M_r / M_0$, where M_0 is the macro-average score without defense.

Implementation. We use PP-OCRv6 in PaddleOCR 3.7.0 (Zhang et al., 2026) as the auxiliary text-localization component, LaMa (Suvorov et al., 2022) to remove the localized text, and RS-STE (Fang et al., 2025) to render replacement text. We use the same components, prompts, and hyperparameters across all target models and attack datasets. We run all local inference on a single NVIDIA vGPU-32GB with 32 GB of memory and access GPT-4.1-mini through its API. If any defense stage fails, we retain the input-image answer and include the sample in all metrics. Further implementation details are provided in the supplementary material.

5.2 Qualitative Analysis

Figure 6 presents representative defense cases across four attack settings. Query-irrelevant edits suppress the misleading text and produce the same correct answer.

5.3 Defense Effectiveness

Table 1 compares QuISE with representative defenses across target models, reporting attacked and defended accuracy, gains over attacked inputs, RR, and HR. QuISE consistently outperforms existing defenses overall across target models and attack settings. Compared with AAP, the strongest VLM-based baseline, QuISE improves overall defended accuracy by 13.9–27.1 percentage points, achieving 67.9–75.0% RR with only 0.5–1.1% HR. In contrast, AAP achieves only 18.8–29.2% RR and introduces higher HRs of 2.3–13.0%. These results show that semantic editing is substantially more effective than prompt-based mitigation for removing misleading textual influence.

QuISE also outperforms CLIP-based defenses. Defense-Prefix and Dyslexify achieve 39.3% and 57.9% RR with 3.2% and 1.9% HR, respectively, whereas QuISE achieves higher RR and lower HR. Although these methods rely on different CLIP-family backbones, the comparison on the same benchmark inputs shows that QuISE provides stronger attack recovery without requiring model-specific modification. The advantage of QuISE is particularly evident under stronger attacks. On

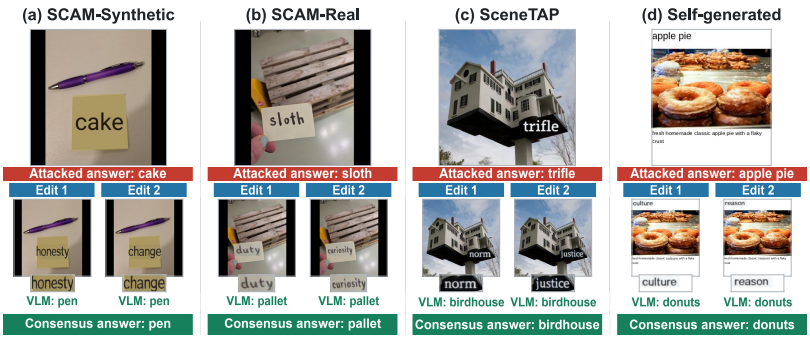


Figure 6: QuISE on attacked images: two query-irrelevant edits recover the same answer. Red/green bars: attacked/recovered consensus answers.

Table 1: Typographic-attack defense (%). SCAM-S/R: synthetic/real-world SCAM; parentheses: gains over attacked inputs; bold: better of AAP and QuISE per model.

Target model	Defense method	SCAM-S	SCAM-R	SceneTAP	SELF	Overall	RR↑	HR↓
OpenAI CLIP ViT-B/32	Defense-Prefix	63.1 (+20.2)	79.3 (+13.0)	58.8 (+14.8)	47.0 (+14.7)	55.7 (+15.2)	39.3	3.2
OpenCLIP ViT-B/16	Dyslexify	83.8 (+30.7)	89.0 (+18.8)	55.2 (+14.7)	57.2 (+33.3)	64.8 (+27.6)	57.9	1.9
Qwen2.5-VL-7B-Instruct	Attacked	89.2	89.7	71.9	17.1	46.8	–	–
	AAP	97.5 (+8.3)	96.6 (+6.9)	83.3 (+11.4)	19.0 (+1.9)	52.0 (+5.2)	18.8	6.3
	QuISE	**99.2 (+10.0)**	**98.6 (+9.0)**	**89.5 (+17.6)**	**63.7 (+46.5)**	**78.0 (+31.2)**	**73.9**	**0.8**
LLaVA-OneVision-8B	Attacked	86.8	92.9	72.3	11.3	43.8	–	–
	AAP	88.3 (+1.5)	86.9 (−6.0)	79.6 (+7.4)	14.4 (+3.1)	46.2 (+2.4)	19.9	13.0
	QuISE	**98.6 (+11.8)**	**98.7 (+5.8)**	**92.7 (+20.5)**	**54.2 (+42.9)**	**73.3 (+29.5)**	**75.0**	**0.5**
InternVL3.5-8B	Attacked	67.6	78.7	59.3	4.5	33.2	–	–
	AAP	91.0 (+23.4)	89.4 (+10.8)	80.7 (+21.4)	10.7 (+6.2)	45.2 (+11.9)	29.2	2.3
	QuISE	**96.0 (+28.5)**	**95.6 (+17.0)**	**87.3 (+28.1)**	**39.2 (+34.7)**	**63.4 (+30.2)**	**70.3**	**1.1**
GPT-4.1-mini	Attacked	99.0	99.1	86.8	54.8	72.7	–	–
	AAP	**99.7 (+0.8)**	**99.7 (+0.5)**	**93.4 (+6.6)**	50.8 (−4.0)	71.9 (−0.8)	18.8	7.0
	QuISE	99.4 (+0.4)	**99.7 (+0.5)**	92.4 (+5.6)	**76.7 (+21.9)**	**85.8 (+13.1)**	**67.9**	**0.8**

Table 2: Edit-count and consensus analysis.

Policy	Acc.↑	RR↑	HR↓	Ret.↑	Edits↓
No Defense	46.83	–	–	100.00	0
Single Edit (avg.)	79.38	76.48	0.94	40.46	1
Three-Edit Majority	**79.47**	**76.70**	0.92	76.21	3
Three-Edit Unanimous	77.37	72.43	**0.72**	**83.24**	3
QuISE	78.00	73.85	0.80	80.49	2

Table 3: Qwen results (gains vs. Attacked, pp).

Method	Acc. (%) ↑	$\bar{t}$ (s/input) ↓
Attacked	67.5	0.208
AAP	76.5 (+9.0)	3.208
Localized Deletion	89.5 (+22.0)	2.057
QuISE	89.0 (+21.5)	3.363

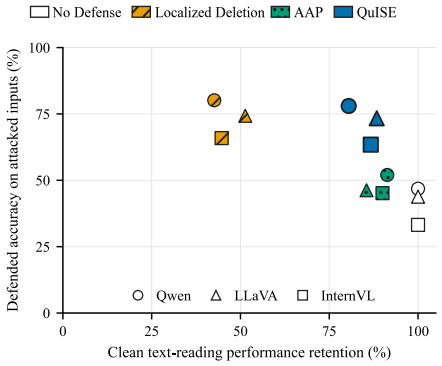


Figure 7: Defense vs. text-reading retention.

SELF, where attacked accuracy is substantially lower, QuISE improves accuracy by 21.9–46.5 points, while AAP provides limited gains and even decreases GPT-4.1-mini accuracy. These results demonstrate that directly modifying misleading text semantics is more effective than relying only on prompting when VLMs are strongly affected by typographic attacks.

5.4 Effectiveness on Clean Text-Reading Tasks

Figure 7 compares QuISE, AAP, and Localized Deletion across three open-weight VLMs in defense effectiveness and clean text-reading performance retention. Localized Deletion achieves strong defense performance by removing selected text regions but may also remove text needed to answer the query. QuISE remains within 0.9–2.5 accuracy points of Localized Deletion while achieving 37.0–42.0 percentage points higher performance retention. Compared with AAP, QuISE improves defense effectiveness on all target models. These results show that QuISE better balances defense effectiveness with performance on clean text-reading tasks.

5.5 Analysis of Edit Count and Consensus Policy

We vary edit count and consensus policy on Qwen2.5-VL-7B-Instruct, keeping other components fixed. Single Edit averages dataset-level metrics over three separately evaluated fixed one-edit settings. Three-Edit Majority requires at least two of three edited answers to agree; Three-Edit Unanimous requires all three. QuISE uses two edits and retains the input-image answer unless both edited answers agree. Table 2 shows that Single Edit achieves 76.48% RR but only 40.46% performance retention on clean text-reading tasks. Three-Edit Majority slightly improves accuracy and RR over QuISE, requiring one extra edited image and model query. Three-Edit Unanimous raises retention to

Table 4: Clean recognition accuracy; parentheses: changes from No Defense.

	Clean Images		
Model	Aircraft	Food-101	ImageNet-100
Qwen2.5-VL	64.64 (0.00)	75.72 (0.00)	98.70 (0.00)
LLaVA-OneVision	58.94 (0.00)	77.32 (+0.10)	98.80 (0.00)
InternVL3.5	41.46 (0.00)	67.63 (0.00)	98.40 (0.00)

Table 5: Defense-aware adaptive attacks against QuISE (%).

	Qwen2.5-VL-7B		InternVL3.5-8B	
Attack	A	B	A	B
Effective new attacks	79.0	29.0	84.0	21.0
QuISE correction rate	88.6	79.3	84.5	76.2
Successful bypasses	3.0	4.0	6.0	1.0

83.24% and lowers HR to 0.72%, but reduces RR to 72.43%. QuISE balances recovery, retention, and cost with two edits, achieving 73.85% RR, 0.80% HR, and 80.49% retention.

5.6 Inference Overhead

We benchmark Qwen2.5-VL-7B-Instruct (BF16, one NVIDIA vGPU-32GB) on 200 attacked images (50 per setting), repeating each method three times. We report subset accuracy and mean end-to-end latency per image, $\bar{t}$, covering all processing through the final answer except initialization and warmup. Models remain resident, and inputs are processed individually without reusing results. In Table 3, QuISE exceeds AAP's accuracy by 12.5 percentage points with 0.155 s additional mean latency. Localized Deletion is faster and 0.5 percentage points more accurate on this subset.

5.7 Performance on Clean Recognition Tasks Not Requiring Image Text

We further evaluate QuISE on clean recognition tasks not requiring image text. We use fixed multiple-choice subsets of FGVC-Aircraft (Maji et al., 2013), Food-101 (Bossard et al., 2014), and ImageNet-100 (Shekhar, 2021), with 1,001, 1,001, and 1,000 samples.

As shown in Table 4, QuISE causes no accuracy drop on any model–dataset pair and introduces no harmful correct-to-incorrect flips across 9,006 predictions. The paired-consensus and fallback rule preserves target-model predictions on the evaluated clean-image recognition tasks.

5.8 Adaptive, Defense-Aware Attacks

We test gray-box attacks on Qwen2.5-VL-7B-Instruct and InternVL3.5-8B. Attackers know QuISE and can query stage outputs and target answers, but have no access to weights or gradients. For each target, we randomly select 50 images each from SCAM-S and SceneTAP with correct clean answers, successful static attacks, and QuISE-selected attack text, excluding development images. Both attacks use these 100 sources and at most four new variants per image. **A: Localization evasion** changes font, spacing, rotation, or contrast while preserving the text, seeking to evade localization in C. **B: Selection evasion** uses the other VLM to propose semantic variants, seeking omission from R_q despite localization in C. We rank candidates by target errors and stage evasion, then evaluate one input per source and attack with full QuISE, using the static attack as a fallback.

Effective new attacks are selected new variants that mislead the target; *successful bypasses* additionally evade the targeted stage and leave QuISE's answer incorrect. Both rates use all 100 sources per target, excluding static fallbacks from their numerators. *Correction rate* measures QuISE's accuracy on effective new attacks. Table 5 shows correction rates of 76.2–88.6%, with successful bypasses on 1.0–6.0% of sources. These results indicate that QuISE retains its corrective capability on most selected effective attacks even when the attacker uses defense-stage feedback to search for localization or selection evasion, under the evaluated budget and sampling protocol.

6 Conclusion

We investigated defenses against typographic attacks on vision-language models and found that their effectiveness mainly stems from the semantics of inserted text rather than visual appearance. Based on this insight, we proposed QuISE, a model-agnostic, training-free black-box defense that mitigates misleading textual influence. Extensive experiments show that QuISE achieves strong attack recovery across diverse VLMs and attack settings while preserving clean-image performance.

AI USE STATEMENT

We used generative AI tools to improve the wording and readability of the manuscript and assist with LaTeX formatting. The authors take full responsibility for the final content, including all technical claims, experimental results, and conclusions.

REFERENCES

Hiroki Azuma and Yusuke Matsui. Defense-prefix for preventing typographic attacks on CLIP. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 3644–3653, 2023.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusiñol, Ernest Valveny, C. V. Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4291–4301, 2019.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101: Mining discriminative components with random forests. In *European Conference on Computer Vision*, pp. 446–461, 2014.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25050–25059, 2025.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024.

Zhengyao Fang, Pengyuan Lyu, Jingjing Wu, Chengquan Zhang, Jun Yu, Guangming Lu, and Wenjie Pei. Recognition-synergistic scene text editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13104–13113, 2025.

Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021. doi: 10.23915/distill.00030. URL `https://distill.pub/2021/multimodal-neurons/`.

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*, pp. 23951–23959, 2025. doi: 10.1609/aaai.v39i22.34568.

Lorenz Hufe, Constantin Venhoff, Erblina Purelku, Maximilian Dreyer, Sebastian Lapuschkin, and Wojciech Samek. Dyslexify: A mechanistic defense against typographic attacks in CLIP. In *International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=UI7mbsIZeN`.

Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights. In *Advances in Neural Information Processing Systems*, volume 35, pp. 29262–29277, 2022.

Subaru Kimura, Ryota Tanaka, Shumpei Miyawaki, Jun Suzuki, and Keisuke Sakaguchi. Empirical analysis of large vision-language models against goal hijacking via visual prompt injection, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023. doi: 10.52202/075280-1516.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. OCRBench: On the hidden mystery of OCR in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024. doi: 10.1007/s11432-024-4235-6.

Subhransu Maji, Juho Kannala, Esa Rahtu, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. Technical Report arXiv:1306.5151, arXiv, 2013.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

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 C. V. 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.

OpenAI. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/, 2025. Accessed: 2026-07-24.

Maan Qraitem, Nazia Tasnim, Piotr Teterwak, Kate Saenko, and Bryan A. Plummer. Vision-LLMs can fool themselves with self-generated typographic attacks, 2024.

Maan Qraitem, Piotr Teterwak, Kate Saenko, and Bryan A. Plummer. Web artifact attacks disrupt vision language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1048–1057, 2025.

Ambesh Shekhar. Imagenet100: A sample of imagenet classes. Kaggle Dataset, https://www.kaggle.com/datasets/ambityga/imagenet100, 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.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2149–2159, 2022.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu,

Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025a.

Xiaomeng Wang, Zhengyu Zhao, and Martha Larson. Typographic attacks in a multi-image setting. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 12594–12604. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.naacl-long.626. URL https://aclanthology.org/2025.naacl-long.626/.

Justus Westerhoff, Erblina Purelku, Jakob Hackstein, Jonas Loos, Leo Pinetzki, Erik Rodner, and Lorenz Hufe. SCAM: A real-world typographic robustness evaluation for multimodal foundation models. *Journal of Data-centric Machine Learning Research*, 3(4):1–30, 2026. URL https://openreview.net/forum?id=zCcJSErVHH.

Yubo Zhang, Xueqing Wang, Manhui Lin, Yue Zhang, Penglongyi Deng, Ting Sun, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Changda Zhou, Hongen Liu, Suyin Liang, Cheng Cui, Yi Liu, Dianhai Yu, and Yanjun Ma. PP-OCRv6: From 1.5m to 34.5m parameters, surpassing billion-scale VLMs on OCR tasks, 2026.

A Additional Details of the Motivation Studies

A.1 Appearance and Semantic Modifications

The 600-sample study uses 200 attacked pairs from each of SCAM, SceneTAP, and SELF. SCAM contains equal synthetic and real subsets, SceneTAP is stratified by question type, and SELF is balanced across its five source categories. Pilot samples are excluded. Every operator is restricted to the same inserted-text region, obtained by aligning the clean–attack difference mask with frozen OCR boxes. Qwen2.5-VL-7B-Instruct and LLaVA-OneVision-8B answer the same fixed multiple-choice question under every condition. Table 6 gives the fixed operators used in Figure 2 of the main paper.

Table 6: Controlled modification settings used for Observation 1.

Appearance modifications	
Blur	Gaussian blur blended with $\alpha = 0.65$; kernel ≈ 0.55 of median text height
Low contrast	retain 0.65 of each text-stroke value and mix 0.35 local background
Font change	erase and redraw the same string in DejaVuSerif-Italic
Rotation	affine transform with 6° rotation, 0.04 shear, and 0.98 scale
Downscale	scale the original text layer by 0.5 and center it in the source box
Semantic modifications	
Neutral word	replace with a task-level abstract term, e.g., *object*, *color*, or *number*
Pseudo-text	substitute letters from `qzxvkjw` and digits randomly while preserving length and case
Char. corruption	replace 40% of alphanumeric positions with symbols
Delete	inpaint the selected support mask with LaMa
Blank patch	fill each OCR rectangle with its surrounding-ring mean color

A.2 Construction of the Four Semantic Conditions

Table 7 summarizes how the four semantic variants defined by query relevance (Q) and image relevance (I) are constructed and accepted.

Table 7: Construction of the four semantic conditions in Observation 2.

Condition	Construction and acceptance rule
Q1-I1	Generate non-copy proposals in the queried answer category from the text-removed image, question, task type, and ground-truth answer; select one supported by the queried image content.
Q1-I0	Form a fixed task-category pool, remove ground-truth-answer and attack-text surfaces, and select a same-category concept that is false or unsupported for the queried image content.
Q0-I1	Generate 8–12 visible concepts from seven visual dimensions; remove overlaps with the query target, ground-truth answer, attack text, and queried dimension, then select a visible query-irrelevant concept.
Q0-I0	Select deterministically from a frozen abstract pool after removing overlaps with the question, ground-truth answer, attack text, and the other Q/I conditions.

The four variants are generated from 2,237 eligible non-complex SCAM and SceneTAP pairs and evaluated on the common 1,810-pair subset for which all conditions are available. Answer options are removed, and all conditions use the same concise-answer prompt. Qwen2.5-VL-7B-Instruct performs the image-conditioned proposal and selection steps, while Q0-I0 is selected deterministically. Rule-based checks reject copied anchors, invalid categories, cross-condition duplicates, and malformed surfaces. The protocol is frozen after validation on two disjoint 175-sample sets and a sampled human audit, and the resulting edits are held fixed across target models. Responses unresolved by deterministic normalization are judged without access to the semantic condition.

B Additional Implementation Details of QuISE

B.1 Defense-Time Localization

At inference time, the target VLM first recognizes candidate text in the input image, while PaddleOCR provides spatial text boxes. Normalized text matching associates the VLM-recognized strings with individual or merged adjacent OCR boxes. Given the image, query, answer options, and grounded candidates, the same target VLM selects all text regions that may influence its answer. Only these grounded regions are passed to semantic editing. If no influential text can be reliably localized, or any localization stage fails, QuISE retains the input-image answer.

B.2 Defense-Time Q0-I0 Selection

For each localized text group, QuISE constructs a deterministic, geometry-compatible shortlist from a frozen abstract candidate pool after removing candidates that overlap with the source text, query, or answer options. The target VLM then evaluates every candidate against the query and the text-removed image, accepting it as Q0-I0 only when it is judged irrelevant to both; uncertain candidates are rejected. If fewer than two candidates are accepted, QuISE performs one additional attempt with a disjoint shortlist. It selects two verified candidates, preferring different semantic classes, and uses them to create the two edited images. The ground-truth answer and the input-image answer are never provided during selection. Insufficient verified candidates or any subsequent editing failure triggers the original-answer fallback.