PDA: TEXT-AUGMENTED DEFENSE FRAMEWORK FOR ROBUST VISION-LANGUAGE MODELS AGAINST ADVERSARIAL IMAGE ATTACKS

Jingning Xu, Haochen Luo, Chen Liu*
Department of Computer Science
City University of Hong Kong
{jingninxu3-c,chester.hc.luo}@my.cityu.edu.hk, chen.liu@cityu.edu.hk

ABSTRACT

Vision–language models (VLMs) are vulnerable to adversarial image perturbations. Existing works based on adversarial training against task-specific adversarial examples are computationally expensive and often fail to generalize to unseen attack types. To address these limitations, we introduce *Paraphrase–Decomposition–Aggregation (PDA)*, a training-free defense framework that leverages text augmentation to enhance VLM robustness under diverse adversarial image attacks. PDA performs prompt paraphrasing, question decomposition, and consistency aggregation entirely at test time, thus requiring no modification on the underlying models. To balance robustness and efficiency, we instantiate PDA as invariants that reduce the inference cost while retaining most of its robustness gains. Experiments on multiple VLM architectures and benchmarks for visual question answering, classification, and captioning show that PDA achieves consistent robustness gains against various adversarial perturbations while maintaining competitive clean accuracy, establishing a generic, strong and practical defense framework for VLMs during inference.

1 Introduction

Vision Language Models (VLMs) have recently achieved substantial success in both general-purpose multimodal systems [4, 19, 23, 49] and applications in specific domains, including medical imaging [2, 21, 42], autonomous driving [30, 37], and robotics control [17, 60]. Despite these advances, VLMs are shown susceptible to small, human-imperceptible image-space perturbations that induce large prediction shifts, posing outsized deployment risks compared with visible textual manipulations [9, 20, 24, 47].

Recently, many approaches are proposed to make VLMs robust against these adversarial image perturbations, including adversarial training [28, 34, 55, 58], adversarial prompt tuning [55, 58] and test-time tuning [36, 45]. Despite steady progress, these defense mechanisms have some practical limitations. (1) **Gradient access**: first-order methods [28, 34, 55, 58] rely on access to input gradients, which is infeasible for many closed-source VLMs where only output tokens can be obtained via APIs. (2) **Generalization to unseen attacks**: training-time approaches [28, 34, 55, 58] often use adversarial perturbations of specific types, such as ℓ_∞-bounded ones. This limits their generalizability, leaving models vulnerable to unseen perturbations or tasks [34, 55, 58]. (3) **Restrictive scope**: some methods [36, 45, 55, 58] rely on specific training objectives or architectural designs of the backbone model, making them unsuitable for general-purpose applications. (4) **Trade-offs with clean inputs**: several methods [28, 34] improve robustness at the expense of degrading performance on clean inputs [39], reducing the overall utility of robust models.

To address the concerns pointed above, we propose **Paraphrase-Decomposition-Aggregation (PDA)**, a generic, training-free framework that only needs black-box access to defend the VLMs against adversarial image perturbations. Our design is inspired by randomized smoothing [6] and its adaptation to large language models (LLMs) [15, 31], where several inputs are sampled in the neighborhood of each input data and the corresponding outputs are aggregated to

*Corresponding author

ensure stabilized decisions under adversarial attacks. In the context of VLMs, many adversarial attacks [50, 57] exploit the similarity between the image input and the text input to craft coordinated adversarial perturbations, suggesting that searching a local neighborhood in the high-dimensional text space can help recover semantics consistent with the image. Building on this intuition, PDA freezes model parameters and operates purely at test time via three steps: (i) **Paraphrase** the user query into multiple semantically equivalent views to exploit linguistic redundancy; (ii) **Decompose** the task into verifiable atomic questions that expose stable, image-aligned evidence; and (iii) **Aggregate** the answers with agreement-/confidence-aware voting to suppress adversarially biased views.

Our proposed framework PDA assumes black-box access to VLMs, it is agnostic to model architectures, tasks and perturbations types. Therefore, it can be applied to various applications in a plug-and-play manner. To assess generality and practicality, we evaluate the framework across common VLM use cases, including visual question answering, zero-shot classification, and image captioning, covering multiple model families and parameter scales. The extensive results demonstrate effectiveness of PDA: it yields consistent gains under adversarial image attack while largely preserving the performance with clean inputs. Our contributions are summarized as follows:

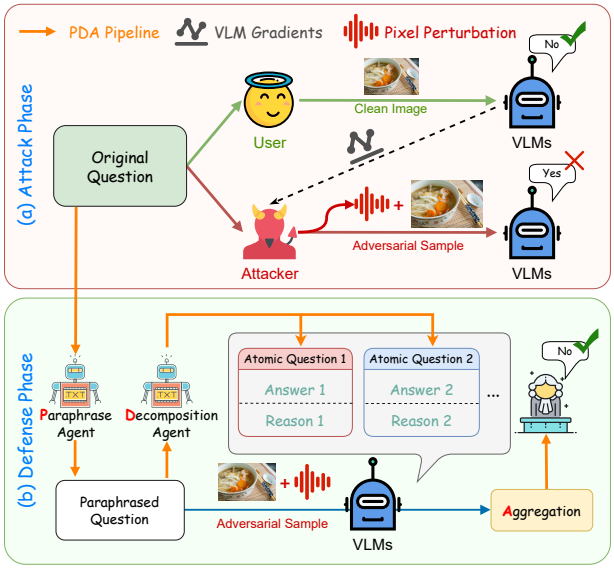

Figure 1: **Overview of the threat model and PDA.** *(a) Attack phase:* an adversary adds pixel-level perturbations guided by VLM gradients, causing the model to flip its answer under the original question. *(b) Defense phase:* our training-free, text-side Paraphrase–Decomposition–Aggregation pipeline, requires only black-box access, and suppresses adversarially biased views by cross-checking stable visual evidence.

- We propose **Paraphrase-Decomposition-Aggregation (PDA)** framework, which supports black-box access and is training-free. It establishes a generic paradigm to defend VLMs against adversarial image perturbations.
- We validate the effectiveness of PDA across a broad spectrum of applications, including different tasks and different model families with varying number of parameters and access controls. The extensive results show consistent robustness gains from PDA against various adversarial image perturbations with small performance degradations in clean inputs, demonstrating the broad applicability and effectiveness of our proposed framework.
- We further introduce several PDA variants that target different efficiency requirements and retain most of the robustness improvements. They substantially reduce the inference cost, yielding favorable accuracy–latency trade-offs across tasks and backbones.

2 Related Work

Attacks against Vision-Language Models. Similar to the uni-modal models [1, 8, 10, 27, 38], a growing body of evidence [9, 47, 50, 56, 57] shows that Vision-Language Models (VLMs) are vulnerable to many kinds of adversarial attacks. Small and human-imperceptible adversarial image perturbations can cause significant performance degradation for VLMs across multiple tasks. Adversarial attacks on pre-trained VLMs can be broadly categorized by the attacker's knowledge and the way multimodal inputs are perturbed. In the white-box setting, adaptations of classical image attacks such as Fast Gradient Sign Method (FGSM) [10], Projected Gradient Descent (PGD) [27] and AutoAttack [8] can be applied directly to the visual encoders. Under black-box constraints, AttackVLM [57] systematically evaluates transfer- and query-based targeted attacks on a wide range of instruction-tuned VLMs. Follow-up black-box methods further strengthen this threat surface by using pre-trained VL encoders to craft image–text pair perturbations and improve transferability and scalability, including VLATTACK [50], Chain of Attack [47], and AnyAttack [56]. All these observations underscore the increasing security concerns of adversarial visual inputs.

Defenses for Vision-Language Models. A large body of work aims to improve the adversarial robustness of CLIP-based VLMs, and existing defenses can be broadly divided into *training-time adversarial tuning* and *test-time adaptation*. At training time, adversarial training methods such as TeCoA [28] and FARE [34] adversarially fine-tune the CLIP vision encoder with text-guided contrastive or robust embedding objectives so that zero-shot robustness transfers to downstream VLMs. Beyond modifying the backbone, adversarial prompt tuning methods [44, 52, 55, 58] keep CLIP

encoders frozen and instead learn robust text prompts or attention modules. Overall, these training-time approaches defensively reshape the visual or language branch but still rely on generating adversarial examples and substantial offline optimization, which limits their applicability when data are restricted.

To avoid the need for additional data and retraining, recent work has begun to explore *test-time defenses* that adapt prompts at inference. Test-Time Adversarial Prompt Tuning (TAPT) [45] learns bi-modal defensive prompts per input by minimizing multi-view entropy and aligning clean–adversarial feature distributions. R-TPT [36] refines this paradigm with a reformulated entropy objective and reliability-weighted aggregation over augmented views, and C-TPT [51] further calibrates prompt updates using text-feature dispersion. Beyond prompt tokens, efficient test-time adaptation such as TDA [18] uses training-free dynamic adapters and key–value caches to reduce overhead under distribution shift. Beyond tuning-based defenses, several studies instead analyze how architectural choices and simple prompt patterns themselves affect robustness [3, 29]. Taken together, these test-time methods rely on adapter optimization over augmented views, are mostly evaluated on CLIP-style zero-shot classification, and still assume white-box gradients or internal representations, which limits their applicability in realistic black-box VLM deployments.

Randomized Smoothing. Randomized smoothing provides instance-wise robustness certificates by predicting with a noise-perturbed ensemble and assigning the class by majority vote under the noise distribution. Classical work establishes tight ℓ_2 guarantees for Gaussian smoothing and ImageNet-scale certification [7], and is further refined by adversarially trained smoothed classifiers, radius-maximizing objectives, and generalized smoothing measures [33, 48, 53]. Originally developed for image classification, randomized smoothing has since been extended to large language models. For example, SmoothLLM [31] perturbs prompts at the character level to stabilize large language models under jailbreak attempts, and SemanticSmooth [15] aggregates predictions over semantically transformed prompts to defend against stronger prompt-level attacks. More recently, smoothing ideas have been explored for vision–language models. PromptSmooth learns zero- or few-shot textual prompts so a fixed Med-VLM remains accurate under Gaussian image noise [13], and Open-Vocabulary Certification accelerates CLIP-style certification via incremental smoothing and caching for novel prompts [29]. However, these methods mostly smooth the visual or embedding space for specific architectures and classification tasks, without reshaping the full multimodal decision process or generated outputs, motivating randomized-smoothing-style defenses that act directly at the VLM decision level in more general black-box settings.

3 Methodology

3.1 Preliminary

Formulation. We use function $f_\theta(x; t) \to y$ to represent a Vision-Language Model (VLM) which is parameterized by θ and maps an image x and a text input t to an output y. Here, the text input t is typically a question or directive related to the image x, and the output y is the corresponding answer generated by the VLM. A correct and coherent response should capture the information from both the visual input and the textual input.

Adversarial Perturbation. Adversarial attacks on the visual input aim to perturb the image x to make the VLM generate incorrect outputs. It is formulated as an optimization problem to minimize a loss objective function $\mathcal{L}$ on the perturbation δ with a constraint on its size, usually based on its l_p norm:

$$\delta^* = \arg\max_{\delta} \mathcal{L}(f_\theta(x + \delta; t), y) \quad \text{subject to} \quad \|\delta\|_p \leq \epsilon \tag{1}$$

In this study, we assume a white-box threat model, meaning that the attacker has full access to the model, including its architecture, and parameters. This assumption allows us to evaluate our defense strategy under the most challenging conditions. The white-box attackers have access to the gradient of the loss with respect to the input perturbation, i.e. $\nabla_\delta \mathcal{L}$, which enables the use of gradient-based adversarial attack methods such as FGSM and PGD to efficiently approximate the optimal perturbation δ^* in Problem (1).

Defense Based on Text Perturbations. As multi-modal models, the output of VLMs depends on both visual and textual inputs. This dependency makes it possible to manipulate textual inputs to counteract the effects of adversarial image perturbations. Inspired by randomized smoothing [7], we expect that adversarial images often lie within the spike regions close to the decision boundary. By introducing small perturbations to the textual inputs without significantly altering their semantic meanings, we can aggregate the corresponding outputs to obtain smoother and thus more robust results.

To formally define these text perturbations, we introduce a perturbation function $\mathcal{T}(\cdot)$, which maps the original text input t to the perturbed text $t' = \mathcal{T}(t)$. In this context, $\mathcal{T}$ paraphrases the original text t, aiming to alter the literal expression of the text t while maintaining its underlying semantic meanings. Compared with existing works [16, 31] which apply

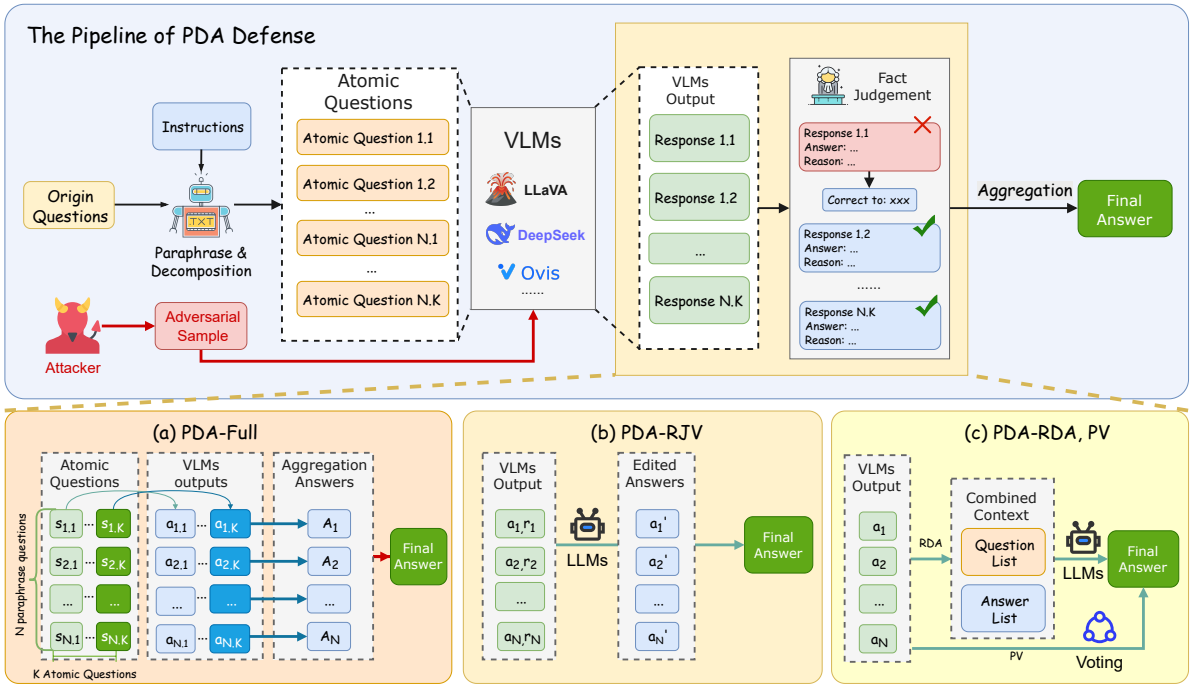

Figure 2: **Pipeline of the proposed PDA defense and its variants.** Given an image x and a query t, the *Paraphrase Agent* produces semantically equivalent prompts. Each prompt is factorized by the *Question Decomposer* into atomic questions that are individually answered by the base VLM. The *Answer Aggregation Agent* performs confidence- and consistency-aware fusion to yield the final prediction. The bottom panels illustrate four PDA variants: (a) **PDA-Full**, (b) **PDA-RJV**, (c) **PDA-RDA** and **PDA-PV**.

randomized smoothing in uni-modal large language models, we consider a richer family of text perturbations. In addition to character-level edits and synonym replacements, we include sentence-level semantic transformations in $\mathcal{T}$, including sentence rephrasing by grammatical restructuring and decomposition of a single complex query into several simpler atomic questions, as shown in Figure 3. These transformations allows us to exploit more high-level, complex and interpretable features extracted by VLMs, which is robust against small and imperceptible image perturbations, since extensive literature [14, 32, 40, 43] demonstrate that adversarial perturbation usually alters high-frequency, low-level and uninterpretable features.

3.2 PDA Defense Framework

Based on the motivations above, we introduce *Paraphrase-Decomposition-Aggregation (PDA)* as a generic, training-free defense framework that only needs black-box access to improve the robustness of VLMs against adversarial image perturbations. As the name indicates, PDA consists of three stages: (1) paraphrase the text inputs to obtain several semantically similar and literally distinct prompts; (2) decompose each prompt into a sequence of several atomic questions; (3) feed atomic questions to VLMs and aggregate the corresponding output to obtain reliable final answers. The overall pipeline is demonstrated in Figure 2 and we discuss the details of each stage below.

Step I: Paraphrase. Paraphrase aims to explore the "neighborhood" in the semantic space of the original textual input t. In this stage, we employ a paraphrase agent $\mathcal{T}_1$, usually a large language model (LLM), to generate a finite set of candidate paraphrases:

$$\mathcal{S} = \{t_1, \ldots, t_n\}, \qquad t_i \sim \mathcal{T}_1(t; \phi) \tag{2}$$

where $\mathcal{T}_1(t; \phi)$ denotes the paraphrase generation distribution and ϕ denotes the prompt. The prompt contains the detailed information for paraphrase, which controls the diversity of the output texts. A higher degree of diversity encourages more aggressive rephrasings and structural variants, while a more deterministic outputs keeps the paraphrases closer to the original wording.

The set $\mathcal{S}$ we obtain in (2) is the collection of admissible text transformations applied to the original textual input. Therefore, the responses from VLMs for the paraphrased texts in $\mathcal{S}$ will establish an *Expectation over Text Transformations* (EoTT) estimator for robust inference, which is analogous to the Expectation over Transformations (EOT) principle [1, 12, 35, 41] commonly used in adversarial vision, but instantiated in the textual domain.

Step II: Decomposition. The success of chain-of-thought prompting [46] indicates that it is more reliable to solve a problem as a series of intermediate steps than directly giving a final answer. Similarly, we decompose the paraphrased question obtained in Step I into a sequence of simpler and atomic questions for more reliable inference. By exploring the answer at the level of finer-grained evidence instead of a single holistic answer, we explicitly make the reasoning process more interpretable, which is shown beneficial to robustness [14, 39]. In the context of VLMs, we reduce a potentially complex, multi-hop query into a sequence of small atomic questions, each targeting a single visual fact like the presence of an object, its attributes, spatial relations, counts, or simple comparisons.

Formally, for each paraphrase t_i obtained in Step I, we employ a decomposition agent $\mathcal{T}_2$, usually a large language model (LLM), to explicitly construct a finite sequence of atomic questions:

$$\mathcal{S}_i \;=\; \{s_{i,1}, \ldots, s_{i,k_i}\} \;\sim\; \mathcal{T}_2(t_i, \varphi), \tag{3}$$

where $s_{i,j}$ is an atomic question and φ is the prompt containing the decomposition instructions for generating the sequence. The prompt φ should ensure that the decomposition is *sufficient* and *faithful*. Specifically, when querying the VLMs with the sequence of atomic questions, the responses $\{a_{i,j} = f_\theta(x; s_{i,j})\}_{j=1}^{k_i}$ should be sufficient to construct a meaningful answer to question t_i. In addition, each atomic question should refer only to evidence that is in principle available in the image x, avoiding the introduction of extraneous assumptions or off-image knowledge. We provide the prompt example to fulfill these two requirements in the supplemental materials.

Step III: Aggregation. In Step I and Step II, we do not use the image for paraphrase and decomposition, since the potential adversarial image perturbation may affect the construction of atomic questions. In Step III, we feed these sequences of atomic questions into the VLM and obtain the corresponding responses. We use the evidence set $\mathcal{B}_i \;=\; \{(s_{i,j}, a_{i,j})\}_{j=1}^{k_i}$, to represent the question-answer pairs for paraphrase t_i, and then entire collection for different paraphrases is $\mathcal{B} = \cup_{i=1}^{n} \mathcal{B}_i$. We use the aggregate agent in Step III to obtain the final results through two-level aggregation: it first summarizes $\mathcal{B}_i$ into a single candidate answer for t_i, then summarizes the answers for different paraphrases to generate the final results.

For structured tasks with a closed candidate set $\mathcal{Y}$ (e.g., multiple choices in VQA or image classification), we instantiate this agent as a lightweight decision head that maps each evidence set to a label $g(\mathcal{B}_i) \;=\; \widehat{y}_i \in \mathcal{Y}$, where g can be implemented as a simple voting rule, such as simple majority, over the sub-answers in $\mathcal{B}_i$. The same rule can be applied to aggregate results from $\mathcal{B}_i$ to obtain the final results for $\mathcal{B}$. For open-form tasks such as image captioning, we use an LLM h as the aggregation agent to directly generate the summary $\widehat{a} = h(x, t, \mathcal{B}, \Phi)$ as the final results. Here, Φ is the prompt describing how the responses in $\mathcal{B}$ are organized and potential task-specific information. In this way, Step III concludes the pipeline of PDA, establishing a generic, training-free and black-box defense framework. The pseudo-code is demonstrated in Algorithm 1.

3.3 Budgeted PDA Variants.

PDA establishes a generic three-step framework for robust inference against adversarial image perturbations. Similar to randomized smoothing, there is an effectiveness-complexity trade-off under PDA framework. In this context, we instantiate four pipelines that differ in where aggregation occurs and how paraphrasing and decomposition are coupled. We summarize the complexity of these variants in Table 1. In the supplemental materials, we provide some concrete examples to demonstrate their workflows.

(1) PDA-Full. This variant executes the full chain. A single LLM is employed to paraphrase the original question t into N paraphrases, and each paraphrase into up to K atomic questions. The atomic questions are fed to VLMs together with the potentially adversarial image input. The VLM responses are then filtered at the probe level based on their consistency. The LLM is employed again in the final aggregation step to produce the decision.

(2) PDA-RJV (Reason–Judge–Vote). This variant skips decomposition step. The VLM directly answers each of the N paraphrases (but with a brief rationale). A small LLM is then employed to *judge* the N paraphrase-level answers one

Table 1: Computation budget by stage (LLM: text-only; VLM: vision–language). Here M is the number of retained paraphrases and K is the number of atomic questions per paraphrase (generated within a single VLM call per paraphrase).

Variant	Paraphrasing	Decomposition		Aggregation
	LLM	VLM	LLM	LLM
PDA-Full	1	$N \times K$	$N \times K + N$	1
PDA-RJV	1	N	N	0
PDA-RDA	1	N	0	1
PDA-PV	1	N	0	0

Algorithm 1 PDA-VLM: training-free PDA wrapper for robust inference with a frozen VLM f_θ

Require: Image x, query t; paraphrase budget n; task type (structured with candidate set $\mathcal{Y}$, or open-form).

// Step I: Paraphrase
Sample candidate paraphrases $\{t_i\}_{i=1}^{n} \leftarrow \mathcal{T}_1(t;\phi)$.
for each retained paraphrase t_i **do**
  // Step II: Decomposition
  $\mathcal{S}_i \leftarrow \mathcal{T}_2(t_i,\varphi)$ {obtain atomic questions}
  for each $s_{i,j} \in \mathcal{S}_i$ **do**
    $a_{i,j} \leftarrow f_\theta(x; s_{i,j})$ {answer atomic question}
  end for
  $\mathcal{B}_i \leftarrow \{(s_{i,j}, a_{i,j})\}_{j=1}^{k_i}$ {evidence for paraphrase t_i}
end for
// Step III: Aggregation
if task is structured with closed candidate set $\mathcal{Y}$ **then**
  for each paraphrase t_i **do**
    $\widehat{y}_i \leftarrow g(\mathcal{B}_i) \in \mathcal{Y}$ {e.g., voting over mapped sub-answers}
  end for
  Compute evidence scores $E(y) \leftarrow \sum_i \mathbb{I}[\widehat{y}_i = y]$ for all $y \in \mathcal{Y}$.
  return $\widehat{y}^\star \leftarrow \arg\max_{y\in\mathcal{Y}} E(y)$
else
  $\mathcal{B} \leftarrow \bigcup_i \mathcal{B}_i$ {collect all atomic question–answer pairs}
  $\widehat{a} \leftarrow h(x,t,\mathcal{B})$ {text-only LLM aggregates into a free-form answer}
  return $\widehat{a}$
end if

by one and summarize the final decision based on the weighted vote across these paraphrase-level answers. Compared with PDA-Full, PDA-RJV employs LLM as judges and apply voting aggregation on the paraphrase level, cutting a significant number of LLM calls while letting the judge correct possible incorrect answers using cross-paraphrase evidence.

(3) PDA-RDA (Reason–Direct–Aggregate). This variant skips the decomposition step and VLM answers each of the N paraphrases with a brief rationale. Compared with PDA-RJV, PDA-RDA directly employs an LLM to jointly reviews all $(\text{paraphrase}, \text{answer}, \text{rationale})$ tuples and outputs the final decision *in one shot* by a single call in aggregation step. It replaces N paraphrase-level judgements with one global pass, reducing the number of LLM calls while assessing the VLM responses from a global perspective.

(4) PDA-PV (Paraphrase–Vote). This variant targets the structured tasks with a closed candidate set. It skips the decomposition step and further integrate the paraphrase and aggregation steps. It generate paraphrases that are not necessary semantically equivalent but logically equivalent to pick the correct answer. For example, paraphrasing "t-shirt / jeans" to "upper wear / lower wear" is not semantically equivalent, but such paraphrase preserves the correct choice. Because of larger semantic perturbations, PDA-PV increases the diversity of paraphrases. It reduces the number of LLM calls to N paraphrases and one voting aggregation.

4 Experiments

4.1 Experimental Setup Overview

Datasets and Metrics. We adopt VQA-v2 [11], ImageNet-D [54], and MS COCO [5] with their community-standard metrics to target three complementary facets of robustness. VQA-v2 is used to probe open-ended multimodal reasoning under reduced language priors, so improvements under attack are less attributable to textual shortcuts. ImageNet-D evaluates recognition under controlled generative shifts that preserve category semantics while perturbing nuisance factors, offering a focused stress test of visual backbones and VL pipelines. For COCO captioning, following recent robustness evaluations [57], we report *CLIPScore* to quantify the similarity between generated captions and ground-truth descriptions. Accordingly, we use *VQA Accuracy* on VQA-v2, *Top-1 Accuracy* on ImageNet-D, and *CLIPScore* on COCO, aligning our protocol with prevailing practice and enabling direct comparability with prior robustness studies.

Models. Primary evaluations use *LLaVA-1.5-7B* and *LLaVA-1.5-13B* [22], widely adopted LVLMs with strong instruction-following capability. To test plug-and-play generality across architectures and parameter scales, we additionally evaluate *DeepSeek-VL-1.3B* [25], *InternVL3-2B* [59], and *Ovis2-4B* [26]. These families differ in visual

encoders, connector designs, and training recipes, providing a diverse set of backbones to examine whether a training-free text-side wrapper transfers without re-training.

Threat model and Baselines. We evaluate robustness under both white-box and black-box adversarial settings. For white-box evaluation, we adopt PGD as the standard attack [27], and further consider adaptive attacks based on expectation over transformations (EOT), which explicitly optimize over PDA's stochastic paraphrase and decomposition pipeline. For black-box evaluation, we include representative transfer- and query-based attack settings, including the transfer-based protocol in [9], *AttackVLM*, and *AnyAttack* [56, 57]. As text-side randomized-smoothing baselines, we include *SmoothLLM* with Swap/Insert/Patch perturbations [31] and *SemanticSmooth* with paraphrase/summarise transformations [16]. As training-time robustification references, we evaluate released checkpoints from *TeCoA* and *FARE* [28, 34]. Our method (*PDA*) is training-free and operates solely on the textual interface; all methods are compared under the same threat model and data splits for fairness.

PDA Configurations. Unless otherwise specified in the corresponding subsection, the primary comparison results use the **PDA-Full** pipeline with `GPT-4.1-mini` as the default LLM agent. For VQA-v2 and ImageNet-D, PDA-Full uses 5 paraphrases per query, and each paraphrase is further decomposed into 3 atomic sub-questions, yielding 15 sub-questions in total. For each paraphrase, the answers to its three sub-questions are aggregated by the LLM into one paraphrase-level answer, and the final prediction is obtained by majority voting over the five paraphrase-level answers. For COCO captioning, we use 2 paraphrases and 5 atomic questions to better probe object/attribute grounding while keeping decoding cost moderate. SmoothLLM and SemanticSmooth both aggregate 10 randomized views by majority/score voting. For evaluation of TeCoA and FARE, the evaluation is run by direct inference under PGD attack with the public weights provided by [34].

we also study several lighter-weight variants in the ablation studies section. *PDA-RJV* keeps the same paraphrase width (5) but removes decomposition: for each paraphrase, the victim VLM outputs an answer and a rationale, and the LLM judges each paraphrase independently; the final answer is determined by majority voting over the five judgments. *PDA-RDA* also removes decomposition while keeping five paraphrases, but instead of judging each paraphrase separately, it feeds all five paraphrase–answer–rationale tuples to the LLM in a single pass to directly produce the final prediction. *PDA-PV* is the lightest configuration: it uses five logically equivalent paraphrases, removes both decomposition and LLM-based aggregation, queries the VLM once per paraphrase without requiring rationales, and obtains the final prediction by majority voting over the five VLM answers. Unless explicitly specified in the corresponding subsection, the main comparison tables use PDA-Full, while variant-specific results are reported separately in the ablation studies.

4.2 Robustness against White-box Attacks

Standard PGD. We start with PGD as the white-box attack, using a perturbation budget $\epsilon = 2/255$. Across the three tasks and both LLaVA scales, *PDA* markedly increases adversarial accuracy while keeping clean accuracy essentially unchanged (Table 2). On VQA-v2, ADV rises to 0.754 (7B) and 0.787 (13B); on ImageNet-D we observe higher clean accuracy alongside large adversarial gains (e.g., 7B clean 0.758 vs. 0.625). This behavior matches the mechanism of PDA's *paraphrase–/decompose–aggregate* pipeline: atomic questions cross-check stable visual evidence and aggregation self-corrects spurious predictions that would otherwise contaminate both robust and nominal outputs. The qualitative examples in Fig. 3 further illustrate this effect: for classification, PDA overturns an incorrect "jeans" decision by paraphrasing the main-object query into checks on prominence, region, and garment type and then voting for "t shirt"; for captioning, PDA decomposes the prompt into factual questions (object count, color, pose) and aggregates the answers into a more faithful caption with a higher CLIP score. Additional qualitative cases are provided in the supplemental materials.

Compared with text-side randomized-smoothing baselines, SmoothLLM and SemanticSmooth yield only modest robustness changes and sometimes degrade clean performance; this is consistent with recent observations that character-level noise or shallow synonym substitutions often induce small movements in the model's semantic space for modern instruction-tuned LLMs, limiting downstream effect when applied naively to VLM prompts. Finally, training-time defenses typically report stronger robustness but can exhibit a clear robustness–utility trade-off on clean inputs. In contrast, PDA achieves large adversarial accuracy improvements without any additional training or parameter updates and requires only black-box access at inference, providing a low-cost and architecture-agnostic alternative to fine-tuning–based defenses.

Larger Perturbation Budgets. We further stress-test PDA under more challenging white-box settings that go beyond the default PGD attack, focusing on both larger perturbation budgets. We evaluate robustness by sweeping the PGD perturbation budget $\epsilon \in \{2, 4, 6, 8\}/255$. This experiment tests whether the robustness gains of PDA are confined to a narrow perturbation regime or persist as the perturbation magnitude increases. As shown in Table 3(a), the no-defense baseline degrades steadily as ϵ grows on both ImageNet-D and VQA-v2, whereas PDA remains comparatively stable

Table 2: **Overall comparison on three tasks.** Results for VQA-v2 (*VQA Accuracy*), ImageNet-D (*Top-1 Accuracy*), and MS COCO (*CLIPScore*). Rows list defenses applied to each victim model (LLaVA-1.5-7B/13B), including SmoothLLM (Swap/Insert/Patch), SemanticSmooth (Paraphrase/Summarise), TeCoA, FARE, and our **PDA**. *Clean* denotes performance on unperturbed inputs; *ADV* denotes performance on adversarially perturbed inputs under the same evaluation protocol.

Defense Method	LLaVA-1.5-7B						LLaVA-1.5-13B					
	VQA-v2		ImageNet-D		COCO		VQA-v2		ImageNet-D		COCO	
	Clean	ADV	Clean	ADV	Clean	ADV	Clean	ADV	Clean	ADV	Clean	ADV
No Defense	0.862	0.445	0.625	0.248	0.858	0.753	**0.886**	0.593	0.646	0.320	**0.865**	0.757
SmoothLLM-Swap	0.866	0.453	0.610	0.260	0.849	0.744	0.881	0.596	0.630	0.327	0.849	0.756
SmoothLLM-Insert	0.870	0.449	0.596	0.254	0.854	0.746	0.884	0.587	0.605	0.330	0.855	0.743
SmoothLLM-Patch	0.864	0.454	0.632	0.261	0.843	0.741	0.877	0.599	0.611	0.314	0.851	0.751
SemanticSmooth-Paraphrase	0.850	0.452	0.638	0.276	0.850	0.748	0.884	0.608	0.657	0.358	0.855	0.751
SemanticSmooth-Summarise	0.846	0.456	0.610	0.289	0.851	0.750	0.870	0.613	0.649	0.379	0.861	0.754
TeCoA	0.787	0.689	0.613	0.569	0.811	0.783	0.830	0.672	0.697	0.611	0.823	0.791
FARE	0.828	0.703	0.640	0.511	0.835	**0.806**	0.838	0.729	0.695	**0.625**	0.833	**0.805**
PDA	**0.871**	**0.754**	**0.758**	**0.570**	**0.860**	0.774	0.879	**0.787**	**0.798**	0.566	0.862	0.778

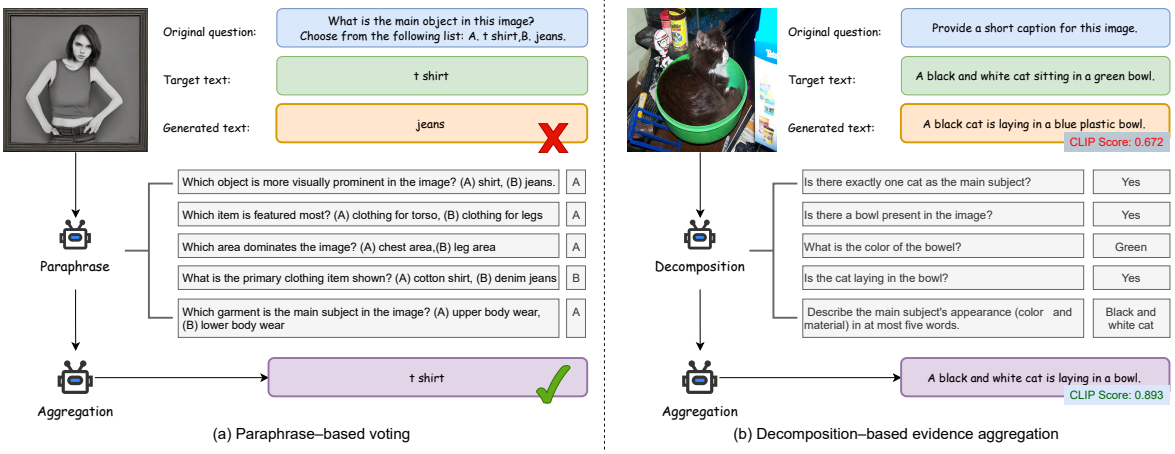

Figure 3: **Qualitative effect of PDA.** (a) The base VLM is attacked into choosing "jeans" in a two-way "t shirt vs. jeans" question, whereas PDA paraphrases the query into targeted checks and recovers the correct "t shirt" label. (b) For a cluttered scene with a black-and-white cat in a green bowl, PDA decomposes the caption prompt into factual sub-questions and aggregates the answers into a faithful caption.

across budgets. These trends indicate that PDA is not tuned to a single attack budget and continues to provide substantial robustness gains even under stronger white-box perturbations.

Adaptive Attacks. In addition to larger adversarial perturbations, we further evaluate robustness against adaptive attacks that explicitly target the stochastic components of PDA, since PDA relies on stochastic paraphrase and decomposition, a vanilla white-box PGD attacker that differentiates through only a single sampled view may be mismatched to the defense. Therefore, we adopt expectation over transformations (EoT) [1] and optimize the expected loss over PDA's stochastic components:

$$\max_{\delta} \; \mathbb{E}_{p,d}[\mathcal{L}(f_{\text{PDA}}(x+\delta; p, d), y)] \,, \tag{4}$$

where the expectation over sampled paraphrases p and decompositions d is approximated by Monte Carlo samples during optimization. We evaluate three variants that isolate different sources of randomness: *EOT-P* (expectation over paraphrases), *EOT-D* (expectation over decompositions), and *EOT-PD* (joint expectation over both).

As shown in Table 3(b), these adaptive attacks are consistently stronger than vanilla PGD, reducing defended accuracy on both ImageNet-D and VQA-v2. However, PDA does not collapse under any EOT variant and retains a clear robustness margin over the no-defense baseline, indicating that its gains are not merely due to attacker mismatch against

Table 3: Robustness under more challenging white-box attacks on ImageNet-D and VQA-v2.

Setting	No Defense		PDA	
	ImageNet-D	VQA-v2	ImageNet-D	VQA-v2
(a) Larger ϵ under PGD				
$2/255$	0.25	0.45	0.57	0.75
$4/255$	0.22	0.42	0.56	0.74
$6/255$	0.19	0.40	0.53	0.74
$8/255$	0.17	0.39	0.54	0.72
(b) Adaptive EOT *attacks*				
PGD	0.25	0.45	0.57	0.75
EOT-P	0.20	0.41	0.53	0.72
EOT-D	0.23	0.40	0.55	0.70
EOT-PD	0.21	0.42	0.54	0.74

a stochastic pipeline. Overall, agreement-based aggregation over diversified textual views remains effective even when the attacker explicitly targets the view distribution induced by PDA.

4.3 Robustness against Black-box Attacks

General Black-box Attacks. Besides white-box robustness, black-box robustness is also critical in practice because many deployed LVLMs are accessed through closed APIs where gradients and model parameters are unavailable. This is particularly relevant to PDA: since our defense is designed to be a test-time wrapper that only interacts with the model through its textual interface, its usefulness should not hinge on privileged white-box access. We therefore complement the white-box evaluation with representative black-box attacks to assess whether PDA's gains persist under realistic constraints.

Following prior LVLM robustness studies, we consider three black-box settings that cover common attacker capabilities (Table 4). The first is a transfer-based black-box setting [9], where adversarial perturbations are generated on a surrogate visual encoder and then evaluated on the target LVLM without access to its gradients. We additionally include AttackVLM [57] and AnyAttack [56], two representative black-box attack frameworks that have been widely used to benchmark practical robustness of multimodal systems. For all three attacks, we use the default configurations from the original papers and keep the datasets, evaluation metrics, and prompting templates identical to our white-box experiments; each entry reports the performance of the same victim model with *No Defense* and with *PDA* enabled.

Table 4: Black-box attacks on ImageNet-D and VQA-v2. Each entry is **No Defense / PDA** robust accuracy.

Attack	ImageNet-D	VQA-v2
Transfer-based [9]	0.58 / 0.62	0.77 / 0.81
AttackVLM [57]	0.56 / 0.61	0.75 / 0.79
AnyAttack [56]	0.61 / 0.64	0.79 / 0.81

As shown in Table 4, PDA consistently improves robustness across all three black-box attacks on both ImageNet-D and VQA-v2. We also observe that the absolute accuracies under black-box attacks are higher than those under strong white-box PGD, because black-box attacks, including transfer-based and finite-query settings, are generally weaker than strong white-box attacks. Importantly, the robustness gains from PDA remain stable across attack families, suggesting that PDA does not rely on fragile white-box assumptions. This behavior aligns with the core mechanism of PDA: paraphrase and decomposition generate multiple semantically equivalent views of the same input, and agreement-based aggregation suppresses sporadic failures that do not persist across views, making the defense effective even when the adversary can only interact with the model through black-box feedback.

Closed-source LVLMs under Transfer Attacks. To reflect realistic deployment where the victim LVLM is only accessible through commercial APIs, we further evaluate PDA on closed-source models using a transfer-attack protocol (Table 5). Specifically, we craft adversarial images with white-box PGD on LLaVA and directly transfer these adversarial images to closed-source APIs for evaluation, keeping the input images and prompts fixed and measuring task performance in the same way as in our open-source experiments. This setting is a standard and reproducible way to probe robustness of API-based systems, since it avoids

Table 5: Closed-source LVLMs under transfer attacks. Each entry is **ImageNet-D / VQA-v2** robust accuracy.

	GPT-5	Claude	Gemini
No Defense	0.67 / 0.81	0.61 / 0.82	0.67 / 0.76
PDA	0.76 / 0.90	0.70 / 0.88	0.72 / 0.86

relying on unrestricted query budgets or API-specific gradient estimators while still capturing adversarial vulnerability under realistic access constraints.

Table 5 shows that PDA consistently improves robustness across all tested closed-source LVLMs on both ImageNet-D and VQA-v2. The gains are observed for GPT-5, Claude, and Gemini, indicating that PDA can be applied as a black-box wrapper even when the underlying model family is proprietary and its internals are unknown. These results complement our query-based black-box evaluation: together, they suggest that PDA's robustness benefits persist in practical deployment scenarios, including both iterative black-box attack settings and transfer-based attacks against commercial LVLM APIs.

4.4 Comparison with Test-time Baselines.

Image-side Defenses. Beyond text-side randomized smoothing and training-time robustification baselines (Table 2), we additionally compare PDA with lightweight *image-side* test-time defenses that are widely used as practical pre-processing modules for vision models. Specifically, we consider JPEG compression and random augmentation (RA), which aim to attenuate pixel-level perturbations by either removing high-frequency artifacts (JPEG) or injecting benign input variation (RA) that can partially wash out adversarial patterns. These methods are attractive in deployment because they are model-agnostic and require no changes to model weights or prompting, but they also operate purely on the visual stream and do not exploit the textual modality that LVLMs naturally provide.

We evaluate all defenses under the same white-box ℓ_∞ PGD threat model ($\epsilon{=}2/255$) with identical datasets, metrics, and prompting templates. As shown in Table 6(a), both JPEG and RA improve robustness over the no-defense baseline, but they remain clearly behind PDA on both ImageNet-D and VQA-v2. More importantly, combining them with PDA yields further gains (e.g., JPEG+PDA and RA+PDA), suggesting that PDA is largely *orthogonal* to image-side defenses: pre-processing reduces the strength of pixel-level corruption, while PDA diversifies the textual views and aggregates answers to suppress sporadic failures that do not persist across views. This complementarity makes PDA a practical plug-in that can be stacked with standard pre-processing without additional training.

Reasoning Baselines. We also compare PDA with *reasoning-based* test-time baselines that attempt to recover correct answers by eliciting more structured inference from the victim LVLM, rather than modifying the input image. We consider CoT prompting (LLaVA-CoT) and self-consistency voting, both of which are commonly used to improve reliability in clean settings by encouraging explicit intermediate reasoning and by aggregating multiple sampled outputs. In the adversarial setting, however, these methods operate on the same corrupted visual evidence and thus primarily target errors *after* the perception stage.

Table 6(b) shows that these reasoning baselines provide limited robustness gains and significantly underperform PDA. This gap highlights a key intuition in adversarial VLM robustness: when the image perturbation biases visual perception, generating longer rationales or sampling multiple reasoning traces cannot reliably restore the missing or distorted visual cues, and may even amplify spurious evidence. In contrast, PDA actively diversifies the *queries* to probe stable visual evidence from multiple angles and then aggregates for agreement, making it more effective for correcting perception-level corruption. Overall, these results position PDA as a stronger test-time defense than purely reasoning-based strategies under the same threat model.

Table 6: Comparisons with additional test-time baselines.

Method	ImageNet-D	VQA-v2
No Defense	0.25	0.45
PDA	0.57 (+0.32)	0.75 (+0.30)
(a) Image-side defenses		
JPEG	0.51 (+0.26)	0.68 (+0.23)
JPEG + PDA	0.61 (+0.36)	0.77 (+0.32)
RA	0.55 (+0.30)	0.70 (+0.25)
RA + PDA	**0.67 (+0.42)**	**0.80 (+0.35)**
(b) Reasoning baselines		
LLaVA-CoT	0.53 (+0.28)	0.51 (+0.06)
Self-consistency	0.32 (+0.07)	0.49 (+0.04)

4.5 Ablation Studies and Design Choices

Variants across Tasks and Architectures. To evaluate the effectiveness of our proposed variants, we conduct controlled comparisons across different tasks and backbone architectures under a consistent threat model and decoding protocol. Table 7 compares the *budgeted* PDA variants across VQA-v2 and ImageNet-D. All PDA variants outperform the no-defense baseline, though the optimal choice depends on the specific task–backbone pair. On ImageNet-D, *PDA–PV* generally performs best: in zero-shot CLIP-style classification, class texts are often closely spaced, making it more reliable to verify a few competing labels than to generate long, detail-heavy rationales, which can introduce non-visual hallucinations and sway the decision. Conversely, in VQA, the answer space is compact while the evidential path is heterogeneous; here, revealing more verifiable cues through richer decomposition (e.g., *PDA–RJV*) is generally beneficial.

Table 7: **Ablation of budgeted PDA variants across backbone architectures.** We report robustness (%) of each victim VLM under the same adversarial threat model, with values shown as "VQA-v2 / ImageNet-D."

Method	**DeepSeek-VL**	**InternVL3**	**LLaVA-1.5-7B**	**Ovis2**
No Defense	59.0 / 46.4	69.6 / 25.8	47.2 / 26.8	88.0 / 39.6
PDA–RJV	63.8 / 53.4	**76.0** / **39.4**	**65.2** / 49.0	87.4 / 53.2
PDA–RDA	**63.4** / 51.8	72.4 / 34.0	61.2 / 45.8	**89.8** / 48.4
PDA–PV	61.6 / **70.2**	72.2 / 31.4	58.4 / **57.0**	89.2 / **56.4**

Although PDA-Full achieves the strongest VQA robustness on DeepSeek-VL and LLaVA-1.5-7B, showing that more exhaustive multi-view reasoning can further improve robustness on some backbones. However, the budgeted variants remain preferable overall: PDA–RJV attains the best VQA accuracy on InternVL3, PDA–RDA is slightly superior on Ovis2, and all three budgeted variants incur substantially lower inference cost. On ImageNet-D, the per-sample runtime of each method is as follows: the model without defense takes 4.5 seconds per sample. Among the PDA variants, PDA-Full is the most expensive at 15.5 seconds, followed by PDA-RJV at 10.5 seconds, PDA-RDA at 6.0 seconds, and PDA-PV, which is the fastest, at 5.5 seconds. All measurements are obtained using LLaVA-1.5-13B as the victim model and DeepSeek-Chat as the aggregator on an RTX A6000 GPU.

Effect of the Number of Paraphrases We vary the paraphrase count K to study the robustness–efficiency trade-off under a fixed agent and threat model. Figure 4 compares the representative settings $K \in \{3, 5\}$ across tasks and backbones. The gap between K=3 and K=5 is generally small, and neither choice uniformly dominates across all settings. To further probe whether larger paraphrase widths help, Table 8 extends the sweep to $K \in \{3, 5, 7, 9\}$ for PDA–RJV and PDA–RDA on ImageNet-D.

Table 8: Effect of the number of paraphrases K on ADV accuracy for PDA–RJV and PDA–RDA on ImageNet-D.

Variant	**Backbone**	$K=3$	$K=5$	$K=7$	$K=9$
PDA–RJV	LLaVA-1.5-7B	49.2	**53.2**	53.0	52.2
	InternVL3	**40.8**	39.2	29.2	39.0
	DeepSeek-VL	53.4	**53.4**	53.2	52.4
	Ovis-2	51.4	**53.2**	52.2	52.4
PDA–RDA	LLaVA-1.5-7B	**47.0**	43.7	43.9	43.5
	InternVL3	**34.2**	29.3	30.3	29.3
	DeepSeek-VL	51.8	51.8	51.2	**52.2**
	Ovis-2	47.0	**48.4**	47.6	**48.4**

Overall, robustness tends to saturate quickly once K reaches a moderate range. Across backbones, most differences remain within a few percentage points, and there is no consistent benefit from very large K: the best or near-best results typically occur at K=3 or K=5, while K=7 and K=9 often bring no further gains and occasionally degrade performance. We attribute this to two factors: (i) longer prompts from additional paraphrases increase context length and dispersion, which can hinder retrieval of the truly relevant cues and dilute aggregation; and (ii) marginal paraphrases are noisier and may invite non-visual hallucination. In practice, these results suggest that performance saturates early; we use K=5 as a stable default in the main experiments, while K=3 remains an attractive lower-cost alternative with very similar robustness.

Table 9: **Effect of LLM agent within the PDA pipeline** using K=5. Each cell reports **VQA-v2 / ImageNet-D** robust accuracy (%) for the same victim VLM under different LLM agents.

Agent	**DeepSeek-VL**	**InternVL3**	**LLaVA-1.5-7B**	**Ovis2**
DeepSeek	**63.4** / 54.4	70.8 / 35.7	60.8 / 49.2	**90.0** / **49.2**
GPT	**63.4** / 51.8	**72.4** / 29.3	**61.4** / 43.7	89.8 / 48.4
LLaMA	58.0 / **54.6**	58.4 / **36.1**	58.0 / **50.4**	61.2 / 41.2
Qwen	63.2 / 50.6	70.2 / 34.8	**61.4** / 47.1	89.2 / 47.2

Effect of LLM Agent. Using the same PDA pipeline, we compare four agents: a commercial small model (*GPT-4.1 mini*), a commercial open API model (*DeepSeek-Chat*), and two locally deployable instruct models (*Llama-3.1-8B-Instruct*, *Qwen2.5-14B-Instruct*), as shown in Table 9. On VQA-v2, stronger instruction followers tend to help

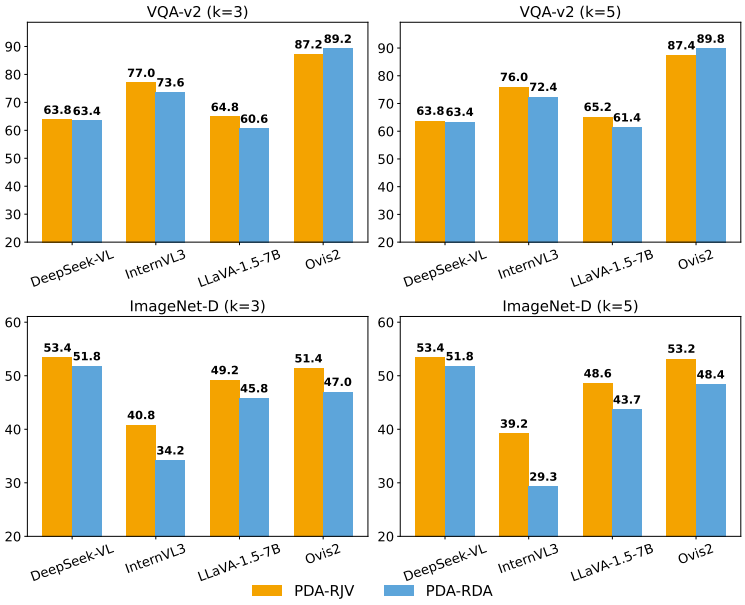

Figure 4: Overall performance of PDA variants on VQA-v2 and ImageNet-D under different paraphrase counts (K=3 and K=5).

decomposition: GPT is best or tied on three backbones and ties DeepSeek-VL. On ImageNet-D, where pairwise checks dominate, capability gaps narrow and the lightweight local models are competitive or best. Practically, this suggests pairing *decomposition-heavy* settings with a stronger agent and *verification-heavy* settings with a compact agent for better accuracy–performance trade-offs.

5 Conclusion

We presented *PDA*, a training-free, text-side defense that stabilizes VLM predictions under adversarial image perturbations by *paraphrasing* the original query into diverse yet semantically equivalent prompts, *decomposing* them into atomic questions, and *aggregating* the resulting answers with agreement-aware voting at inference. To validate practicality and generality, we evaluated propossed framework across different tasks and multiple backbones, observing consistent robustness gains with minimal clean degradation. And on some datasets like ImageNet-D, PDA even improved clean accuracy due reducing spurious predictions by cross-checked evidence. For efficiency, we further explored compressed variants that retain most of the robustness while lowering latency, yielding task-dependent recipes. We hope this perspective on training-free, inference-time robustness will spur research on principled certifications and stronger attack settings, ultimately advancing the safety of VLMs in real-world use.

References

[1] A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283. PMLR, 2018.

[2] S. Bannur, S. Hyland, Q. Liu, F. Perez-Garcia, M. Ilse, D. C. Castro, B. Boecking, H. Sharma, K. Bouzid, A. Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15016–15027, 2023.

[3] R. Bhagwatkar, S. Nayak, P. Bashivan, and I. Rish. Improving adversarial robustness in vision-language models with architecture and prompt design. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17003–17020, 2024.

[4] X. Chen, J. Djolonga, P. Padlewski, B. Mustafa, S. Changpinyo, J. Wu, C. R. Ruiz, S. Goodman, X. Wang, Y. Tay, et al. On scaling up a multilingual vision and language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14432–14444, 2024.

[5] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[6] J. Cohen, E. Rosenfeld, and Z. Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019.

[7] J. Cohen, E. Rosenfeld, and Z. Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019.

[8] F. Croce and M. Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020.

[9] X. Cui, A. Aparcedo, Y. K. Jang, and S.-N. Lim. On the robustness of large multimodal models against image adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24625–24634, 2024.

[10] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. 2014.

[11] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[12] M. Horváth, M. Müller, M. Fischer, and M. Vechev. (de-) randomized smoothing for decision stump ensembles. volume 35, pages 3066–3081, 2022.

[13] N. Hussein, F. Shamshad, M. Naseer, and K. Nandakumar. Promptsmooth: Certifying robustness of medical vision-language models via prompt learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 698–708. Springer, 2024.

[14] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

[15] J. Ji, B. Hou, A. Robey, G. J. Pappas, H. Hassani, Y. Zhang, E. Wong, and S. Chang. Defending large language models against jailbreak attacks via semantic smoothing. *arXiv preprint arXiv:2402.16192*, 2024.

[16] J. Ji, B. Hou, A. Robey, G. J. Pappas, H. Hassani, Y. Zhang, E. Wong, and S. Chang. Defending large language models against jailbreak attacks via semantic smoothing. *arXiv preprint arXiv:2402.16192*, 2024.

[17] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan. Vima: Robot manipulation with multimodal prompts. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research (PMLR)*, pages 14975–15022, 2023.

[18] A. Karmanov, D. Guan, S. Lu, A. El Saddik, and E. Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14162–14171, 2024.

[19] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[20] Y. Li, H. Guo, K. Zhou, W. X. Zhao, and J.-R. Wen. Images are achilles' heel of alignment: Exploiting visual vulnerabilities for jailbreaking multimodal large language models. In *European Conference on Computer Vision*, pages 174–189. Springer, 2024.

[21] W. Lin, Z. Zhao, X. Zhang, C. Wu, Y. Zhang, Y. Wang, and W. Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer, 2023.

[22] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.

[23] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[24] X. Liu, Y. Zhu, Y. Lan, C. Yang, and Y. Qiao. Safety of multimodal large language models on images and texts. *arXiv preprint arXiv:2402.00357*, 2024.

[25] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, H. Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

[26] S. Lu, Y. Li, Q.-G. Chen, Z. Xu, W. Luo, K. Zhang, and H.-J. Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, 2024.

[27] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[28] C. Mao, S. Geng, J. Yang, X. Wang, and C. Vondrick. Understanding zero-shot adversarial robustness for large-scale models. *arXiv preprint arXiv:2212.07016*, 2022.

[29] A. Nirala, A. Joshi, S. Sarkar, and C. Hegde. Fast certification of vision-language models using incremental randomized smoothing. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 252–271. IEEE, 2024.

[30] T. Qian, J. Chen, L. Zhuo, Y. Jiao, and Y.-G. Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4542–4550, 2024.

[31] A. Robey, E. Wong, H. Hassani, and G. J. Pappas. Smoothllm: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*, 2023.

[32] H. Salman, A. Ilyas, L. Engstrom, A. Kapoor, and A. Madry. Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems*, 33:3533–3545, 2020.

[33] H. Salman, J. Li, I. Razenshteyn, P. Zhang, H. Zhang, S. Bubeck, and G. Yang. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in neural information processing systems*, 32, 2019.

[34] C. Schlarmann, N. D. Singh, F. Croce, and M. Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *arXiv preprint arXiv:2402.12336*, 2024.

[35] J. Schuchardt, T. Wollschläger, A. Bojchevski, and S. Günnemann. Localized randomized smoothing for collective robustness certification. 2022.

[36] L. Sheng, J. Liang, Z. Wang, and R. He. R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29958–29967, 2025.

[37] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, J. Beißwenger, P. Luo, A. Geiger, and H. Li. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pages 256–274. Springer, 2024.

[38] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[39] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

[40] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

[41] S. Ugare, T. Suresh, D. Banerjee, G. Singh, and S. Misailovic. Incremental randomized smoothing certification. 2023.

[42] Z. Wan, C. Liu, M. Zhang, J. Fu, B. Wang, S. Cheng, L. Ma, C. Quilodrán-Casas, and R. Arcucci. Med-unic: Unifying cross-lingual medical vision-language pre-training by diminishing bias. *Advances in Neural Information Processing Systems*, 36:56186–56197, 2023.

[43] H. Wang, X. Wu, Z. Huang, and E. P. Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8684–8694, 2020.

[44] R. Wang, S. An, M. Cheng, T. Zhou, S. J. Hwang, and C.-J. Hsieh. One prompt is not enough: Automated construction of a mixture-of-expert prompts. 2024.

[45] X. Wang, K. Chen, J. Zhang, J. Chen, and X. Ma. Tapt: Test-time adversarial prompt tuning for robust inference in vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19910–19920, 2025.

[46] J. Wei, X. Wang, D. Schuurmans, M. Bosma, brian ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou. Chain of thought prompting elicits reasoning in large language models. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[47] P. Xie, Y. Bie, J. Mao, Y. Song, Y. Wang, H. Chen, and K. Chen. Chain of attack: On the robustness of vision-language models against transfer-based adversarial attacks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14679–14689, 2025.

[48] G. Yang, T. Duan, J. E. Hu, H. Salman, I. Razenshteyn, and J. Li. Randomized smoothing of all shapes and sizes. In *International conference on machine learning*, pages 10693–10705. PMLR, 2020.

[49] Q. Ye, H. Xu, J. Ye, M. Yan, A. Hu, H. Liu, Q. Qian, J. Zhang, and F. Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 13040–13051, 2024.

[50] Z. Yin, M. Ye, T. Zhang, T. Du, J. Zhu, H. Liu, J. Chen, T. Wang, and F. Ma. Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models. *Advances in Neural Information Processing Systems*, 36:52936–52956, 2023.

[51] H. S. Yoon, E. Yoon, J. T. J. Tee, M. Hasegawa-Johnson, Y. Li, and C. D. Yoo. C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. *arXiv preprint arXiv:2403.14119*, 2024.

[52] L. Yu, H. Zhang, and C. Xu. Text-guided attention is all you need for zero-shot robustness in vision-language models. *Advances in Neural Information Processing Systems*, 37:96424–96448, 2024.

[53] R. Zhai, C. Dan, D. He, H. Zhang, B. Gong, P. Ravikumar, C.-J. Hsieh, and L. Wang. Macer: Attack-free and scalable robust training via maximizing certified radius. *arXiv preprint arXiv:2001.02378*, 2020.

[54] C. Zhang, F. Pan, J. Kim, I. S. Kweon, and C. Mao. Imagenet-d: Benchmarking neural network robustness on diffusion synthetic object. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21752–21762, 2024.

[55] J. Zhang, X. Ma, X. Wang, L. Qiu, J. Wang, Y.-G. Jiang, and J. Sang. Adversarial prompt tuning for vision-language models. In *European conference on computer vision*, pages 56–72. Springer, 2024.

[56] J. Zhang, J. Ye, X. Ma, Y. Li, Y. Yang, Y. Chen, J. Sang, and D.-Y. Yeung. Anyattack: Towards large-scale self-supervised adversarial attacks on vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19900–19909, 2025.

[57] Y. Zhao, T. Pang, C. Du, X. Yang, C. Li, N.-M. M. Cheung, and M. Lin. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36:54111–54138, 2023.

[58] Y. Zhou, X. Xia, Z. Lin, B. Han, and T. Liu. Few-shot adversarial prompt learning on vision-language models. *Advances in Neural Information Processing Systems*, 37:3122–3156, 2024.

[59] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, H. Tian, Y. Duan, W. Su, J. Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[60] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

A Overview

This supplementary material provides additional details and results for our PDA framework beyond what is reported in the main paper. It is organized into two parts: (i) full prompt specifications for all text-only LLM agents used in PDA, including semantic and logical paraphrasing, decomposition, and caption verification/aggregation and (ii) additional qualitative case studies on VQA-v2, ImageNet-D, and COCO captioning that illustrate typical failure modes and how PDA corrects them.

B Prompts for LLM Agents

In this section, we document the exact prompts used to interact with the text-only LLMs in PDA. For reproducibility, we report system, user, and (when applicable) few-shot example templates. All prompts follow the same high-level design principles as discussed in Section 3 of the main paper, but are presented here in full for completeness.

B.1 Prompts for Paraphrase

The paraphrase agent is prompted as a "Paraphrase Engine" that performs only local synonym or short-phrase substitutions, so that we gain textual diversity without changing the underlying decision problem. The strict rules fix negation, quantities, and proper nouns and preserve the question type, which prevents the LLM from implicitly altering the task (e.g., by re-framing or summarising the query).

We first provide the system role description, followed by the user instruction and an illustrative example. The final prompt used in our experiments can be pasted here as a verbatim code block. The adjustable parameters (`change_intensity`, `num_candidates`) expose a simple knob for controlling diversity, while the JSON-only output constraint removes free-form reasoning text and makes it easy to parse and filter the N candidates.

(a) System prompt (Semantic paraphrase).

```
You are a Paraphrase Engine.

Task: Rewrite the input sentence using synonym / short-phrase substitutions
while preserving the same meaning. Keep the structure and length roughly similar.

Strict rules:
- Do not add or remove facts.
- Preserve negation, comparison, quantities, times, names, and proper nouns
  (keep their casing).
- Keep the sentence type unchanged (questions stay questions, statements stay statements).
- Prefer part-of-speech aligned swaps (noun<->noun, verb<->verb, etc.).
  If no good swap exists, keep the original token.
- No explanations, no reasoning - output only the rewritten sentence(s).
- Match the language of the input (English in -> English out; Chinese in -> Chinese out).

Adjustable parameters:
- change_intensity: low | medium | high
- num_candidates: integer K

Output format:
Return a single JSON object: {"candidates": [...]}
- The array length must equal K.
- Each element is one paraphrased sentence (string).
- No extra text before/after the JSON.
```

(b) User prompt (Semantic paraphrase).

```
Now start:
- change_intensity: medium
- num_candidates: 4
- input_sentence: <ORIGINAL SENTENCE HERE>
```

(c) System prompt (Logical paraphrase for PDA-PV).

```
Request:
Input:
(Example) Which object is more visually prominent in the image?
Choose from the following list: jeans, t shirt.

Generation steps:
To generate logically equivalent contrasting questions, consider the
following attributes:
- Material: different types of materials the objects are made from
  (e.g., denim vs cotton).
- Shape: the overall shape or outline of the object
  (e.g., pant shape vs shirt shape).
- Color: the color or color pattern of the object
  (e.g., blue jeans vs white t-shirt).
- Functional components: elements related to functionality
  (e.g., fitted pants vs fitted shirt).
- State of use: how the object is used or worn
  (e.g., worn on lower body vs worn on upper body).
- Position/area: the area of the body or region in focus
  (e.g., leg area vs chest area).
- Pairing: how the objects are paired with other items
  (e.g., shoes with pants vs accessories with shirt).

Logic equivalence:
The new questions must be logically equivalent to the original question.
If the first option should be selected in the original question, the
first option in every generated question should also be the logically
correct one. This preserves the decision boundary while changing the
surface form.

Output requirements:
- Generate TEN logically equivalent multiple-choice questions.
- Each question should use only the object names in the options, without
  extra explanations in parentheses.
- Return a single JSON object with the structure:

{
  "original_question":
    "Which object is more visually prominent in the image? \
Choose from the following list: (A) jeans, (B) t shirt.",
  "generated_questions": [
    {
      "question":
        "Which object is more visually prominent in the image? \
Choose from the following list: (A) lower body wear, (B) upper body wear.",
      "options": ["lower body wear", "upper body wear"]
    },
    {
      "question":
        "Which object is more visually prominent in the image? \
Choose from the following list: (A) pants, (B) shirt.",
      "options": ["pants", "shirt"]
    },
    {
      "question":
        "Which object is more visually prominent in the image? \
Choose from the following list: (A) leg area, (B) chest area.",
      "options": ["leg area", "chest area"]
    },
    {
      "question":
```

```
       "Which object is more visually prominent in the image? \
Choose from the following list: (A) denim pants, (B) cotton shirt.",
      "options": ["denim pants", "cotton shirt"]
    },
    {
      "question":
       "Which object is more visually prominent in the image? \
Choose from the following list: (A) clothing for legs, (B) clothing for torso.",
      "options": ["clothing for legs", "clothing for torso"]
    }
    ... 5 more question objects ...
  ]
}
```

For PDA-PV, aggregation is reduced to simple voting, so we need the paraphrase layer itself to explore richer but still *logically equivalent* variants of the original multiple-choice query. The above prompt asks the LLM to rewrite the question by contrasting different attributes (material, position, body area, etc.) while preserving the option-wise decision boundary: whenever option A is correct originally, option A must also be correct in every generated question. Constraining the output to a JSON list of questions and option strings makes it easy to parse these logical paraphrases and reuse them directly as VLM queries in PDA-PV.

B.2 Prompts for Decomposition

We now describe the prompts used to decompose a main image question or caption into sub-questions. For generic VQA-style tasks (VQA-v2, ImageNet-D), the LLM directly designs factual sub-questions for the given query; for captioning, we first extract structured atomic claims from the short caption and then generate verification sub-questions conditioned on these claims.

(a) System prompt (Decomposition for VQA/ImageNet-D).

```
You are a visual reasoning expert. Generate 3-5 factual sub-questions to answer an image-based
     question WITHOUT seeing the image. Follow these rules:

Security & Robustness Rules:

Use descriptive phrases instead of specific nouns from the original question (e.g., "long-necked
      animal" instead of "giraffe")
Design redundancy: Key attributes should be cross-verified by multiple questions
Include error-checking questions to catch VLM inconsistencies
Use negative verification questions where applicable
Make questions independent to isolate potential errors

Output Format:
{
  "original_question": "[exact input]",
  "sub_questions": [
    {
      "id": 1,
      "question": "[simple complete question]",
      "answer_type": "yes_no/choice/phrase",
      "options": "[if choice]"
    },
    ...
  ],
  "answer_logic": "[Concise 1-sentence reasoning]"
}

Example:
Input: "Is this rice noodle soup?"
Output:
{
```

```
  "original_question": "Is this rice noodle soup?",
  "sub_questions": [
    {
      "id": 1,
      "question": "Are thin noodle strands visible?",
      "answer_type": "yes_no"
    },
    {
      "id": 2,
      "question": "What noodle texture appears?",
      "answer_type": "choice",
      "options": "translucent/opaque/other"
    },
    {
      "id": 3,
      "question": "Is liquid broth present?",
      "answer_type": "yes_no"
    },
    {
      "id": 4,
      "question": "What color are the noodles?",
      "answer_type": "phrase"
    }
  ],
  "answer_logic": "If Q1=yes and Q2=translucent and Q3=yes -> likely rice noodle soup; otherwise
       not."
}

Now process:
```

For generic VQA and ImageNet-D, this decomposition prompt treats the LLM as a visual reasoning expert that proposes 3–5 factual, robustness-aware sub-questions plus a concise `answer_logic` sentence, encouraging descriptive rephrasings, redundancy, and explicit decision rules so that the downstream VLM+LLM pipeline can cross-verify key attributes instead of relying on a single potentially fragile query.

(b) System prompt (Claim extraction for Caption Verification).

```
You extract atomic visual claims from a short COCO-style caption.

You MUST return a single JSON object with EXACTLY these keys:

- "subject_head": a short noun for the main subject, such as
  "woman", "man", "dog", "cat", "car", "bicycle", "kitchen",
  "bathroom", or "unknown" if you are not sure.
- "subject_count": one of ["one", "two", "many", "unknown"].
- "key_object": a short noun for ONE most important secondary object,
  such as "umbrella", "bench", "truck", "clock". If there is no clear
  secondary object, use "none" or "unknown".
- "relation": a short phrase for the relation between subject and
  key_object, such as "holding", "sitting on", "next to",
  "leaning against", "in front of", or "none"/"unknown" if unclear.
- "scene": one or two words for coarse scene type, such as
  "street", "kitchen", "bathroom", "bedroom", "field", or "unknown"
  if the scene type is not stated.

Rules:
- Base everything ONLY on the given caption.
- Do NOT invent objects or places not mentioned in the caption.
- If something is not stated, use "unknown" (or "none" where appropriate).
- Be concise; prefer single words when possible.

Output format (IMPORTANT):
```

```
- Return ONLY the JSON object, no extra text.
```

User prompt template (Claim extraction).

```
Short caption:
---
{SHORT_CAPTION}
---

Extract the JSON object with keys:
"subject_head", "subject_count", "key_object", "relation", "scene".
```

For caption-based tasks, we first run this claim-extraction stage to convert the short COCO-style caption into a structured set of atomic claims (main subject, count, key object, relation, scene); these claims are then passed as part of the input to the verification-oriented decomposition prompt below, which allows us to design targeted sub-questions that directly stress-test the caption's core semantics.

(c) System prompt (Decomposition for Caption Verification).

```
You design verification sub-questions for an image, focusing on finding possible conflicts with
    a short COCO-style caption.

Inputs:
- Parsed short-caption claims (subject_head, subject_count, key_object, relation, scene).
- One detailed caption of the image (high-recall, possibly noisy).

Goal:
Write 3-5 short questions that will be asked to a vision-language model that CAN see the image,
    to verify the short-caption's core claims:
- main subject identity (from subject_head),
- subject count (one/two/many),
- key secondary object (key_object),
- the relation between the subject and the key object (relation),
- a coarse scene type (scene) if the caption clearly asserts one.

Design them so that:
- If the short caption is wrong about these aspects, the answers are likely to contain "no" or "
    unclear".
- At least one question explicitly checks the number of main subjects.
- At least one question explicitly checks the presence and identity of the key_object (if it is
    not "none"/"unknown").
- At least one question explicitly checks the relation between subject and key_object (if
    relation is not "none"/"unknown").
- Optionally, one question checks the coarse scene type when the caption asserts a clear scene (
    e.g., "kitchen", "bathroom", "street").

COUNT-FOCUSED DESIGN (very important):
- Let claims.subject_count belong {"one", "two", "many", "unknown"}.
- If claims.subject_count is "one", "two", or "many", you MUST create at least two different
    questions about the number of main subjects:
  1) One generic count question whose last sentence is:
     "Answer exactly: 'one', 'two', 'many', 'none', or 'unclear'."
  2) One yes/no question that directly tests the asserted count, for example:
     "Does the image show exactly two women as the main subjects?"
     "Is there exactly one dog as the main subject?"
     and this yes/no question must end with:
     "Answer exactly: 'yes', 'no', or 'unclear'."
- If claims.subject_count is "unknown", you may include only the generic count question (or even
     skip it) and focus more on subject identity and key object.
```

```
APPEARANCE CHECKS FOR MAIN OBJECT:
- Use claims.subject_head as the main subject type.
- If claims.subject_head is very generic (e.g., "person", "people", "man", "woman", "child", "
    animal", "vehicle", "object", "thing", "room"), you may keep appearance questions very short
     or skip one of them.
- If claims.subject_head is more specific (e.g., "toilet", "bathtub", "telephone", "sink", "
    bicycle", "bench", "stove", "laptop", "jeep"), you MUST include appearance checks so that
    the model is forced to describe what the main object looks like.

In particular, when subject_head is specific, you MUST add:
- One question asking the model to briefly describe the main subject's visual appearance (color,
     material, rough shape), for example:
  "Describe the main subject's appearance (color and material) in at most five words. Answer in
      up to five words."
- Optionally, one yes/no question asking whether the main subject looks like a typical <
    subject_head>, for example:
  "Does the main subject look like a typical toilet fixture you would see in a bathroom? Answer
      exactly: 'yes', 'no', or 'unclear'."
- These appearance questions are intended to catch confusions such as mistaking a toilet for an
    old telephone, or a sink for a bathtub.

Constraints:
- Most questions must be yes/no/unclear with EXACT answer format:
  "Answer exactly: 'yes', 'no', or 'unclear'."
- Include at least one generic count question with the fixed answer set:
  "Answer exactly: 'one', 'two', 'many', 'none', or 'unclear'."
- You may add 1--3 short fill-in questions (appearance, color, scene) with answer formats like:
  "Answer in up to two words."
  "Answer in at most five words."
- Do NOT introduce new specific objects that do not appear in the short caption or detailed
    caption.

Probe for contradictions:
- If the short caption claims something unusual or very specific (e.g., "two women", "three
    giraffes", "a bird pulling a wagon", "an old telephone in a bathtub"), design questions that
     directly test those claims, especially the number of subjects and the main object's
    appearance.
- For relations (e.g., "holding", "sitting on", "next to", "leaning against"), ask explicitly
    whether this relation is true.

Output format (IMPORTANT):
- Return ONLY a single JSON object:
  {
    "sub_questions": [ "...", ... ]
  }
- "sub_questions" must be a list of 3-5 question strings.
- Each question string must include its own answer-format instruction.
- Do NOT add any explanation or extra keys.
```

For caption-based evaluation, the decomposition prompt is specialized to verify the structured claims produced in the previous step: given the extracted subject, count, key object, relation, and scene, it asks the LLM to generate 3-5 focused verification questions with fixed answer formats, making it easy to detect contradictions between the short caption and the image while keeping the output as a clean JSON list of queries for the VLM.

B.3 Prompts for Aggregation

We finally specify the prompts that aggregate sub-question answers into a single prediction. For generic VQA-style tasks, the aggregator reasons over all sub-answers to produce a short, task-aligned answer; for captioning, the aggregator acts as a COCO-style caption judge that decides whether to keep or minimally correct the original caption.

(a) System prompt (Aggregation for VQA/ImageNet-D).

```
Answer the question by reasoning over ALL sub-question answers. Follow these rules:

1. Output format:
   - Quantity questions: ONLY the number (e.g., "2")
   - Yes/no questions: ONLY "yes" or "no"
   - Action questions ("what is doing"): ONLY the verb/phrase (e.g., "crossing" not "crossing
     the street")
   - Gaze questions ("where looking"): ONLY direction (up/down/left/right, not "-wards" or "
     upwards")
   - NO articles: a/an/the
   - Others: single word/phrase (max 3--4 words)

2. Reasoning process (internal):
   a. Synthesize ALL answers to form a coherent conclusion
   b. Make logical inferences (e.g., trees + leaves -> forest)
   c. Avoid being too dependent on any single sub-question

3. For location/context questions:
   - Combine environmental clues (e.g., trees + leaves -> forest)
   - Consider object interactions (e.g., beach + waves -> ocean)

Question:
{question}

Sub-questions and answers:
{sub}

Output ONLY the answer with no additional text (no explanations):
```

For VQA and ImageNet-D, the aggregation prompt instructs the LLM to treat sub-question answers as evidence, integrate them with lightweight internal reasoning, and return a minimal, format-controlled answer string (number, yes/no, or short phrase), which keeps the output easy to score while reducing sensitivity to any single noisy or adversarial probe.

(b) System prompt (Aggregation for Caption Verification).

```
COCO-style caption judge and editor (no trimming + confidence gate).

Inputs:
1. A short COCO-style caption (one sentence).
2. Parsed claims from that caption: subject_head, subject_count, key_object, relation, scene.
3. A small set of sub-questions and their answers from a VLM that can see the image.

Critical slots to check:
- main subject identity (head noun)
- subject count (one/two/many)
- key object (if any)
- relation between subject and key object
- coarse scene type (if the caption asserts one)

Strict decision rule with confidence gate:

For each slot that the short caption clearly asserts:
- Look for support in the answers.
  - "yes" on a matching question = positive support.
  - Any "no" that contradicts the slot = conflict.
  - Only "unclear" (and no supporting "yes") = unsupported.

Confidence gate (edit only when clearly confident):
You may modify the caption only if at least one is true:
1. Direct conflict: a "no" explicitly contradicts the asserted slot.
```

```
2. Strong counter-evidence for a different value: at least two consistent, relevant "yes"
    answers indicating a specific alternative, with no conflicting "no" answers for that slot.

If a slot is merely unsupported (only "unclear" or weak/mixed signals) and not contradicted, do
    not change that slot.

If all asserted slots are non-conflicting and at least one asserted slot has positive support,
    keep the original caption.

Global decision:
- Set has_conflict = true only when you actually change the caption under the confidence gate
    above.
- Otherwise set has_conflict = false and keep the caption exactly as given.

Edit rules (no trimming policy):

When has_conflict = true:
- Preserve all original descriptive content whenever possible.
  - Only change words that are explicitly proven wrong by the sub-question answers.
  - Do not shorten, simplify, or remove adjectives, modifiers, or secondary objects unless they
      are directly contradicted.
  - If uncertain, keep the original content instead of deleting it.

Allowed edits:
1. Fix only the specific critical slot(s) that conflict:
   - subject_head (e.g., "dog" -> "cat")
   - subject_count (e.g., "two" -> "one")
   - key_object (e.g., "table" -> "bench")
   - relation (e.g., "on" -> "next to")
   - scene (e.g., "kitchen" -> "living room")
2. Keep all other content exactly unchanged.
3. You may rephrase minimally to maintain grammatical correctness after fixing the conflicts.
4. Do not trim sentence length, remove descriptors, or generalize nouns unless you are
    absolutely certain that the original word is wrong.

Style:
- One English sentence, 8--20 words (slightly longer allowed for full retention).
- Preserve original detail level and structure whenever possible.
- Use natural, descriptive COCO-style language.

Output format (IMPORTANT):
Return only a JSON object with exactly these keys:
{
  "has_conflict": <true or false>,
  "caption": "<final short caption>"
}

Important:
- If "has_conflict" is false, "caption" must be exactly the original short caption (no
    rephrasing, no added words, no removed words).
- If "has_conflict" is true, "caption" must be the corrected sentence following the rules above,
     changing only the slots that are clearly contradicted while preserving all other
    descriptive content.
```

For caption-based aggregation, the prompt casts the LLM as a conservative COCO-style caption judge that uses VLM-derived answers to check a small set of critical slots and only edits the caption when there is clear, slot-specific evidence of conflict, otherwise returning the original sentence verbatim; the JSON output with a boolean conflict flag and a single caption string makes it straightforward to integrate this decision into our overall PDA pipeline.

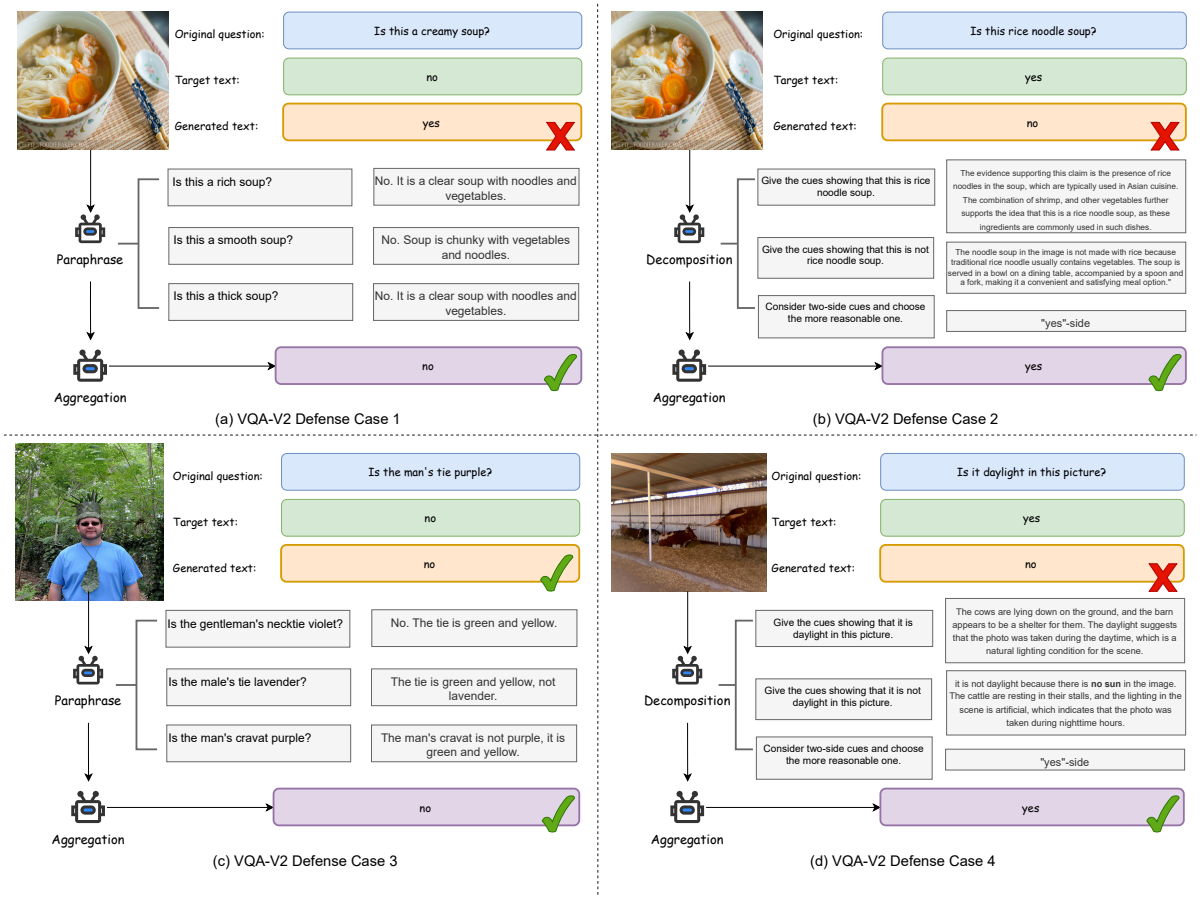

Figure 5: Additional VQA-v2 qualitative examples. For each adversarial image we compare the answer of the undefended VLM with the output of PDA and visualize key paraphrases and sub-questions that support the corrected decision.

C Additional Qualitative Examples

This section provides additional qualitative examples across VQA-v2, ImageNet-D, and COCO captioning to complement the quantitative results in the main paper. For each dataset we highlight typical failure modes exhibited by undefended VLMs and how PDA corrects them through paraphrasing, decomposition, and aggregation. All examples are drawn directly from our evaluation pipeline.

C.1 VQA-v2: Defense Case Studies

Figure 5 shows several representative VQA-v2 cases where adversarial perturbations cause the victim models to produce incorrect answers on seemingly easy questions, such as misclassifying food categories or confusing simple attributes (e.g., soup type, clothing color, or whether the scene is in daylight). PDA corrects these failures by enforcing semantic consistency across multiple paraphrases and their associated sub-questions: even when some prompts are hijacked by the attack, the majority of paraphrase–sub-question pairs still support the correct answer, and the aggregation stage downweights inconsistent evidence and recovers the underlying semantics.

C.2 ImageNet-D: Fine-Grained Recognition Under Shift

On ImageNet-D, the combination of distribution shift and adversarial noise often pushes the baseline models toward visually nearby but incorrect categories. As illustrated in Figure 6, undefended VLMs tend to overfit to corrupted

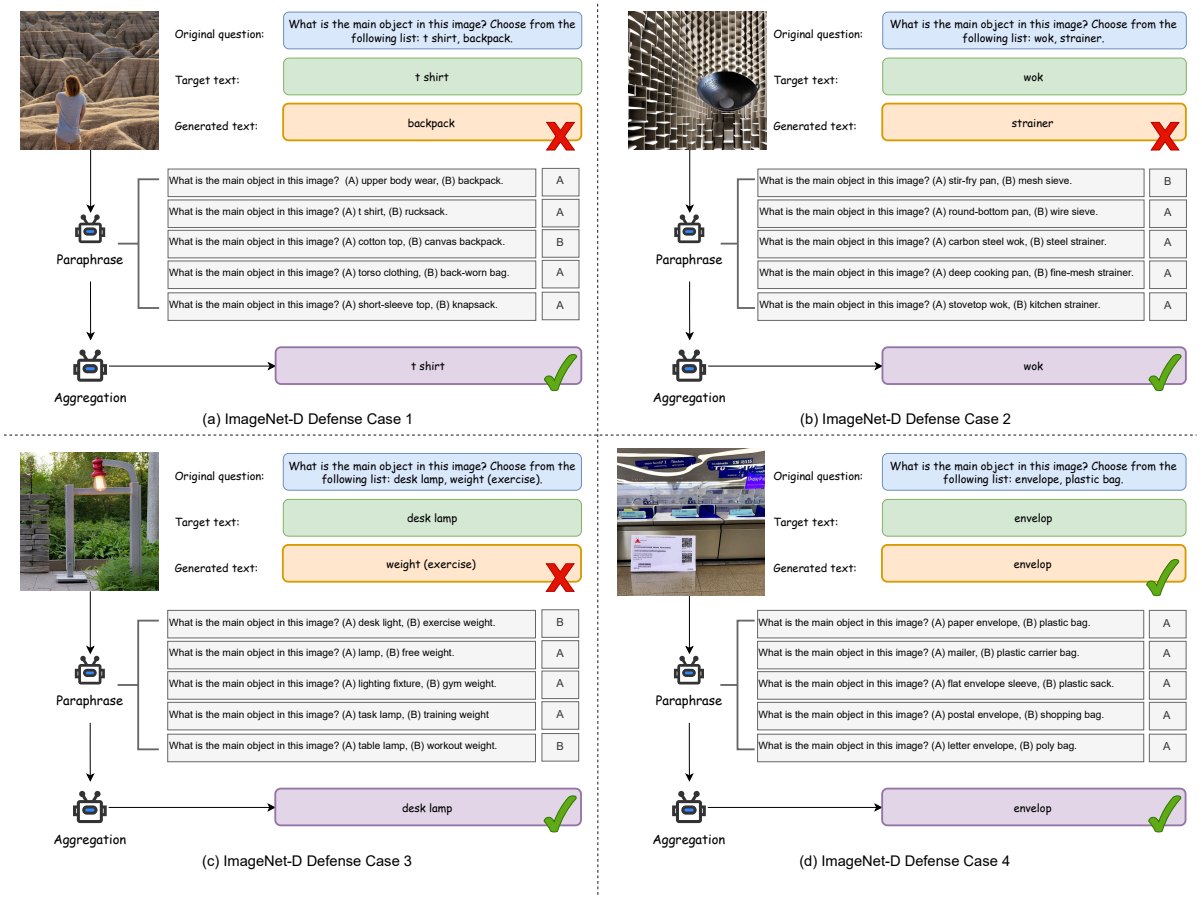

Figure 6: Additional ImageNet-D qualitative examples. PDA stabilizes fine-grained recognition by aggregating evidence from multiple paraphrases that describe shape, part configuration, and material, instead of relying on a single corrupted view.

local textures or background cues, leading to fine-grained misclassification. PDA mitigates this by forcing agreement between paraphrase-derived cues that describe shape, parts, material, and coarse category; inconsistent labels that cannot be reconciled with these multi-view descriptions are rejected, and the final aggregated prediction aligns more closely with the true object class, even when the raw logits of the victim model are heavily distorted.

C.3 COCO Captioning: Caption Repair and Conflict Resolution

For COCO captioning, the task is to repair a short caption that has been corrupted by adversarial perturbations. Figure 7 presents several examples where the baseline caption either hallucinates non-existent objects, misstates the subject identity or count, or confuses relations and scene type. PDA first extracts atomic claims from the short caption (main subject, subject count, key object, relation, and scene) and then verifies each claim using a set of structured sub-questions posed to a VLM that can see the image. Claims that conflict with the answers are minimally edited, while unsupported but non-contradicted content is left unchanged. As a result, PDA produces corrected captions that fix concrete visual errors without aggressively rewriting or shortening the original sentence, leading to more reliable and faithful descriptions under adversarial conditions.

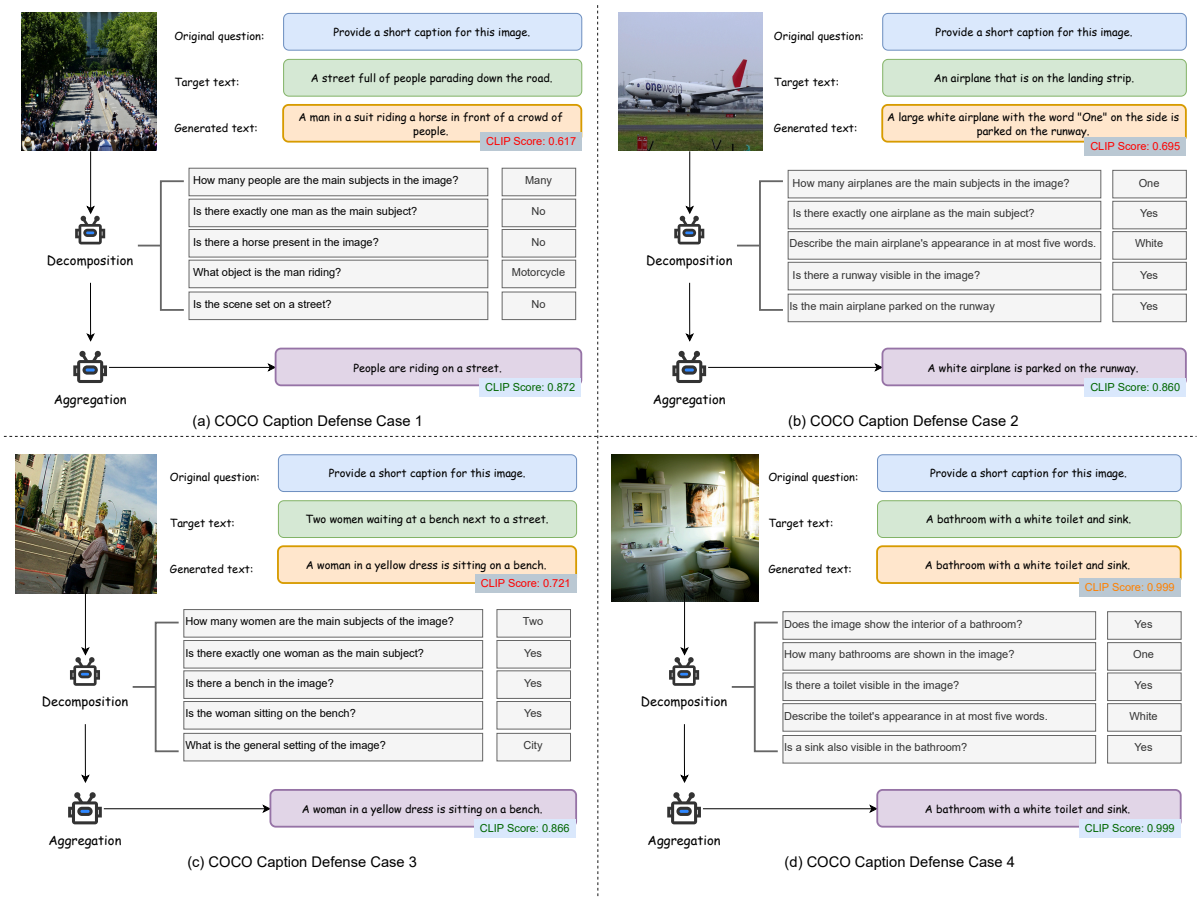

Figure 7: Additional COCO captioning examples. PDA detects and repairs conflicts between a corrupted short caption and VLM answers to verification sub-questions, yielding minimally edited but visually consistent captions.