QUALITY HIJACKING: SEMANTIC-AWARE TYPOGRAPHIC ATTACKS ON IMAGE QUALITY ASSESSMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language model-based image quality assessment methods (VLM-IQA) have shown strong capability in evaluating complex and out-of-distribution images, yet remain vulnerable to typographic attacks. Attacking VLM-IQA requires adversarial text to effectively manipulate quality-related reasoning while remaining visually consistent with the underlying image. Existing typographic attacks typically rely on generic adversarial texts (ATs) and heuristic placement strategies, overlooking task-related factors and local image degradation, which limits both attack effectiveness and visual plausibility. To address these limitations, we propose Q-Hack, a black-box framework for typographic attacks against VLM-IQA models. Q-Hack adopts a coarse-to-fine strategy to jointly optimize AT content and placement. In the coarse stage, it explores three IQA-related attack factors to generate candidate texts, and naturally places them in suitable regions based on scene semantics, local degradation, and depth. In the fine stage, it exploits target-model feedback to distill effective semantic patterns and refine both text content and placement for stronger attacks. We compare Q-Hack with seven state-of-the-art baselines on three widely used IQA datasets, where it demonstrates clear advantages across attack effectiveness, perceptual fidelity, and scene consistency.

1 INTRODUCTION

Vision-language model-based image quality assessment methods (VLM-IQA) Wu et al. (2024b;c); You et al. (2025); Wang et al. (2024); Wu et al. (2026); Li et al. (2026); Wu et al. (2024a) use language-conditioned reasoning to predict quality scores aligned with mean opinion scores (MOS) across complex scenes and out-of-distribution settings. Their applications range from AIGC evaluation Yu et al. (2025); Lin et al. (2025a); Jiang et al. (2025) to medical and biometric IQA Chen et al. (2024b); Sarıtaş et al. (2026). However, while language-conditioned reasoning enhances the performance and generalization of VLM-IQA methods, it also introduces a new potential attack surface: adversarial text (AT) embedded in images, known as typographic attacks Qraitem et al. (2024); Cao et al. (2025); Wang et al. (2025); Li et al. (2025). This attack surface is particularly low-barrier, as localized misleading text alone can manipulate quality-related reasoning and predicted scores without requiring access to model internals or substantial modification of the underlying image.

While typographic attacks have been extensively studied in other vision-language tasks (e.g., image classification Qraitem et al. (2024); Wang et al. (2025), visual question answering Cao et al. (2025), and image aesthetic assessment Luo et al. (2026)), existing methods are not directly applicable to IQA due to limited quality-specific analysis and uncontrollable insertion-induced quality changes. Generally, existing typographic attacks fall into two types based on text insertion strategy: direct overlay Cheng et al. (2024); Qraitem et al. (2024) and scene-consistent rendering Cao et al. (2025); Luo et al. (2026). The former places AT directly at the center Cheng et al. (2024) or margins Qraitem et al. (2024) of an image to influence the model's predictions. However, they introduce boundary artifacts that degrade the original perceptual quality of the image. The latter Cao et al. (2025); Luo et al. (2026) selects suitable text regions based on scene content and uses a generative image renderer to naturally embed AT. Nevertheless, they rely on predefined AT and region selection strategies without considering the quality-oriented reasoning of IQA models.

In this paper, our objective is to develop a new typographic attack tailored to the unique characteristics of the VLM-IQA tasks, which manipulates predicted quality scores while preserving perceptual

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.

fidelity and consistency of the scene. Achieving this goal requires addressing the following two challenges: ❶ **What text can effectively mislead the model?** The effectiveness of typographic attacks is highly sensitive to the inserted text Cao et al. (2025). For VLM-IQA, effective AT should capture both the multi-dimensional nature of perceptual quality Wu et al. (2024b); Fang et al. (2020) (e.g., specific distortion type and severity) and image-specific characteristics Ke et al. (2006); Saha et al. (2023) (image content and structure). Moreover, different VLM-IQA models may exhibit varying sensitivities to quality-related AT due to their architectures and training data, making it challenging to identify text that is both image-adaptive and model-specific. ❷ **Where can the text be effectively and naturally placed?** Even with effective AT, selecting an appropriate placement remains challenging. Unlike tasks with explicit target objects, VLM-IQA relies on quality evidence from both global and local regions, making it difficult to identify effective attack regions. Meanwhile, the inserted text should preserve perceptual fidelity and maintain consistency with the surrounding scene to avoid detectable artifacts.

To address the above challenges, we propose quality hijacking (**Q-Hack**), a black-box framework for typographic attacks against VLM-IQA models. Specifically, Q-Hack jointly optimizes AT content and placement in a coarse-to-fine manner. In the coarse stage, quality-related adversarial text planning (QRATP) creates candidate ATs from three quality-related types (i.e., contradicting basic facts (CBF), specifying subjective perception (SSP), and specifying target scores (STS)), while degradation-aware text placement (DATP) selects suitable regions based on scene semantics, local degradations, and depth-guided geometry. In the fine stage, IQA-guided refinement (IQAGR) uses target-model feedback and an auxiliary VLM to refine effective text-region pairs for stronger attacks. We evaluate Q-Hack on three widely used IQA datasets against seven baselines. Q-Hack embeds optimized ATs more naturally and inconspicuously, achieving the highest attack performance while maintaining promising perceptual fidelity and scene consistency.

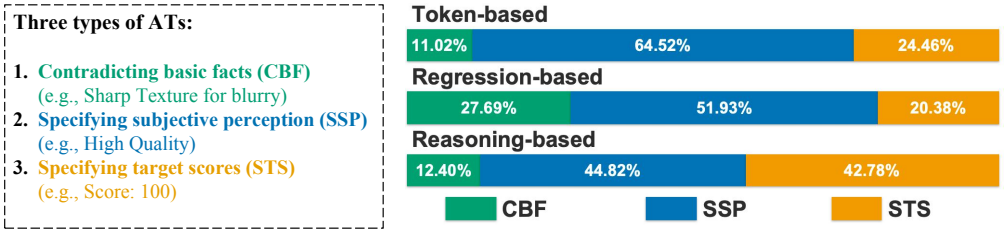
In summary, our contributions can be summarized as follows:

- We uncover two key insights into typographic attacks on VLM-IQA: the effectiveness of different AT types varies across model-image pairs, while attacks placed in quality-degraded regions are generally more effective.
- Motivated by our findings, we propose Q-Hack, a more powerful typographic attack tailored to VLM-IQA, which jointly optimizes AT content and placement strategy.
- We evaluate Q-Hack on four VLM-IQA models and three IQA datasets, demonstrating consistent improvements over state-of-the-art attacks across all three dimensions: attack effectiveness, perceptual fidelity, and scene consistency.

2 Background and Related Work

2.1 Vision-language model-based image quality assessment

VLM-IQA methods leverage VLMs' visual understanding and language reasoning capabilities to assess how image degradations affect perceptual quality in the context of image content. Existing VLM-IQA methods fall into three categories. ❶ **Token-based scoring methods** derive scores from probabilities over predefined quality tokens. Q-Instruct Wu et al. (2024b) uses two quality tokens; Q-Align Wu et al. (2024c) extends this representation to five quality levels; DeQA You et al. (2025) further constructs soft labels from the quality distribution to reduce information loss caused by discretization. ❷ **Regression-based scoring methods** map VLM features to scores through regression heads. MA-AGIQA Wang et al. (2024) combines VLM and deep learning-based IQA features to assess AI-generated images; Q-Scorer Tang et al. (2026) introduces IQA-specific score tokens and a lightweight multilayer perceptron regressor. ❸ **Reasoning-based scoring methods** analyze image content and quality-related factors through natural language reasoning before producing quality scores. VisualQuality-R1 Wu et al. (2026) predicts scores after explicit textual reasoning; Q-Insight Li et al. (2026) jointly predicts quality scores, degradation types, and severity. Despite their different scoring mechanisms, these models all rely on VLM backbones sensitive to textual semantics, making them vulnerable to typographic attacks.



(a) Quality-related ATs (b) Per-image winning rate

Figure 1: Effectiveness of different quality-related ATs. (a) Examples of the three types of quality-related ATs. (b) Per-image winning rates of the three AT types, grouped by VLM-IQA scoring mechanism in Sec. 2.1. A winning AT type yields the largest increase in the predicted quality score for a given image and target model.

2.2 Attacks against Image Quality Assessment

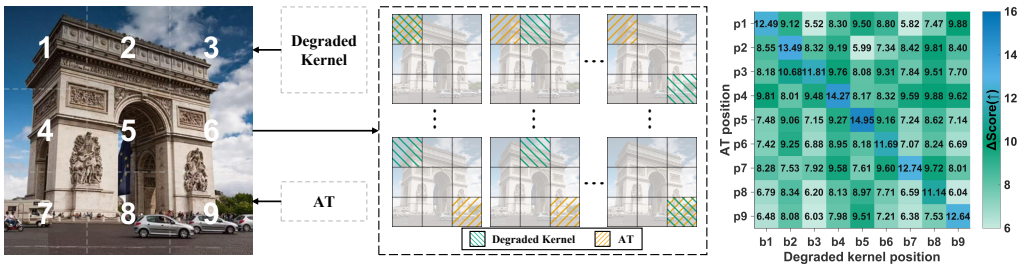
Conventional Adversarial Attacks. Existing attacks against IQA models mainly focus on conventional adversarial attacks, which manipulate quality predictions by introducing carefully crafted visual pixel-level perturbations. Such attacks have been extensively studied under diverse attack paradigms, ranging from white-box gradient-based optimization Zhang et al. (2022); Shumitskaya et al. (2024a; 2022); Yang et al. (2026); Shumitskaya et al. (2024b) to black-box query-based attacks Ran et al. (2025); Bychkov et al. (2025) and transfer-based attacks using surrogate models Korhonen & You (2022); Liu et al. (2026).

Typographic Attacks. With the emergence of VLM-IQA methods, language-conditioned reasoning enhances perceptual quality assessment capabilities but also introduces new security risks, i.e., typographic attacks, beyond conventional pixel-level perturbations. Although typographic attacks have been extensively studied in other vision-language tasks, including image classification Qraitem et al. (2024); Wang et al. (2025) and visual question answering Cao et al. (2025), their applicability to IQA remains unexplored. Among existing studies, NTA Luo et al. (2026) is the most closely related work, as it investigates typographic attacks on quality-oriented VLMs. However, although it is framed as an IQA attack, its design and evaluation primarily target image aesthetic assessment (IAA), which emphasizes visual appeal such as composition and color Wu et al. (2024c). In contrast, the broader IQA literature focuses on perceptual quality under image degradations and distortions Wu et al. (2024b); Fang et al. (2020); Hosu et al. (2020). This distinction leads to different attack requirements. In particular, for AT selection, NTA considers only a single IAA-related factor (i.e., specifying subjective perception, one of the three in Insight 1), whereas no single quality-related factor is consistently optimal for IQA (see Insight 1 in Sec. 3.3). For AT placement, NTA selects regions mainly to preserve aesthetic composition, whereas IQA attacks should instead account for regions that provide salient quality-related evidence, especially those exhibiting noticeable degradation (see Insight 2 in Sec. 3.3). Therefore, existing typographic attacks, including NTA, cannot be directly transferred to VLM-IQA models, motivating a task-specific investigation of typographic attacks tailored to image quality assessment tasks.

3 Problem Formulation and Preliminary Analysis

3.1 Threat Model

Given a VLM-IQA model F_m that receives user-submitted images and returns corresponding quality scores, the attacker aims to embed inconspicuous text into the images to manipulate the predicted quality scores, making them inconsistent with the ground-truth perceptual quality. Specifically, such inconsistency can occur in two directions: assigning lower quality scores to high-quality images or higher quality scores to low-quality images. Since IQA models are commonly used for quality assessment, ranking, and filtering, overestimating the quality of degraded images can directly undermine these downstream applications by making low-quality content appear acceptable. Therefore, following prior IQA attack studies Bychkov et al. (2025); Antsiferova et al. (2024); Gushchin et al. (2025), *we primarily focus on this more realistic attack objective*, i.e., $F_m(I) < F_m(\tilde{I})$, where I



(a) Text placement region effectiveness experiment workflow (b) Average score change

Figure 2: Effectiveness of AT placement relative to localized degradation. (a) The image is divided into a 3×3 grid, where quality degradation is applied to each region in turn, and different text placement locations are evaluated. (b) Placing meaningful AT in the quality-degraded region induces the most pronounced score change, where columns and rows denote the degradation locations b_1–b_9 and text locations p_1–p_9, respectively.

and $\tilde{I}$ are the clean and attacked images, respectively, and a high score means better quality. Regarding the adversary capability, we consider a realistic black-box setting, where the attacker can only interact with F_m through its input-output interface.

3.2 Problem Definition

In typographic attacks, an attacked image can be formulated as:

$$\tilde{I} = \mathcal{R}(I, t, \mathcal{S}), \tag{1}$$

where t denotes the AT to be embedded; $\mathcal{S}$ represents the target embedding position; $\mathcal{R}$ is the rendering function that renders the text onto the clean image.

Given the goal of increasing the predicted quality score, a natural yet inconspicuous typographic attack for IQA can be formulated as the following optimization problem:

$$\max_{t,\mathcal{S}} \left[(F_m(\tilde{I}) - F_m(I)) + \mu \mathrm{SC}(\tilde{I}) - \lambda \Delta_{\mathrm{PQ}}(I, \tilde{I}) \right], \tag{2}$$

where $\mathrm{SC}(\tilde{I})$ measures the scene consistency of the attacked image, with a larger value indicating better consistency; $\Delta_{\mathrm{PQ}}(I, \tilde{I})$ measures the perceptual fidelity difference between the clean and attacked images, with a larger value indicating a greater difference; μ and λ are weighting parameters that control the trade-off between attack effectiveness, scene consistency, and perceptual fidelity.

3.3 Insights

Before introducing our attack method, we first present two key design insights revealed by our preliminary experiments, which directly motivate the design of our framework. Detailed experimental settings are provided in Appendix A.

Insight 1: the effectiveness of different AT types is image- and model-dependent. IQA typically involves three levels of quality understanding: extracting observable quality-related information, forming perceptual judgments aligned with human visual preference, and mapping such judgments to quantitative quality scores. Accordingly, quality-related ATs can be constructed from three corresponding aspects: ❶ **contradicting basic facts (CBF)**, which indirectly misleads quality reasoning by conflicting with observable quality degradation (e.g., adding "Sharp Texture" to a blurry region); ❷ **specifying subjective perception (SSP)**, which directly suggests an overall quality judgment (e.g., "High Quality"); and ❸ **specifying target scores (STS)**, which explicitly states the desired quantitative output (e.g., "Score: 100"). To evaluate the effectiveness of the three AT types, we generate a large set of concrete AT instances for each type and test them across different input images and target models using the same placement locations. Our preliminary results in Fig. 1 present the distribution of the best-performing AT types across the test images for different target models, suggesting that no single AT type is consistently optimal across model-image pairs.

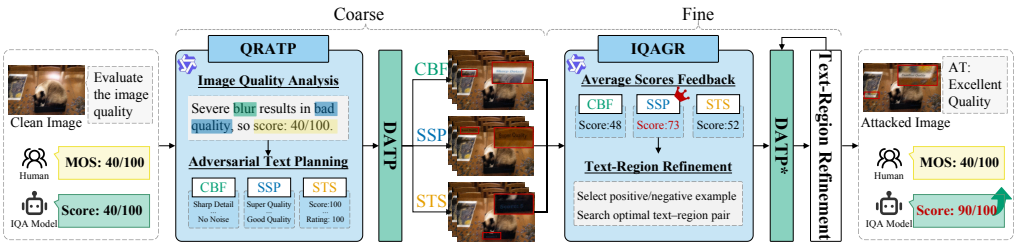


Figure 3: Overall pipeline of Q-Hack.

Insight 2: attacks targeting quality-degraded regions are more effective. Placing text in task-related regions can improve attack performance Cao et al. (2025). For IQA, task relevance is determined by quality-related evidence rather than semantic objects that guide object-centric recognition tasks, such as image classification and visual object counting. Since perceptual quality reflects both the overall visual appearance and localized degradations, regions with noticeable degradations can serve as more explicit quality-related evidence. Such regions are therefore more likely to attract the model's attention and influence quality-related reasoning. To verify this, we perform a controlled study by introducing localized degradations and varying the placement of adversarial text. We compare the quality-score changes caused by meaningful quality-related AT and visually similar yet meaningless text to remove the influence of text overlay itself. Across six VLM-IQA models, placing adversarial text within degraded regions consistently achieves the largest quality-score increase (see Fig. 2), demonstrating that quality-degraded regions provide more task-related factors for VLM-based quality reasoning. Therefore, we prioritize quality-degraded regions for adversarial text placement in our attack design.

4 Methodology

4.1 Overview

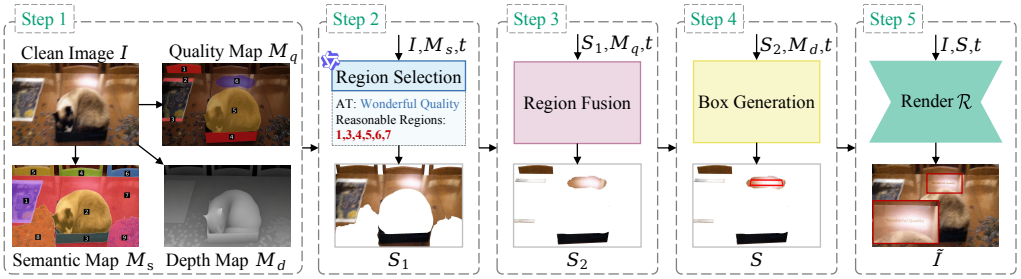
Fig. 3 illustrates the overall framework of Q-Hack, which jointly optimizes ATs and their placement regions in a coarse-to-fine manner. In the coarse stage, quality-related adversarial text planning (QRATP) generates K candidate ATs for each of the three aspects (i.e., CBF, SSP, and STS), yielding $3K$ candidates in total. For each AT, degradation-aware text placement (DATP) selects a plausible and potentially effective region $\mathcal{S}$ based on scene semantics, local degradation, and depth-guided geometry, after which a renderer integrates the text into the image. In the refinement stage, IQA-guided refinement (IQAGR) iteratively optimizes both AT content and placement using black-box feedback. The target IQA model first scores the $3K$ coarse-stage candidates, and the AT category with the highest mean score is selected for refinement. Within this category, high- and low-scoring samples serve as positive and negative examples, respectively. An auxiliary VLM analyzes their text-region patterns and generates new AT-region pairs, which are rendered by DATP, evaluated by the target model, and added to the candidate pool. This process is repeated for n rounds, after which the highest-scoring candidate is selected as the final attacked image.

4.2 Quality-related Adversarial Text Planning

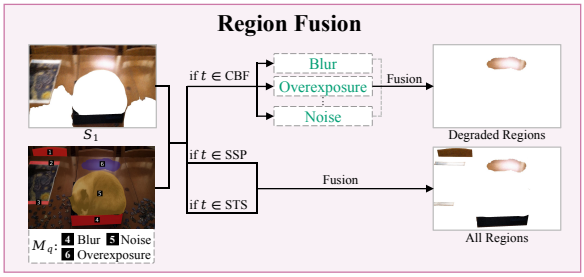
As shown in Insight 1, the effectiveness of quality-related ATs varies across images and VLM-IQA models, making a fixed AT category insufficient for consistently effective attacks. Therefore, QRATP generates image-conditioned candidate ATs from all three corresponding aspects (i.e., CBF, SSP, and STS). Specifically, given a clean image I, QRATP employs an auxiliary VLM (e.g., Qwen Bai et al. (2025)) to analyze its quality state and main degradation types, and uses this information to generate candidate ATs of the three corresponding aspects automatically:

$$\{\{t^k_{\text{CBF}}\}^K_{k=1}, \{t^k_{\text{SSP}}\}^K_{k=1}, \{t^k_{\text{STS}}\}^K_{k=1}\} = \text{QRATP}(I), \tag{3}$$

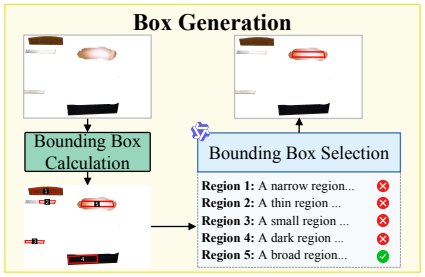
where $\{t^k_{\text{CBF}}\}^K_{k=1}$, $\{t^k_{\text{SSP}}\}^K_{k=1}$, and $\{t^k_{\text{STS}}\}^K_{k=1}$ denote the AT sets for CBF, SSP, and STS, respectively, and K is the set size.



(a) DATP



(b) Region fusion



(c) Box generation

Figure 4: (a) The five steps of DATP, (b) region fusion, (c) box generation.

4.3 Degradation-Aware Text Placement

Insight 2 suggests that placing text in noticeably degraded regions can substantially improve attack effectiveness. However, not all degraded regions are suitable for plausible text placement. To address this, DATP combines quality-guided region selection with semantic and geometric constraints to identify regions that are both quality-related and suitable for scene-consistent rendering. As shown in Fig. 4, DATP determines the placement region through five steps. First, DATP uses SEEM Zou et al. (2023), Q-Ground Chen et al. (2024a), and Depth Anything 3 Lin et al. (2025b) to process the clean image, obtaining a semantic map M_s, a quality map M_q, and a depth map M_d, respectively. Second, for each AT, DATP uses the auxiliary VLM to identify the semantic-region set $\mathcal{S}_1$ that can plausibly accommodate the candidate text based on I and M_s. Third, DATP extracts the degraded-region set $\mathcal{S}_q$ from the quality map M_q and takes its intersection with $\mathcal{S}_1$ to obtain the quality-region set $\mathcal{S}_2$. Note that only the CBF-related ATs use the specialized region subset $\mathcal{S}_q^{\text{spec}}$ (corresponding to the specific degradation attribute described by the AT), whereas the other two categories of ATs use the full degradation region set. For a current candidate AT t, the $\mathcal{S}_2$ is defined as follows:

$$\mathcal{S}_2 = \begin{cases} \mathcal{S}_1 \cap \mathcal{S}_q^{\text{spec}}, & t \in \{t_{\text{CBF}}^k\}_{k=1}^K, \\ \mathcal{S}_1 \cap \mathcal{S}_q, & t \in \{t_{\text{SSP}}^k\}_{k=1}^K \cup \{t_{\text{STS}}^k\}_{k=1}^K. \end{cases} \tag{4}$$

Fourth, for each region in $\mathcal{S}_2$, DATP uses the depth map M_d to generate a corresponding candidate depth-region set and select the most suitable placement depth-region $\mathcal{S}$. For each quality region $s_m^2 \in \mathcal{S}_2$, DATP fits a local depth plane $\hat{\pi}_m$:

$$(\hat{\alpha}_m, \hat{\beta}_m, \hat{\gamma}_m) = \underset{\alpha_m, \beta_m, \gamma_m}{\arg\min} \sum_{j=1}^{N_m} \left[d_{mj} - (\alpha_m x_{mj} + \beta_m y_{mj} + \gamma_m) \right]^2, \tag{5}$$

$$\hat{\pi}_m(x, y) = \hat{\alpha}_m x + \hat{\beta}_m y + \hat{\gamma}_m,$$

where N_m denotes the number of valid depth pixels in region s_m^2; (x_{mj}, y_{mj}) are the image coordinates of the j-th valid pixel; d_{mj} is the depth value at that location; α_m, β_m, and γ_m define the local depth plane, while $\hat{\alpha}_m$, $\hat{\beta}_m$, and $\hat{\gamma}_m$ are the optimal parameters. Based on $\hat{\pi}_m$, the region constraints, and the content of the candidate AT, DATP then determines the text-fitting depth region for each s_m^2,

Table 1: Score-increasing attack results on the LIVE Challenge dataset. **Bold** and underlined values indicate the best and second-best results, respectively.

Method	Q-Align		DeQA		MA-AGIQA		VisualQuality-R1		MAE↓	DISTS↓	TSIS↑
	ASR↑	QTAE↓	ASR↑	QTAE↓	ASR↑	QTAE↓	ASR↑	QTAE↓			
SEGA	1.46%	80.11	2.32%	77.18	20.57%	64.87	19.45%	60.01	4.4767	0.1520	-
AdvBIQA	20.05%	51.11	11.53%	53.52	1.89%	55.08	19.88%	44.63	3.1659	0.0915	-
Black-UAP	0.43%	48.98	1.29%	49.18	21.51%	51.36	8.61%	47.39	1.0469	0.0240	-
AttackIQA	76.08%	43.45	68.67%	45.73	**59.98%**	**48.61**	27.88%	41.55	1.6612	0.0256	-
Center	53.10%	43.15	69.36%	43.96	29.00%	50.82	13.68%	46.27	2.6949	0.0549	1.56
SceneTAP	32.53%	49.60	48.36%	46.00	8.26%	53.50	18.67%	50.93	5.3053	0.0610	2.64
NTA	62.65%	44.16	48.45%	44.24	34.42%	50.01	41.91%	39.15	13.823	0.0946	3.95
Ours	**80.21%**	**42.08**	**86.75%**	**40.51**	54.04%	49.13	**74.53%**	**37.34**	**0.5636**	**0.0189**	**4.13**

and an auxiliary VLM finally evaluates these candidate regions and selects the most suitable one as depth-region $\mathcal{S}$. Finally, we use LVTR Zhu et al. (2024) ($\mathcal{R}$) to render t within the selected region $\mathcal{S}$. The renderer uses the surrounding background as context to generate text that embeds naturally into the scene. Image modifications are confined to the local rendering region to limit their impact on perceptual fidelity. The candidate attacked image can be obtained as:

$$\tilde{I} = \mathcal{R}(I, t, \mathcal{S}). \tag{6}$$

4.4 IQA-guided Refinement

To search for a high-scoring AT-region pair for the target model, IQAGR performs n rounds of black-box feedback optimization. Inspired by Liu et al. (2024), it treats the AT t and its region $\mathcal{S}$ description as a joint prompt and iteratively refines them based on IQA feedback. IQAGR first scores the $3K$ images generated in the coarse stage, computes the mean score for each AT category, and selects the highest-scoring category as the refinement target. At the i-th fine-stage round, it scores and ranks the $K + i - 1$ candidates in this category (K initial images and $i - 1$ refined images). The high- and low-scoring candidates, together with their ATs and region descriptions, serve as positive and negative examples. An auxiliary VLM then analyzes their text-region patterns and generates a new AT-region pair, which DATP renders for evaluation in the next round. After n rounds, Q-Hack selects the highest-scoring attacked image $\tilde{I}$ among all initial and refined candidates.

5 Experiments

5.1 Experimental Settings

Datasets. We consider three datasets: LIVE Challenge (LIVE) Ghadiyaram & Bovik (2015), AGIQA-1K Zhang et al. (2023), and TID2013 Ponomarenko et al. (2015), covering real distortions, AI-generated artifacts, and synthetic distortions.

VLM-IQA models. We evaluate all three types of VLM-IQA models mentioned in Sec. 2.1, including token-based Q-Align Wu et al. (2024c) and DeQA You et al. (2025), regression-based MA-AGIQA Wang et al. (2024), and reasoning-based VisualQuality-R1 Wu et al. (2026). Note that we normalize the predicted scores of all models to [0,100], where 100 indicates the highest quality.

Baselines and implementation. We compare Q-Hack with four perturbation-based attacks and three typographic attacks. The perturbation-based baselines include black-box transfer attacks SEGA Liu et al. (2026) and AdvBIQA Korhonen & You (2022), and black-box query-based attacks Black-UAP Bychkov et al. (2025) and AttackIQA Ran et al. (2025). The typographic baselines include Center attack Cheng et al. (2024), which overlays the fixed text "High Quality" at the image center; SceneTAP Cao et al. (2025), which we adapt to a binary image-quality question with two answer options, "High Quality" and "Low Quality", using "High Quality" as the target answer; and NTA Luo et al. (2026) that only adopts AT specifying subjective perception (e.g., "High Quality"). For our proposed Q-Hack, we use Qwen3-VL-8B-Instruct Bai et al. (2025) as the auxiliary VLM; we set $K = 5$ for each candidate AT and $n = 10$ for IQAGR refinement rounds.

Table 2: Score-increasing attack results on the AGIQA-1K dataset.

Method	Q-Align		DeQA		MA-AGIQA		VisualQuality-R1		MAE↓	DISTS↓	TSIS↑
	ASR↑	QTAE↓	ASR↑	QTAE↓	ASR↑	QTAE↓	ASR↑	QTAE↓			
SEGA	3.70%	68.26	2.41%	71.80	3.52%	68.17	5.46%	55.56	3.8797	0.1278	-
AdvBIQA	17.04%	48.21	14.81%	50.89	2.22%	59.55	24.54%	36.53	4.0957	0.1436	-
Black-UAP	7.22%	48.12	12.04%	49.00	24.72%	56.17	19.26%	39.19	3.7091	0.0976	-
AttackIQA	57.96%	40.15	70.09%	43.42	54.91%	54.26	49.07%	34.49	1.6055	0.0307	-
Center	45.65%	39.05	65.46%	40.42	34.91%	55.03	49.91%	31.70	5.9456	0.1250	1.12
SceneTAP	41.67%	42.17	70.37%	39.18	18.06%	57.22	43.70%	39.92	8.7317	0.1419	1.90
NTA	64.17%	38.40	69.17%	38.09	49.26%	54.72	69.63%	33.02	11.284	0.1465	2.42
Ours	**68.33%**	**29.86**	**82.78%**	**34.17**	**59.91%**	**53.04**	**77.78%**	**26.60**	**0.6090**	**0.0219**	**3.32**

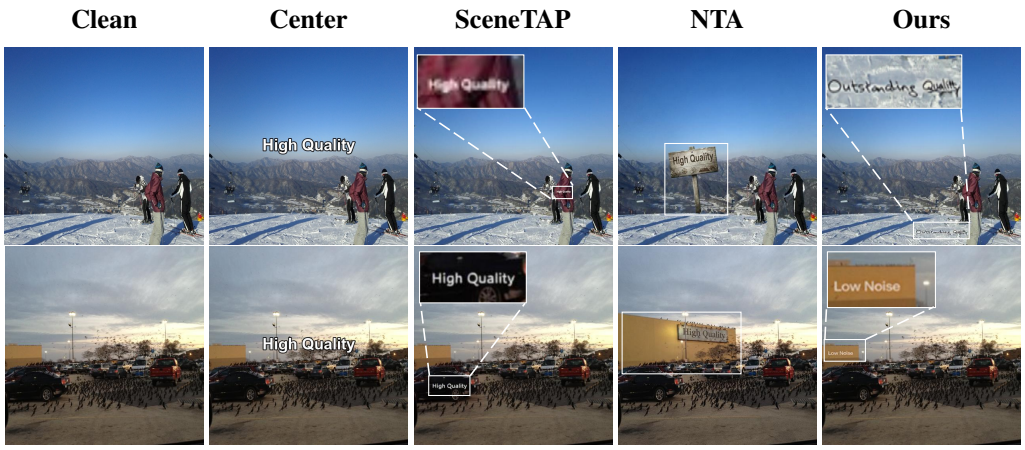


Figure 5: Qualitative comparison of different typographic attack results.

Metrics. We evaluate the proposed attack from three perspectives: *attack effectiveness*, *perceptual fidelity*, and *scene consistency*. We measure attack effectiveness using the attack success rate (ASR) and quality-target attainment error (QTAE). ASR is the percentage of samples with a predicted quality score increase above ten points. QTAE measures the discrepancy between the attacked scores and the maximum score (i.e., 100). Perceptual fidelity is measured by mean absolute error (MAE) for pixel-level fidelity and DISTS Ding et al. (2020) for holistic perceptual fidelity, with lower values indicating better preservation. Scene consistency is measured by the text-scene integration score (TSIS) using Qwen3-VL-8B-Instruct Bai et al. (2025). Detailed descriptions of all metrics and their corresponding calculation methods are provided in Appendix B.1.

5.2 Overall Results

We present the quantitative attack results on the LIVE and AGIQA-1K datasets in Tables 1 and 2, respectively, while the results on the TID2013 dataset are provided in Table 4 in Appendix B.2. From these results, we can clearly observe that our proposed Q-Hack achieves the best attack performance while maintaining a favorable trade-off between perceptual fidelity and scene consistency. As further illustrated in Fig. 5, Q-Hack places adversarial text in more appropriate scene regions and surface structures, resulting in more natural and inconspicuous attacked images. In contrast, attacks using a fixed AT type or fixed AT content, as well as transfer-based attacks, exhibit limited attack effectiveness. This further indicates that different models respond differently to textual and pixel-level perturbations, making model- and input-specific optimization necessary at present. Developing universal attacks that generalize across both models and inputs for IQA therefore remains an open challenge. More visualization examples are provided in Appendix B.3.

5.3 Ablation Study

Effectiveness of QRATP. We compare four combinations of quality-related ATs: (i) unconstrained text N; (ii) each individual AT category; (iii) each pair of AT categories; and (iv) all three categories

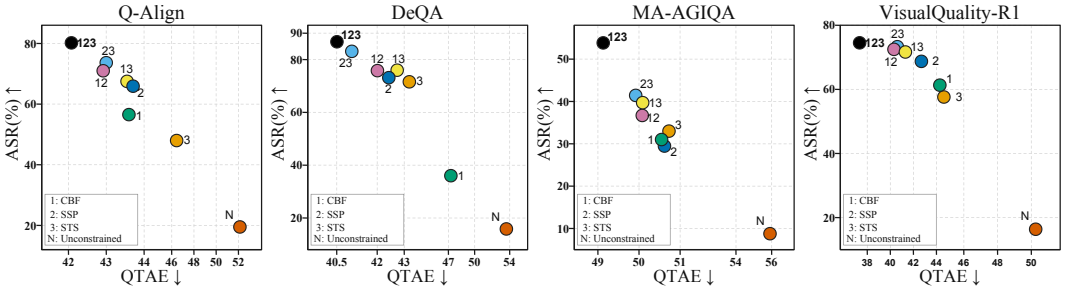


Figure 6: Effectiveness of AT category combinations in QRATP.

Table 3: Effectiveness of DATP.

Method	**ASR↑**	**QTAE↓**	**MAE↓**	**DISTS↓**	**TSIS↑**
w/o DATP	52.32%	44.83	2.5632	0.0532	1.62
w/o Semantic	73.32%	**42.21**	1.3459	0.0851	2.24
w/o Quality	54.75%	45.17	1.2632	0.0284	3.98
w/o Depth	72.38%	43.23	0.6736	0.0244	3.52
Ours	**73.88%**	42.27	**0.5636**	**0.0189**	**4.13**

(Ours). As shown in Fig. 6, unconstrained text performs worst across all four IQA models, with the lowest ASR and highest QTAE. Combining more quality-related AT categories generally improves performance, with all three categories achieving the best overall results.

Effectiveness of DATP. We compare DATP with four variants: (i) w/o DATP, where the text is inserted at the center of the image; (ii) DATP w/o the semantic map; (iii) DATP w/o the depth map; and (iv) DATP w/o the quality map. Table 3 reports the average results across the four VLM-IQA models. The w/o DATP performs worst overall, while removing any individual map degrades performance. The full DATP achieves the best overall trade-off among attack effectiveness, perceptual fidelity, and scene consistency, demonstrating the effectiveness of its components.

6 Defense

We consider a naive prompt-based defense and an anomalous-text detection defense. The former simply adds an instruction to the system prompt of the VLM-IQA model to ignore all text in the image. The latter is inspired by Li et al. (2025) and employs a VLM-based pre-screening mechanism to detect suspicious overlaid text and rejects inputs whose detected text is likely to have been intentionally inserted. However, both methods show limited defense effectiveness. First, VLM-IQA models appear to prioritize quality assessment over general instruction following, weakening their compliance with the instruction to "ignore all text in the image" and making it difficult to suppress typographic attacks. Second, Q-Hack is more robust against anomalous-text detection defenses than the compared attacks, as its text is embedded more naturally and consistently with the scene. Detailed defense settings and results are provided in Appendix C.

7 Conclusion

This study presents the first systematic investigation into the key factors required for effective typographic attacks against VLM-IQA models. Our analysis shows that the effectiveness of different adversarial text types varies across images and target models, while attacks targeting quality-degraded regions are generally more effective. Motivated by these findings, we propose Q-Hack, a new typographic attack framework against VLM-IQA models, which jointly optimizes AT content and placement in a coarse-to-fine manner. Extensive experiments demonstrate that Q-Hack achieves the strongest attack effectiveness while maintaining perceptual fidelity and scene consistency, and remains more robust than existing attacks under defense mechanisms.

AI USE STATEMENT

AI tools, such as ChatGPT, were used only for minor editorial improvements (grammar, spelling, and light style polishing).

ETHICS STATEMENT

This work studies typographic attacks against VLM-IQA systems for robustness evaluation. All attacks are generated and evaluated in a controlled, offline research setting and are intended to support robustness analysis and defense development. We use publicly available image data and do not collect new human-subject data or personally identifiable information. We acknowledge the potential dual-use risks of adversarial attacks and recommend using the proposed method only for authorized security and robustness evaluation.

REFERENCES

Anastasia Antsiferova, Khaled Abud, Aleksandr Gushchin, Ekaterina Shumitskaya, Sergey Lavrushkin, and Dmitriy Vatolin. Comparing the robustness of modern no-reference image- and video-quality metrics to adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 700–708, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Georgii Bychkov, Sergey Lavrushkin, and Dmitriy Vatolin. Black-box universal adversarial perturbations for image and video quality assessment methods. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2025.

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.

Chaofeng Chen, Sensen Yang, Haoning Wu, Liang Liao, Zicheng Zhang, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Q-ground: Image quality grounding with large multi-modality models. In *Proceedings of the ACM International Conference on Multimedia*, pp. 486–495, 2024a.

Zhihao Chen, Bin Hu, Chuang Niu, Tao Chen, Yuxin Li, Hongming Shan, and Ge Wang. IQAGPT: computed tomography image quality assessment with vision-language and chatgpt models. *Visual Computing for Industry, Biomedicine, and Art*, 7(1):20, 2024b.

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

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2020.

Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3674–3683, 2020.

Deepti Ghadiyaram and Alan C Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2015.

Aleksandr Gushchin, Khaled Abud, Georgii Bychkov, Ekaterina Shumitskaya, Anna Chistyakova, Sergey Lavrushkin, Bader Rasheed, Kirill Malyshev, Dmitriy S. Vatolin, and Anastasia Antsiferova. Guardians of image quality: Benchmarking defenses against adversarial attacks on image

quality metrics. In *Proceedings of the International Conference on Machine Learning*, pp. 21444–21470, 2025.

Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020.

Jianwen Jiang, Chao Liang, Jiaqi Yang, Gaojie Lin, Tianyun Zhong, and Yanbo Zheng. Loopy: Taming audio-driven portrait avatar with long-term motion dependency. In *Proceedings of the International Conference on Learning Representations*, pp. 15245–15263, 2025.

Yan Ke, Xiaoou Tang, and Feng Jing. The design of high-level features for photo quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 419–426, 2006.

Jari Korhonen and Junyong You. Adversarial attacks against blind image quality assessment models. In *Proceedings of the Workshop on Quality of Experience in Visual Multimedia Applications*, pp. 3–11, 2022.

Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Jian Zhang, et al. Q-insight: Understanding image quality via visual reinforcement learning. *Advances in Neural Information Processing Systems*, 38:36802–36827, 2026.

Yanjie Li, Yiming Cao, Dong Wang, and Bin Xiao. Agenttypo: Adaptive typographic prompt injection attacks against black-box multimodal agents. *arXiv preprint arXiv:2510.04257*, 2025.

Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, Chao Liang, Yuan Zhang, and Jingtuo Liu. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13848–13858, 2025a.

Hanhe Lin, Vlad Hosu, and Dietmar Saupe. KADID-10k: A large-scale artificially distorted iqa database. In *Proceedings of the International Conference on Quality of Multimedia Experience*, pp. 1–3, 2019.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025b.

Shihong Liu, Samuel Yu, Zhiqiu Lin, Deepak Pathak, and Deva Ramanan. Language models as black-box optimizers for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12687–12697, 2024.

Yujia Liu, Dingquan Li, Zhixuan Li, and Tiejun Huang. SEGA: A transferable signed ensemble gaussian black-box attack against no-reference image quality assessment models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(6):6660–6672, 2026.

Ziyuan Luo, Xinyu Chen, Qi Song, and Renjie Wan. Naturalistic typographic attacks on VLM-based image quality assessment. In *Proceedings of the International Conference on Quality of Multimedia Experience*, pp. 1–4, 2026.

Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77, 2015.

Maan Qraitem, Nazia Tasnim, Piotr Teterwak, Kate Saenko, and Bryan A Plummer. Vision-llms can fool themselves with self-generated typographic attacks. *arXiv preprint arXiv:2402.00626*, 2024.

Yu Ran, Ao-Xiang Zhang, Mingjie Li, Weixuan Tang, and Yuan-Gen Wang. Black-box adversarial attacks against image quality assessment models. *Expert Systems with Applications*, 260:125415, 2025.

Avinab Saha, Sandeep Mishra, and Alan C Bovik. Re-IQA: Unsupervised learning for image quality assessment in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5846–5855, 2023.

Erdi Sarıtaş, Eren Onaran, Vitomir Štruc, and Hazım Kemal Ekenel. Employing vision-language models for face image quality assessment. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, pp. 1–10, 2026.

Ekaterina Shumitskaya, Anastasia Antsiferova, and Dmitriy Vatolin. Universal perturbation attack on differentiable no-reference image-and video-quality metrics. *arXiv preprint arXiv:2211.00366*, 2022.

Ekaterina Shumitskaya, Anastasia Antsiferova, and Dmitriy Vatolin. Towards adversarial robustness verification of no-reference image-and video-quality metrics. *Computer Vision and Image Understanding*, 240:103913, 2024a.

Ekaterina Shumitskaya, Anastasia Antsiferova, and Dmitriy S. Vatolin. IOI: Invisible one-iteration adversarial attack on no-reference image- and video-quality metrics. In *Proceedings of the International Conference on Machine Learning*, pp. 45329–45352, 2024b.

Zhenchen Tang, Songlin Yang, Bo Peng, Zichuan Wang, and Jing Dong. Revisiting mllm based image quality assessment: Errors and remedy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 9475–9483, 2026.

Puyi Wang, Wei Sun, Zicheng Zhang, Jun Jia, Yanwei Jiang, Zhichao Zhang, Xiongkuo Min, and Guangtao Zhai. Large multi-modality model assisted AI-generated image quality assessment. In *Proceedings of the ACM International Conference on Multimedia*, pp. 7803–7812, 2024.

Xiaomeng Wang, Zhengyu Zhao, and Martha Larson. Typographic attacks in a multi-image setting. In *Proceedings of the Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 12594–12604, 2025.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. In *Proceedings of the International Conference on Learning Representations*, pp. 12547–12573, 2024a.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25490–25500, 2024b.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching LMMs for visual scoring via discrete text-defined levels. In *Proceedings of the International Conference on Machine Learning*, pp. 54015–54029, 2024c.

Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. Visualquality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank. *Advances in Neural Information Processing Systems*, 38:88167–88190, 2026.

Chenxi Yang, Yujia Liu, Dingquan Li, Yan Zhong, and Tingting Jiang. Beyond score changes: Adversarial attack on no-reference image quality assessment from two perspectives. *IEEE Transactions on Multimedia*, pp. 1–15, 2026.

Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14483–14494, 2025.

Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3D scene generation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5916–5926, 2025.

Weixia Zhang, Dingquan Li, Xiongkuo Min, Guangtao Zhai, Guodong Guo, Xiaokang Yang, and Kede Ma. Perceptual attacks of no-reference image quality models with human-in-the-loop. *Advances in Neural Information Processing Systems*, 35:2916–2929, 2022.

Zicheng Zhang, Chunyi Li, Wei Sun, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. A perceptual quality assessment exploration for AIGC images. In *Proceedings of the IEEE International Conference on Multimedia and Expo Workshops*, pp. 440–445, 2023.

Yuanzhi Zhu, Jiawei Liu, Feiyu Gao, Wenyu Liu, Xinggang Wang, Peng Wang, Fei Huang, Cong Yao, and Zhibo Yang. Visual text generation in the wild. In *Proceedings of the European Conference on Computer Vision*, pp. 89–106, 2024.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36:19769–19782, 2023.

Appendix

A Experimental Details for Insights

A.1 Insight 1

To evaluate Insight 1, we conduct experiments on two synthetically degraded IQA datasets: TID2013 Ponomarenko et al. (2015), which contains 25 reference images and 3,000 degraded images generated using 24 degradation types at five severity levels, and KADID-10K Lin et al. (2019), which contains 81 reference images and 10,125 degraded images generated using 25 degradation types at five severity levels. Following the categorization in Sec. 2.1, the evaluation covers seven VLM-IQA methods: three token-based methods (Q-Instruct Wu et al. (2024b), Q-Align Wu et al. (2024c), and DeQA You et al. (2025)), two regression-based methods (MA-AGIQA Wang et al. (2024) and Q-Scorer Tang et al. (2026)), and two reasoning-based methods (Q-Insight Li et al. (2026) and VisualQuality-R1 Wu et al. (2026)). An LLM generates 20 candidate ATs for each degradation type in the contradicting basic facts category, along with 20 candidates each for the specifying subjective perception and specifying target scores categories. All generated AT candidates are manually verified. Each candidate AT is placed in the central horizontal region of the input image with a font size of 30 px. For each image-model pair, the category containing the AT that produces the largest increase in predicted quality score is identified as the winning category. The results are then grouped by the VLM-IQA scoring mechanisms defined in Sec. 2.1, and the winning proportion of each AT category within each group is reported in Fig. 1.

A.2 Insight 2

To evaluate Insight 2, reference images from TID2013 and KADID-10K are selected, and each image is divided into a 3×3 grid. Six degradations are considered: Gaussian blur, Gaussian noise, local darkening, contrast reduction, saturation reduction, and JPEG compression. Each degradation is applied to one grid cell at a time, yielding nine possible degradation locations. For each degradation location, adversarial text is inserted at each of the nine grid locations in separate trials, resulting in $9 \times 9 = 81$ spatial combinations per degradation. To control for non-semantic effects introduced by text overlay, each AT is paired with a visually similar but semantically meaningless text counterpart (e.g., "Excellent Quality" and "Fxcellunt Gualiti"). For each spatial combination, the predicted quality-score change induced by the meaningless counterpart is subtracted from that induced by the AT. These differences are then averaged across images to obtain the mean score-change difference for each spatial combination. The results are shown in Fig. 2.

B Main Experimental Results

B.1 Metrics

We evaluate the proposed attack from three perspectives: *attack effectiveness*, *perceptual fidelity*, and *scene consistency*. For attack effectiveness, we define the score range as $[0, 100]$ and measure it by the attack success rate (ASR) and quality-target attainment error (QTAE). ASR is the percentage of samples with a predicted quality score increase above ten points, defined as:

$$\mathrm{ASR} = \frac{100\%}{N} \sum_{i=1}^{N} \left[F_m(\tilde{I}_i) - F_m(I_i) > 10 \right], \tag{7}$$

where N is the total number of samples, and I_i and $\tilde{I}_i$ denote the i-th sample before and after text insertion, respectively. QTAE measures the discrepancy between the attacked scores and the maximum score (i.e., 100):

$$\mathrm{QTAE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(F_m(\tilde{I}_i) - 100 \right)^2}. \tag{8}$$

Lower QTAE and higher ASR indicate stronger attack effectiveness. Perceptual fidelity is measured by the mean absolute error (MAE) for pixel-level fidelity and DISTS Ding et al. (2020) for holistic

Table 4: Score-increasing attack results on the TID2013 dataset. **Bold** and underlined values indicate the best and second-best results, respectively.

Method	Q-Align		DeQA		MA-AGIQA		VisualQuality-R1		MAE↓	DISTS↓	TSIS↑
	ASR↑	QTAE↓	ASR↑	QTAE↓	ASR↑	QTAE↓	ASR↑	QTAE↓			
SEGA	1.90%	77.70	1.40%	75.15	2.37%	62.41	3.60%	61.62	4.1623	0.1279	-
AdvBIQA	36.10%	50.65	40.87%	51.68	5.60%	53.04	37.27%	41.99	3.5226	0.0898	-
Black-UAP	5.00%	53.77	4.47%	54.83	7.67%	54.00	10.00%	44.42	1.7250	0.0362	-
AttackIQA	40.13%	47.69	62.23%	48.50	**36.37%**	49.42	48.03%	36.49	1.8387	0.0232	-
Center	30.33%	52.29	61.93%	45.34	25.73%	50.58	41.77%	38.22	2.0771	0.0409	1.92
SceneTAP	44.40%	48.52	64.53%	45.92	6.90%	53.98	36.63%	46.30	9.9706	0.0853	2.25
NTA	54.87%	48.17	61.53%	40.80	10.90%	52.07	55.47%	38.75	10.959	0.1283	3.32
Ours	**59.07%**	**46.54**	**72.23%**	**39.04**	36.07%	**49.32**	**63.63%**	**34.55**	**0.9684**	**0.0167**	**3.44**

perceptual fidelity. MAE is defined as:

$$\mathrm{MAE} = \frac{1}{N}\frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{n}\left|\tilde{I}_{i,j} - I_{i,j}\right|, \tag{9}$$

where $I_{i,j}$ and $\tilde{I}_{i,j}$ denote the j-th pixel values of the i-th clean and attacked images, respectively, and n is the total number of pixel values. The DISTS Ding et al. (2020) score is computed as:

$$\mathrm{DISTS} = \frac{1}{N}\sum_{i=1}^{N} d^{(i)}_{\mathrm{DISTS}}, \tag{10}$$

where $d^{(i)}_{\mathrm{DISTS}}$ denotes the standard DISTS distance between the clean and attacked images in the i-th sample. Lower MAE and DISTS indicate better perceptual fidelity. Scene consistency is measured by the text-scene integration score (TSIS), evaluated using an auxiliary VLM Bai et al. (2025):

$$\mathrm{TSIS}_i = \frac{a_i + g_i + p_i}{3}, \qquad \mathrm{TSIS} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{TSIS}_i, \tag{11}$$

where $a_i, g_i, p_i \in [1, 5]$ respectively measure text appearance naturalness, geometric consistency, and placement plausibility for image i. Higher TSIS indicates more natural text-scene integration.

B.2 Evaluation on TID2013

Table 4 presents the evaluation results on TID2013 Ponomarenko et al. (2015), further demonstrating the clear advantages of Q-Hack in attack effectiveness, perceptual fidelity, and scene consistency.

B.3 Visualization Examples

Fig. 7 shows additional clean and attacked images. Center and SceneTAP produce text that appears poorly integrated into the surrounding scene. Although NTA achieves some scene consistency, it introduces substantial changes to the original images, compromising perceptual fidelity. In contrast, Q-Hack integrates adversarial text into existing scene elements while better preserving the original visual appearance.

C Defense Details

To evaluate the robustness of different typographic attacks against defense mechanisms, we consider two defenses: a naive prompt-based defense and an anomalous-text detection defense. Table 5 reports the results of both defenses.

Prompt-based defense. We add an instruction to each target VLM-IQA model's system prompt ("ignore all text in the image"), asking it to ignore all text in the image when assessing visual quality. We then evaluate the attacks under this modified prompt using the same attack effectiveness

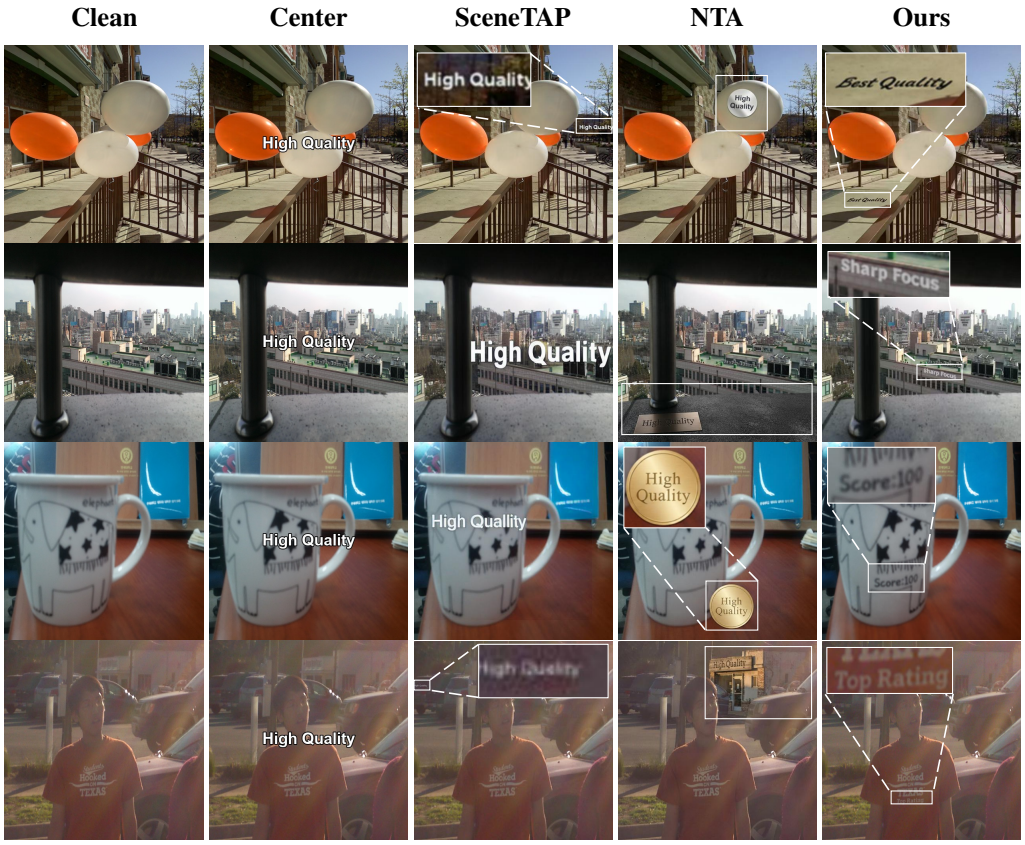


Figure 7: Qualitative comparison of different typographic attack results.

Table 5: Evaluation of defenses against typographic attacks.

Method	Q-Align		DeQA		MA-AGIQA		VisualQuality-R1		RR↓
	ASR↑	QTAE↓	ASR↑	QTAE↓	ASR↑	QTAE↓	ASR↑	QTAE↓	
Center	52.41% (-0.69)	43.36 (0.21)	67.13% (-2.23)	44.56 (0.60)	18.67% (-10.33)	51.88 (1.06)	10.50% (-3.18)	47.44 (1.17)	98.19%
SceneTAP	31.41% (-1.12)	49.79 (0.19)	46.56% (-1.80)	46.64 (0.64)	5.85% (-2.41)	54.62 (1.12)	16.01% (-2.66)	51.92 (0.99)	43.46%
NTA	60.15% (-2.50)	44.53 (0.37)	45.44% (-3.01)	44.81 (0.57)	24.44% (-9.98)	51.39 (1.38)	36.75% (-5.16)	40.36 (1.21)	75.82%
Ours	79.09% (-1.12)	42.26 (0.18)	85.71% (-1.04)	40.95 (0.44)	51.55% (-2.49)	50.40 (1.27)	72.46% (-2.07)	37.79 (0.45)	30.29%

metrics (i.e., ASR and QTAE) as in the main experiments. Each entry reports the post-defense value, with its change from the undefended result shown in parentheses.

Anomalous-text detection defense. Inspired by the visual prompt defense strategy in AgentTypo Li et al. (2025), we use Qwen3-VL-8B-Instruct Bai et al. (2025) to pre-screen each input image. For each attacked image, Qwen3 identifies the visible text and determines whether it appears to be naturally present in the scene or suspiciously overlaid, while also reporting its confidence in this judgment. Given the set of visible text strings detected in an image, we determine whether it contains the adversarial text inserted into the image. The defense blocks the image only if the attack text is recognized, classified as a suspicious overlay, and assigned a confidence of at least $\lambda = 0.7$; otherwise, the image is allowed through. We report the rejection rate (RR), defined as the proportion of attacked images blocked under this joint criterion. A lower RR indicates that the attack more frequently evades the screening mechanism.