Do Not Leave a Gap: Hallucination-Free Object Concealment in Vision-Language Models

Amira Guesmi and Muhammad Shafique

eBrain Lab, Engineering Division, New York University Abu Dhabi, UAE

Abstract. Vision-language models (VLMs) have recently shown remarkable capabilities in visual understanding and generation, but remain vulnerable to adversarial manipulations of visual content. Prior object-hiding attacks primarily rely on suppressing or blocking region-specific representations, often creating semantic gaps that inadvertently induce hallucination, where models invent plausible but incorrect objects. In this work, we demonstrate that hallucination arises not from object absence per se, but from semantic discontinuity introduced by such suppression-based attacks. We propose a new class of *background-consistent object concealment* attacks, which hide target objects by re-encoding their visual representations to be statistically and semantically consistent with surrounding background regions. Crucially, our approach preserves token structure and attention flow, avoiding representational voids that trigger hallucination. We present a pixel-level optimization framework that enforces background-consistent re-encoding across multiple transformer layers while preserving global scene semantics. Extensive experiments on state-of-the-art vision-language models show that our method effectively conceals target objects while preserving up to 86% of non-target objects and reducing grounded hallucination by up to 3× compared to attention-suppression-based attacks. Qualitative results further confirm that our approach maintains scene coherence and avoids spurious object insertion. Our findings highlight semantic continuity as a key factor in hallucination behavior and introduce a new direction for adversarial analysis of generative multimodal models.

Keywords: VLM · Hallucination · Object Concealment

1 Introduction

Vision–Language Models (VLMs) have become foundational components of modern AI systems, enabling image captioning, visual question answering, and multimodal reasoning [1, 6, 12]. Despite their impressive capabilities, VLMs remain prone to hallucination, generating objects, attributes, or relationships that are not supported by the visual input [3, 8]. Hallucination has emerged as one of the central reliability challenges in multimodal AI, particularly in applications where visual evidence must be faithfully preserved.

A setting in which hallucination becomes particularly pronounced is *object concealment*, where specific objects or regions are intentionally hidden from a

model while preserving the overall scene semantics [5, 9, 11]. Such concealment techniques are increasingly relevant for privacy preservation, content protection, and controlled information disclosure. Existing approaches typically achieve concealment by suppressing visual evidence associated with a region of interest (ROI), for example through masking, patch removal, or attention suppression mechanisms [11, 15]. While these methods can successfully prevent recognition of the target object, they frequently produce an undesirable side effect: the model compensates for the missing visual evidence by generating plausible but visually unsupported content [3, 8]. As a result, object concealment often trades recognition failure for hallucination.

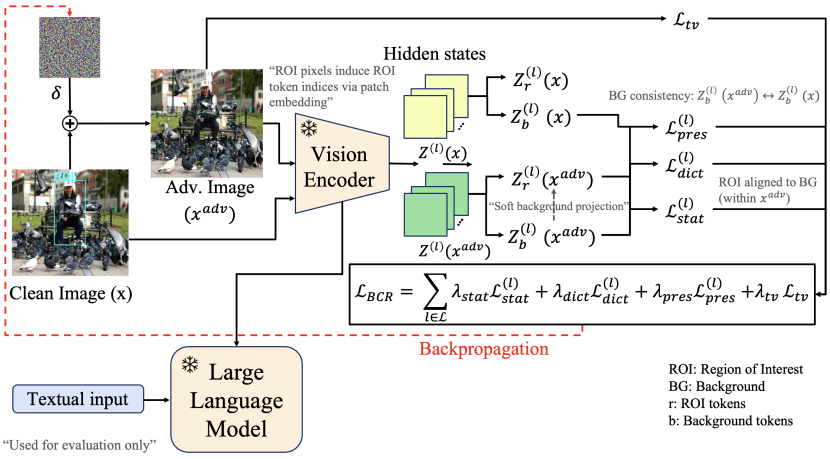

Fig. 1: Overview of **Background-Consistent Re-encoding (BCR)**. A pixel-level perturbation δ is optimized to produce an adversarial image $\mathbf{x}^{adv}$ from a clean image $\mathbf{x}$ with a specified ROI. A frozen vision encoder extracts layer-wise hidden states, from which ROI and background tokens are identified via patch embedding. BCR enforces semantic continuity by (i) aligning ROI and background statistics, (ii) softly projecting ROI features onto background representations, and (iii) preserving background tokens between clean and adversarial images. A total variation regularizer encourages smooth perturbations. The objective is optimized across multiple transformer layers, while the language model remains frozen and is used only for evaluation.

In this work, we investigate a fundamental question: *why does object concealment induce hallucination?* We argue that hallucination is not primarily caused by object absence itself, but by the *representational discontinuities* introduced when visual evidence is aggressively suppressed. Suppression-based concealment methods create feature-space gaps between ROI representations and their surrounding visual context. These discontinuities propagate through the visual encoder and cross-modal alignment modules, forcing the language model to compensate for missing or anomalous representations through prior-driven

inference [8]. Consequently, the model often generates objects or scene elements that are not visually grounded.

To test this hypothesis, we introduce *Background-Consistent Re-encoding (BCR)* (Figure 1), a continuity-preserving object concealment framework. Rather than removing ROI information, BCR re-encodes ROI representations such that they become statistically and semantically consistent with surrounding background features. Specifically, BCR enforces statistical alignment between ROI and background tokens, projects ROI features onto the background feature manifold, and explicitly preserves non-ROI representations. This formulation maintains continuity within the visual representation while concealing object-specific semantics. BCR serves both as a practical concealment method and as an experimental tool for studying the relationship between representation continuity and hallucination. Through quantitative analyses, we show that reductions in feature discontinuity are consistently accompanied by reductions in grounded hallucination. Across multiple VLM architectures, BCR achieves concealment performance comparable to suppression-based methods while substantially improving global semantic preservation and reducing hallucination. These findings provide empirical evidence that hallucination in concealment settings is closely linked to representational discontinuity rather than object removal alone.

This paper makes the following contributions:

- We identify representational discontinuity as a key driver of hallucination in object concealment settings and provide empirical evidence linking continuity preservation to hallucination reduction.
- We introduce *Background-Consistent Re-encoding (BCR)*, a continuity-preserving object concealment framework that hides target objects while maintaining statistical and semantic consistency with the surrounding background.
- We propose a multi-layer optimization formulation that enforces statistical alignment, dictionary-based semantic projection, and background preservation within the visual representation.
- We introduce hallucination-aware evaluation protocols and demonstrate that continuity-preserving concealment substantially reduces grounded hallucination while maintaining strong concealment performance across diverse vision–language models.

2 Related Work

2.1 Adversarial Attacks on Vision–Language Models

Vision–language models (VLMs) have been shown to inherit and amplify adversarial vulnerabilities from their visual backbones while introducing new cross-modal failure modes. Early work adapted classical image-space attacks such as FGSM and PGD [4, 10] to multimodal architectures, demonstrating that small, imperceptible perturbations can significantly degrade image captioning and visual question answering (VQA) performance [9, 15]. These studies established that multimodal grounding does not inherently confer robustness and that

adversarial perturbations can propagate through cross-modal alignment layers to corrupt downstream language generation. Subsequent research has explored more structured attack strategies that explicitly exploit multimodal fusion mechanisms. While effective, these attacks generally aim to *break* the model's predictions, often at the cost of destroying overall image semantics and utility.

2.2 Adversarial Attacks for Privacy and Information Protection

More recently, adversarial attacks have been explored as tools for privacy preservation. VIP (Visual Information Protection) [11] frames privacy as an adversarial objective and proposes selectively masking regions of interest (ROIs) by suppressing attention and value activations in early vision layers, effectively preventing VLMs from recognizing sensitive content. VIP demonstrates strong concealment performance across multiple VLMs, but relies on explicit ROI suppression, which creates a representational "gap" that can encourage the language model to hallucinate or infer missing content. Other methods explicitly manipulate internal representations rather than output tokens. For instance, PRM (Patch Representation Misalignment) [5] disrupts hierarchical patch representations by enforcing feature divergence between clean and adversarial images, leading to global semantic corruption. However, these methods typically alter pixel-level appearance or remove information entirely, making them unsuitable for scenarios where global semantic coherence must be preserved. In contrast to prior work, our method does not suppress or erase visual information within the ROI. Instead, we propose *Background-Consistent Re-encoding (BCR)* manipulation, which reshapes the *distributional relationship* between ROI and background tokens inside the vision encoder.

3 Threat Model

We consider a *white-box, image-level adversary* targeting vision–language models (VLMs). The adversary is given access to the model architecture and parameters, and can compute gradients with respect to the visual input. The attack operates by applying a bounded perturbation to the input image at the pixel level, subject to an ℓ_∞ norm constraint. The adversary is additionally provided with a region of interest (ROI), specified as a bounding box, corresponding to an object whose visual presence should be concealed. Such ROIs naturally arise from object detectors, annotations, or user-defined sensitive regions.

The adversary's objective is not to induce misclassification or nonsensical outputs, but rather to *remove the semantic evidence of the target object* from the model's internal representation while preserving global visual coherence. Importantly, the attacker seeks to avoid introducing explicit visual artifacts or representational gaps that could trigger compensatory hallucinations by the language decoder. Unlike object removal or inpainting approaches, our objective is not to alter the visual scene but to modulate representational grounding within the vision–language model. This enables privacy-preserving concealment

while maintaining perceptual readability and evidentiary integrity for human observers. The adversary does not modify the model weights, prompts, or decoding strategy, and has no control over the language model beyond its dependence on visual features. All evaluations are performed under this fixed threat model.

4 Continuity-Preserving Object Concealment

Problem Setting. Vision–language models (VLMs) rely on a continuous visual representation to ground language generation, where aligned visual tokens provide the evidential basis for downstream captioning and reasoning [1,6,12]. When salient visual evidence is abruptly removed or heavily suppressed (e.g., via masking, erasing, or zeroing localized regions), the model is forced to infer missing content from contextual priors encoded in the language decoder [8, 13]. This inference process frequently manifests as *hallucination*, where the model fabricates objects or attributes that are not visually present. Existing object-hiding or privacy-oriented attacks often rely on explicit removal or severe suppression of visual features corresponding to a target region [11, 15]. While effective at concealing the object itself, such approaches introduce a representational "gap" in the visual embedding space that the model attempts to compensate for during decoding, leading to unstable or hallucinated outputs [8]. This effect is particularly pronounced in large-scale VLMs, where strong language priors and generative biases amplify missing or ambiguous visual evidence [8].

Principle of Background-Consistent Re-encoding (BCR). Instead of removing or suppressing the target object, we propose to *re-encode* its visual representation such that it becomes statistically and semantically consistent with the surrounding background. The goal is not to erase information, but to *blend* the target region into the background feature manifold. Concretely, BCR enforces three complementary constraints:

- Statistical alignment: The first- and second-order feature statistics of target-region tokens are matched to those of background tokens.

- Dictionary consistency: Target-region features are reconstructed as convex combinations of background features, ensuring that they lie within the background feature span.

- Background preservation: Features outside the target region are explicitly preserved to prevent global degradation or collateral semantic drift.

By maintaining continuity in the visual feature space, the global image representation remains coherent and does not signal missing content to the language decoder. As a result, the model neither mentions the concealed object nor compensates for its absence via hallucination.

Contrast with Suppression-Based Attacks. Unlike masking- or suppression-based attacks (e.g., patch removal, attention blocking, or zeroing), BCR does not introduce visually or representationally empty regions. Instead, the target object is *absorbed* into the background representation. This distinction is critical: while suppression creates a void that invites hallucination, background-consistent re-encoding preserves representational smoothness and stabilizes downstream lan-

guage generation. The resulting adversarial image conceals the target object while preserving global semantics and minimizing hallucinated content. Importantly, BCR operates entirely at the representation level while being implemented via pixel-space optimization, making it applicable to black-box or frozen VLMs without modifying model parameters.

5 Continuity-Preserving Concealment Formulation

We consider a vision–language model (VLM) composed of a vision encoder f_θ and a language decoder g_ϕ. Given an input image $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$, the vision encoder produces a sequence of visual tokens $\mathbf{Z} = f_\theta(\mathbf{x}) \in \mathbb{R}^{T\times D}$, where the first token corresponds to a global representation (CLS) and the remaining tokens encode spatial image regions. The language decoder conditions on this representation to generate a caption or answer. Let $\mathcal{R}$ denote a target region of interest (ROI) corresponding to an object that the attacker wishes to conceal, specified by bounding boxes. Let $\mathcal{I}_r \subset \{1, \ldots, T-1\}$ be the indices of visual tokens whose receptive fields intersect the ROI, and $\mathcal{I}_b$ the indices of background tokens. Our goal is to produce an adversarial image $\mathbf{x}^{\text{adv}}$ that conceals the target object while preserving global semantic consistency and avoiding hallucination.

5.1 Background-Consistent Re-encoding Objective

Unlike suppression-based attacks that remove or mask ROI features, we aim to *re-encode* ROI tokens such that they become statistically and semantically indistinguishable from background tokens. Formally, we optimize the following objective:

$$\min_{\mathbf{x}^{\text{adv}}} \ \mathcal{L}_{\text{BCR}}(\mathbf{x}^{\text{adv}}) = \sum_{l\in\mathcal{L}} \Big(\lambda_{\text{stat}}\mathcal{L}^{(l)}_{\text{stat}} + \lambda_{\text{dict}}\mathcal{L}^{(l)}_{\text{dict}} + \lambda_{\text{pres}}\mathcal{L}^{(l)}_{\text{pres}}\Big) + \lambda_{\text{tv}}\mathcal{L}_{\text{tv}}, \quad (1)$$

subject to a pixel-level perturbation budget $\|\mathbf{x}^{\text{adv}} - \mathbf{x}\|_\infty \le \epsilon$. The losses are computed across a set of intermediate vision encoder layers $\mathcal{L}$ to enforce multi-level consistency.

5.2 Statistical Alignment Loss

To eliminate distributional discrepancies that could signal the presence of a concealed object, we align the first- and second-order statistics of ROI and background features. For layer l, let $\mathbf{Z}^{(l)}_r = \{\mathbf{z}^{(l)}_i \mid i \in \mathcal{I}_r\}$ and $\mathbf{Z}^{(l)}_b = \{\mathbf{z}^{(l)}_j \mid j \in \mathcal{I}_b\}$. We define:

$$\mathcal{L}^{(l)}_{\text{stat}} = \big\|\mu(\mathbf{Z}^{(l)}_r) - \mu(\mathbf{Z}^{(l)}_b)\big\|_2^2 + \big\|\sigma(\mathbf{Z}^{(l)}_r) - \sigma(\mathbf{Z}^{(l)}_b)\big\|_2^2, \quad (2)$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote per-dimension mean and standard deviation. This loss enforces that ROI features follow the same distribution as background features, preventing detectable anomalies.

5.3 Dictionary Projection Loss

While statistical alignment controls low-order moments, it does not ensure semantic consistency. To explicitly constrain ROI features to lie on the background feature manifold, we introduce a soft background projection loss. For each ROI token, we compute a soft assignment over background tokens using scaled dot-product similarity: $\alpha_{ij}^{(l)} = \frac{\exp\left(\mathbf{z}_{r,i}^{(l)} \cdot \mathbf{z}_{b,j}^{(l)} / \tau\right)}{\sum_{k \in \mathcal{I}_b} \exp\left(\mathbf{z}_{r,i}^{(l)} \cdot \mathbf{z}_{b,k}^{(l)} / \tau\right)}$. Each ROI feature is then projected onto the background feature dictionary: $\hat{\mathbf{z}}_{r,i}^{(l)} = \sum_{j \in \mathcal{I}_b} \alpha_{ij}^{(l)} \mathbf{z}_{b,j}^{(l)}$.

The dictionary loss is defined as:

$$\mathcal{L}_{\text{dict}}^{(l)} = \frac{1}{|\mathcal{I}_r|} \sum_{i \in \mathcal{I}_r} \left\| \mathbf{z}_{r,i}^{(l)} - \hat{\mathbf{z}}_{r,i}^{(l)} \right\|_2^2. \tag{3}$$

This loss encourages ROI tokens to be re-encoded as convex combinations of background features, enforcing semantic continuity without removing token structure.

5.4 Background Preservation Loss

To avoid global semantic drift and unintended distortions, we explicitly preserve background features by penalizing deviations from their clean counterparts:

$$\mathcal{L}_{\text{pres}}^{(l)} = \frac{1}{|\mathcal{I}_b|} \sum_{j \in \mathcal{I}_b} \left\| \mathbf{z}_{b,j}^{(l)}(\mathbf{x}^{\text{adv}}) - \mathbf{z}_{b,j}^{(l)}(\mathbf{x}) \right\|_2^2. \tag{4}$$

This constraint ensures that only the ROI representation is altered, while the remainder of the image remains perceptually and semantically stable.

5.5 Pixel-Level Regularization

Finally, we apply total variation regularization to encourage spatially smooth perturbations within the ROI:

$$\mathcal{L}_{\text{tv}} = \sum_{c,h,w} \left| \mathbf{x}_{c,h,w+1}^{\text{adv}} - \mathbf{x}_{c,h,w}^{\text{adv}} \right| + \left| \mathbf{x}_{c,h+1,w}^{\text{adv}} - \mathbf{x}_{c,h,w}^{\text{adv}} \right|. \tag{5}$$

By jointly enforcing statistical alignment, dictionary-based semantic projection, and background preservation across multiple layers, BCR removes object-specific information without introducing representational gaps. As a result, the global image representation remains coherent, preventing both object mention and compensatory hallucination during language generation.

Why Representation Continuity Prevents Hallucination. Vision–language models rely on the assumption that visual token representations form a continuous and semantically coherent manifold. During generation, the language decoder implicitly performs inference over this manifold: when a region-specific signal is

missing or anomalous, the decoder compensates by extrapolating from learned visual–semantic priors, often resulting in hallucinated content. This behavior is particularly pronounced in suppression-based attacks, where masking, zeroing, or attention blocking creates a representational "gap" that violates the continuity assumptions learned during pretraining. Our attack avoids this failure mode by enforcing *representation continuity* rather than suppression. By re-encoding ROI tokens to match the statistical distribution and semantic span of background tokens, we ensure that the visual embedding remains locally smooth and globally consistent. From the decoder's perspective, the ROI does not appear as missing information, but as background-consistent evidence. Consequently, there is no incentive for the model to hypothesize or hallucinate alternative objects to explain an apparent void. Importantly, this continuity is enforced across multiple layers of the vision encoder, preventing object semantics from re-emerging through higher-level abstractions or cross-token interactions. Thus, hallucination is reduced not by restricting generation, but by eliminating representational discontinuities that would otherwise trigger compensatory reasoning in the language model.

6 Evaluation Metrics

Evaluating object concealment attacks on vision–language models requires more than measuring caption similarity or task accuracy. A successful concealment attack must simultaneously (i) remove evidence of a target object, (ii) preserve non-target visual semantics, and (iii) avoid inducing hallucinated content. To capture these distinct objectives, we introduce a set of complementary evaluation metrics combining caption analysis with grounded visual verification.

Object Sets. Given a clean image I and its adversarial counterpart I', we generate captions $c = f(I)$ and $c' = f(I')$ using the same prompting strategy. From each caption, we extract a set of object-like tokens using a dependency-based noun phrase parser: $\mathcal{O}(c), \quad \mathcal{O}(c')$.

All object strings are normalized (lemmatized and lowercased) prior to comparison.

Concealment Success. Concealment Success measures whether the target object is successfully removed from the adversarial caption. Let o^* denote the target object. We define:

$$\text{ConcealmentSuccess} = \mathbb{1}\big[o^* \in \mathcal{O}(c) \ \wedge \ o^* \notin \mathcal{O}(c')\big]. \tag{6}$$

This metric isolates the primary attack objective independently of other caption changes.

Global Preservation. Global Preservation measures how well non-target visual content is retained. It is defined as the fraction of clean-caption objects that remain present after the attack:

$$\text{GlobalPreservation} = \frac{|\mathcal{O}(c) \cap \mathcal{O}(c')|}{|\mathcal{O}(c)|}. \tag{7}$$

A high preservation score indicates that the attack does not indiscriminately erase visual semantics.

Grounded Hallucination Rate. Caption-level comparison alone cannot determine whether newly mentioned objects are visually supported. To address this limitation, we introduce a *grounded hallucination* metric that verifies adversarial caption objects against visual evidence using an external grounding model.

Let $\mathcal{H} = \mathcal{O}(c') \setminus \mathcal{O}(c)$ denote the set of newly introduced objects in the adversarial caption. For each object $h \in \mathcal{H}$, we query a grounding model (GLIP [7]) to determine whether the object can be visually localized in the image. Formally, let

$$\text{detect}(h, I') = \begin{cases} 1 & \text{if the grounding model localizes object } h \text{ with confidence } > \tau, \\ 0 & \text{otherwise.} \end{cases}$$

Objects that cannot be grounded are considered hallucinated. The grounded hallucination rate is defined as:

$$\text{GroundedHallucinationRate} = \frac{|\{h \in \mathcal{H} \mid \text{detect}(h, I') = 0\}|}{|\mathcal{O}(c')|}. \tag{8}$$

This formulation ensures that hallucination is measured relative to actual visual evidence rather than only caption differences.

Head-Noun Grounded Hallucination. To account for paraphrasing and compound nouns (e.g., *"salt shaker"* vs. *"shaker"*), we additionally report a head-noun grounded hallucination rate. Each object phrase is reduced to its syntactic head noun prior to grounding verification, yielding a more semantically robust estimate.

Semantic Drift. While hallucination focuses on unsupported object introduction, Semantic Drift measures global caption deviation. Using a text embedding function $\phi(\cdot)$ (e.g., a sentence encoder), we compute: $\text{SemanticDrift} = 1 - \cos\big(\phi(c), \phi(c')\big)$. This metric ensures that reduced hallucination is not achieved by collapsing captions into generic or uninformative descriptions. Together, these metrics provide a fine-grained evaluation of concealment attacks.

Unlike CHAIR [13] and related hallucination metrics [8], which assume static object presence and operate at coarse category granularity, our approach explicitly evaluates grounded object support and is designed for adversarial object concealment scenarios.

7 Experimental Setup

7.1 Models

To evaluate the effectiveness and generality of our attack, we consider three widely used vision–language models (VLMs) with distinct architectural designs: LLaVA-1.5 which combines a CLIP-based visual encoder with the Vicuna-7B large language model, BLIP-2 paired with the Flan-T5-XL language model via a

Q-Former alignment module, and InstructBLIP which also employs a Vicuna-7B language model with instruction-tuned visual–language alignment. These models have been extensively adopted in prior work on adversarial attacks and robustness evaluation for VLMs [5, 9, 11]. They represent diverse design choices in terms of visual encoders, cross-modal fusion mechanisms, and decoding strategies. For brevity, we refer to these models as LLaVA, BLIP2-T5, and InstructBLIP throughout the remainder of the paper.

7.2 Datasets

We evaluate our attack on two standard large-scale vision datasets with object-level annotations. *ImageNet:* We randomly sample 1,000 images from the ImageNet validation set [14]. For each image, we treat a ground-truth bounding box as the region of interest (ROI) and define the attack objective as preventing the VLM from detecting or describing the object within this region. *COCO:* In addition, we evaluate our method on images from the COCO dataset [2], which contains complex multi-object scenes with diverse object categories and spatial layouts. COCO allows us to assess the robustness of our attack under more challenging conditions, including overlapping objects and crowded scenes. As with ImageNet, ground-truth bounding boxes are used to define ROIs, and the attack targets a single object category per image while preserving the remaining scene content.

7.3 Implementation Details

Unless otherwise stated, all experiments use a perturbation budget of $\epsilon = 0.2$. The loss weights are set to $\lambda_{\text{stat}} = 1$, $\lambda_{\text{dict}} = 1$, $\lambda_{\text{pres}} = 1$, and $\lambda_{\text{tv}} = 10^{-3}$, with a temperature parameter $\tau = 0.07$. BCR is applied to later vision transformer layers. Specifically, we optimize layers $\{22, 23, 24, 25\}$ for Instruct-BLIP, $\{21, 22, 23, 24\}$ for LLaVA, and $\{22, 23, 24, 25\}$ for BLIP2-T5. For BLIP-2, we apply BCR losses to the frozen ViT-g/14 vision encoder, prior to the Q-Former bottleneck. We do not operate on Q-Former tokens, as they discard spatial correspondence. The choice of targeted layers is discussed in Appendix B.

Baselines We compare our BCR attack against attention-suppression-based concealment methods, including VIP. All baseline methods are implemented using their official or publicly available code and evaluated under the same threat model and perturbation budget.

Caption Generation and Prompts. For caption-based evaluation, we adopt the same prompts used in VIP, including: **“**`Describe this picture.`**”** For binary verification tasks, we use object-specific yes/no prompts (e.g., **“**`Is there a {object} in the image?`**”**). All captions are generated using greedy decoding with a fixed maximum token budget.

Evaluation Protocol. Each attack is evaluated using the metrics introduced in Section 6, including Concealment Success, Global Preservation, Hallucination Rate, and Semantic Drift. Results are reported as averages over all test images and target objects. To isolate the effect of object concealment, all comparisons

between clean and adversarial images are performed using identical prompts and decoding settings. Further details and algorithm are presented in Appendix A.

8 Main Results

We evaluate our Background-Consistent Re-encoding (BCR) attack against existing ROI-based suppression methods, with a particular focus on object concealment effectiveness, grounded hallucination behavior, and semantic consistency. All results are averaged over the evaluation set described in Section 7. Table 1 summarizes the main quantitative results across three vision–language models. While both VIP and BCR achieve high object concealment rates, their failure modes differ substantially. VIP relies on suppressing attention and value flow from the ROI tokens, which frequently introduces a semantic gap in the visual representation. This disruption weakens the visual grounding available to the language model. As a result, the model often compensates by generating plausible but visually unsupported objects, leading to a high *grounded hallucination* rate and increased semantic drift. In contrast, BCR achieves comparable concealment performance while substantially reducing grounded hallucinations. Instead of suppressing ROI tokens, BCR re-encodes them to be statistically and semantically consistent with surrounding background tokens. This preserves representational continuity in the visual encoder and prevents the emergence of anomalous feature patterns that trigger compensatory language generation. The improvements are consistent across models and datasets. For example, on ImageNet with Instruct-BLIP, BCR reduces grounded hallucination from 0.43 (VIP) to 0.19 while increasing global preservation from 0.57 to 0.81. Similar trends are observed across BLIP2-T5 and LLaVA, as well as on the more complex COCO scenes. These results suggest that hallucination under concealment attacks is closely tied to representational discontinuity. By maintaining feature-level continuity rather than removing information outright, BCR enables effective object hiding while preserving coherent scene understanding.

Table 1: Multi-model evaluation on ImageNet and COCO subsets. **C**: Concealment success, **GP**: Global Preservation, **GH**: Grounded Hallucination rate (lower is better), **SD**: Semantic Drift.

Dataset	Method	Instruct-BLIP				BLIP2-T5				LLaVA			
		C ↑	GP ↑	GH ↓	SD ↓	C ↑	GP ↑	GH ↓	SD ↓	C ↑	GP ↑	GH ↓	SD ↓
ImageNet	No Attack	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00
	Masking	1.00	0.41	0.59	0.54	1.00	0.52	0.48	0.40	1.00	0.55	0.45	0.41
	PRM	0.66	0.38	0.62	0.44	0.70	0.31	0.69	0.41	0.73	0.34	0.66	0.39
	VIP	0.93	0.57	0.43	0.35	0.96	0.59	0.41	0.32	0.98	0.62	0.48	0.29
	BCR (Ours)	**0.92**	**0.81**	**0.19**	**0.13**	**0.95**	**0.83**	**0.17**	**0.10**	**0.98**	**0.86**	**0.14**	**0.08**
COCO	No Attack	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00
	Masking	1.00	0.37	0.63	0.46	1.00	0.39	0.61	0.44	1.00	0.41	0.59	0.41
	PRM	0.63	0.54	0.46	0.37	0.77	0.58	0.42	0.34	0.80	0.60	0.40	0.33
	VIP	0.80	0.42	0.58	0.48	0.83	0.45	0.55	0.45	0.88	0.48	0.52	0.43
	BCR (Ours)	**0.89**	**0.77**	**0.23**	**0.16**	**0.92**	**0.79**	**0.21**	**0.13**	**0.95**	**0.82**	**0.18**	**0.11**

8.1 Qualitative Comparison with Attention Suppression

We present qualitative comparisons between our Background-Consistent Re-encoding (BCR) attack and attention-suppression-based methods such as VIP. Figure 2 shows representative examples where the target object lies within the annotated region of interest (ROI). For each case, we report the clean caption together with captions generated after applying VIP and BCR.

In the first example, both methods successfully conceal the target person. However, VIP substantially alters the global scene description, introducing unrelated objects such as a *fire hydrant* and describing the scene as an abstract painting. This behavior reflects a common failure mode of attention suppression: by removing the contribution of ROI tokens, the visual representation becomes incomplete, prompting the language model to hallucinate visually unsupported content. In contrast, BCR removes the person while preserving the surrounding visual context. The generated caption continues to reference grounded elements of the scene, such as the pigeons and park bench, without introducing spurious objects. The second example highlights a different failure mode. VIP

Fig. 2: Qualitative comparison of object concealment attacks on InstructBLIP.

again removes the target object but produces a caption describing an unrelated fireworks scene, which bears little resemblance to the underlying image. This reflects severe semantic drift caused by the disruption of the visual representation. In contrast, BCR replaces the concealed object with a visually plausible marine entity (a sea urchin). Although this represents a semantic substitution rather than a strict omission, the resulting description remains consistent with the visual environment (coral reef and marine background) and does not introduce unrelated objects. Across examples, attention suppression often creates a representational void that the language model compensates for by generating plausible but visually unsupported content. BCR mitigates this effect by preserving representational continuity within the visual encoder. Instead of removing ROI tokens, BCR re-encodes them to align with the statistical structure

of surrounding background tokens. This prevents anomalous feature patterns that would otherwise trigger compensatory hallucination during language generation. Overall, BCR consistently conceals the target object while maintaining coherent scene descriptions, favoring contextually grounded substitutions over hallucinated or unrelated content. Additional qualitative results are provided in Appendix C.

8.2 Empirical Analysis of the Continuity–Hallucination Relationship

Our central hypothesis is that hallucination in object-concealment attacks is driven by representational discontinuity between the concealed region and the surrounding visual context, rather than by object removal itself. To directly evaluate this claim, we measure feature discontinuity between ROI and background tokens and analyze its relationship with grounded hallucination.

We consider two complementary discontinuity measures computed from the intermediate visual representations of the vision encoder:

- **Statistical discrepancy**: difference between ROI and background feature distributions, measured using the mean and variance alignment error (Eq.2).
- **Dictionary projection error**: the reconstruction error obtained when ROI features are projected onto the background feature dictionary (Eq. 3).

Table 2: Empirical analysis of the continuity–hallucination relationship. Lower values indicate greater representation continuity between the ROI and the surrounding background.

Method	Statistical Discrepancy ↓	Dictionary Projection Error ↓	Grounded Hallucination ↓
Clean	0.0948	1.0113	0.000
VIP	0.0773	1.0500	0.440
BCR (Ours)	**0.0297**	**0.0777**	**0.167**

Table 2 reveals a consistent relationship between representation continuity and hallucination behavior. While VIP slightly reduces statistical discrepancy relative to the clean representation, it often fails to reduce dictionary projection error and can even increase it, indicating that ROI features remain outside the background feature manifold. In contrast, BCR substantially reduces both discontinuity measures, decreasing statistical discrepancy by approximately 50–75% and dictionary projection error by up to 90%. Importantly, these reductions coincide with large decreases in grounded hallucination across all evaluated models. The results suggest that hallucination is not primarily caused by the absence of an object, but by discontinuities introduced into the visual representation when object evidence is aggressively suppressed. Maintaining continuity between ROI and background features therefore provides a mechanism for reducing hallucination while preserving concealment.

8.3 Comparison with Pixel-Space Obfuscation Baselines

Beyond attention-suppression attacks, another class of concealment baselines operates directly in pixel space, including *blurring*, *masking*, and *inpainting* of the target region. These techniques are widely used in visual privacy and anonymization, as they remove explicit visual evidence of the target object. However, their interaction with vision–language models can introduce unintended semantic artifacts. Figure 3 shows representative examples. When the target region is blurred or masked, the resulting image often contains ambiguous or unnatural patterns that the vision encoder struggles to interpret. As a result, the language model attempts to explain these signals by generating plausible but visually unsupported descriptions. For example, in a scene originally depicting chickens near a fence, pixel-space obfuscation leads the model to hallucinate unrelated objects such as dogs or cows. This illustrates a key limitation of pixel-space concealment: removing visual evidence alone can introduce ambiguity that triggers compensatory reasoning in the language model, resulting in hallucinated scene interpretations.

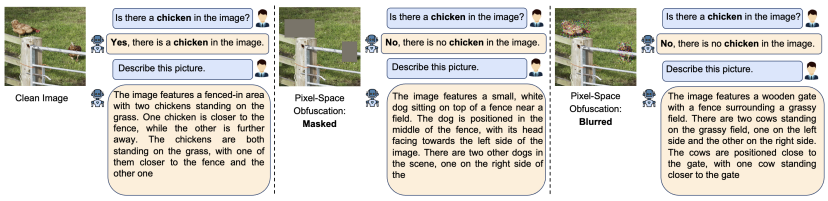

Fig. 3: Failure cases of pixel-space obfuscation methods. When the target object is masked or blurred, the resulting visual artifacts introduce ambiguous signals that vision–language models attempt to explain. Despite the absence of the original object, the model generates hallucinated descriptions containing unrelated entities such as dogs or cows.

8.4 Perceptual Fidelity

In addition to semantic preservation, effective object concealment should maintain high visual fidelity so that perturbations remain unobtrusive. We therefore measure perceptual similarity between clean and adversarial images using Structural Similarity (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS). Table 3 shows that BCR consistently achieves higher SSIM and lower LPIPS across models compared to attention-suppression attacks. Although BCR modifies pixels within the ROI, the perturbations remain visually coherent with surrounding content, avoiding the artifacts commonly introduced by suppression-based methods. These results indicate that BCR preserves both semantic consistency and perceptual realism, producing adversarial images that remain visually close to the original inputs.

Table 3: Perceptual fidelity between clean and adversarial images. Higher SSIM and lower LPIPS indicate better visual similarity.

Model	Instruct-BLIP		BLIP2-T5		LLaVA	
Method	SSIM ↑	LPIPS ↓	SSIM ↑	LPIPS ↓	SSIM ↑	LPIPS ↓
PRM	0.93	0.06	0.93	0.06	0.92	0.07
VIP	0.71	0.34	0.804	0.175	0.65	0.32
BCR (Ours)	**0.96**	**0.02**	**0.95**	**0.03**	**0.94**	**0.04**

9 Ablation Study

We conduct an ablation study to isolate the contribution of each component in BCR. All variants are evaluated on the same images and ROIs with identical optimization and evaluation settings. Table 4 shows that the full objective achieves the best balance between concealment success and semantic stability. Removing either the statistical alignment loss ($\mathcal{L}_{\text{stat}}$) or the dictionary projection loss ($\mathcal{L}_{\text{dict}}$) substantially increases hallucination, indicating that both distributional matching and semantic anchoring are necessary for stable representations. Similarly, disabling background preservation ($\mathcal{L}_{\text{pres}}$) degrades global semantics, reducing GP scores. These results confirm that BCR's effectiveness stems from enforcing representational continuity at both statistical and semantic levels.

Table 4: Ablation study of BCR components on Instruct-BLIP. ✓ indicates the loss term is enabled. We report Concealment Success (CS ↑), Hallucination Rate (HR ↓), and Global Preservation (GP ↑).

$\mathcal{L}_{\text{stat}}$	$\mathcal{L}_{\text{dict}}$	$\mathcal{L}_{\text{pres}}$	$\mathcal{L}_{\text{tv}}$	CS ↑	HR ↓	GP ↑
✓	✓	✓	✓	**0.91**	**0.08**	**0.83**
✓	✓	✓		0.90	0.09	0.81
✓	✓		✓	0.86	0.17	0.69
✓		✓	✓	0.79	0.21	0.74
	✓	✓	✓	0.72	0.28	0.66
VIP-style suppression baseline				0.88	0.42	0.51

10 Conclusion

We introduced Background-Consistent Re-encoding (BCR), a hallucination-aware object concealment framework for vision–language models. Through extensive experiments, we showed that concealment-induced hallucination is strongly associated with representational discontinuities between concealed regions and their surrounding context. By preserving continuity rather than suppressing visual evidence, BCR achieves effective object concealment while substantially reducing grounded hallucination. Our results suggest that hallucination arises largely from representational discontinuities rather than object absence itself. This insight highlights continuity-aware manipulation as a promising direction for adversarial analysis and robust multimodal perception.

Acknowledgements

This work was supported in parts by the NYUAD Center for Cyber Security (CCS), funded by Tamkeen under the NYUAD Research Institute Award G1104.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022) 1, 5
2. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1209–1218 (2018) 10
3. Datta, S., Sundararaman, D.: Evaluating hallucination in large vision-language models based on context-aware object similarities. arXiv preprint arXiv:2501.15046 (2025) 1, 2
4. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014) 3
5. Hu, A., Gu, J., Pinto, F., Kamnitsas, K., Torr, P.: As firm as their foundations: Can open-sourced foundation models be used to create adversarial examples for downstream tasks? arXiv preprint arXiv:2403.12693 (2024) 2, 4, 10
6. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023) 1, 5
7. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Chang, K.W., Gao, J.: Grounded language-image pre-training (2022), https://arxiv.org/abs/2112.03857 9
8. Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., Peng, W.: A survey on hallucination in large vision-language models. arXiv preprint arXiv:2402.00253 (2024) 1, 2, 3, 5, 9
9. Luo, H., Gu, J., Liu, F., Torr, P.: An image is worth 1000 lies: Adversarial transferability across prompts on vision-language models. arXiv preprint arXiv:2403.09766 (2024) 2, 3, 10
10. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017) 3
11. Meftah, H.F., Hamidouche, W., Fezza, S.A., Déforges, O.: Vip: Visual information protection through adversarial attacks on vision-language models. arXiv preprint arXiv:2507.08982 (2025) 2, 4, 5, 10
12. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021) 1, 5
13. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. arXiv preprint arXiv:1809.02156 (2018) 5, 9
14. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015) 10

15. Zhang, T., Wang, L., Zhang, X., Zhang, Y., Jia, B., Liang, S., Hu, S., Fu, Q., Liu, A., Liu, X.: Visual adversarial attack on vision-language models for autonomous driving. arXiv preprint arXiv:2411.18275 (2024) 2, 3, 5

A. Additional Implentation Details

A.1. Algorithm Overview

Algorithm 1 summarizes the proposed Background-Consistent Re-encoding (BCR) attack. Given an input image and a region of interest (ROI), the algorithm optimizes a pixel-level adversarial image under a bounded perturbation budget. At each iteration, visual features are extracted from selected layers of the vision encoder. ROI token representations are then encouraged to (i) match the first- and second-order statistics of background tokens, (ii) lie within the background feature manifold via soft dictionary projection, and (iii) preserve non-ROI features to maintain global scene semantics. A total variation regularizer enforces spatial smoothness of the perturbation. The optimization proceeds via gradient descent in pixel space, producing an adversarial image that conceals the target object while preserving semantic continuity and reducing hallucination.

Algorithm 1 Background-Consistent Re-encoding (BCR)

Require: Image $\mathbf{x}$, ROI boxes $\mathcal{B}$, vision encoder f_θ, layers $\mathcal{L}$, step size η, steps T, perturbation budget ϵ
Ensure: Adversarial image $\mathbf{x}^{adv}$

Compute ROI pixel mask $\mathbf{M}$ from $\mathcal{B}$
Compute ROI token indices $\mathcal{I}_r$ and background indices $\mathcal{I}_b$
Initialize $\mathbf{x}^{adv} \leftarrow \mathbf{x}$
for each layer $l \in \mathcal{L}$ **do**
  Store clean features $\mathbf{Z}^{(l)}(\mathbf{x})$
end for
for $t = 1$ to T **do**
  for each layer $l \in \mathcal{L}$ **do**
    Extract ROI features $\mathbf{Z}_r^{(l)}$ and background features $\mathbf{Z}_b^{(l)}$
    Compute statistical loss $\mathcal{L}_{\text{stat}}^{(l)}$
    Compute dictionary projection loss $\mathcal{L}_{\text{dict}}^{(l)}$
    Compute background preservation loss $\mathcal{L}_{\text{pres}}^{(l)}$
  end for
  Compute total variation loss $\mathcal{L}_{\text{tv}}$ on ROI
  $\mathcal{L} \leftarrow \sum_{l \in \mathcal{L}} \left(\lambda_{\text{stat}} \mathcal{L}_{\text{stat}}^{(l)} + \lambda_{\text{dict}} \mathcal{L}_{\text{dict}}^{(l)} + \lambda_{\text{pres}} \mathcal{L}_{\text{pres}}^{(l)} \right) + \lambda_{\text{tv}} \mathcal{L}_{\text{tv}}$
  Update $\mathbf{x}^{adv}$ using gradient descent on $\mathcal{L}$
  Project perturbation to ℓ_∞ budget ϵ
end for
return $\mathbf{x}^{adv}$

A.2. Grounded Hallucination Verification with GLIP

To ensure that hallucination measurements reflect the presence of visually unsupported objects rather than linguistic variation, we augment our caption-based evaluation with a grounding-based verification step.

Caption comparison alone can overestimate hallucination due to paraphrasing or synonym usage. For example, an object may appear in a caption under a different lexical form (e.g., "automobile" vs. "car"). Conversely, models may introduce object terms that are not visually supported in the image. To disambiguate these cases, we verify whether newly mentioned objects can be grounded in the image.

Given a clean image I and its adversarial counterpart I', we generate captions $c = f(I)$ and $c' = f(I')$ using the same prompt and decoding settings. Object candidates are extracted from both captions using a dependency-based noun phrase parser. After lemmatization and normalization, we obtain the object sets

$$O(c), \quad O(c').$$

The set of newly introduced objects is defined as

$$H_{\text{cand}} = O(c') \setminus O(c).$$

For each candidate object $o \in H_{\text{cand}}$, we query the Grounded Language–Image Pretraining (GLIP) detector to verify whether the object can be localized in the adversarial image I'. GLIP predicts bounding boxes conditioned on the textual query corresponding to the object name. If no detection with confidence above a predefined threshold τ_g is produced, the object is considered visually unsupported.

Formally, the hallucination set is defined as

$$H = \{\, o \in H_{\text{cand}} \mid \text{GLIP}(I', o) = \varnothing \,\}.$$

Using the verified hallucination set H, we compute the grounded hallucination rate as

$$\text{HallucinationRate}_{\text{grounded}} = \frac{|H|}{|O(c')|}.$$

This metric measures the fraction of objects mentioned in the adversarial caption that cannot be visually grounded in the image.

We use a publicly available GLIP model pretrained on large-scale grounding datasets. The grounding confidence threshold is set to $\tau_g = 0.3$. For multi-word phrases, the head noun is used as the textual query to improve grounding robustness. This grounding-based verification reduces ambiguity in hallucination evaluation by ensuring that newly introduced object tokens correspond to detectable visual evidence. By combining caption-based object extraction with grounding verification, the evaluation more accurately captures hallucinations caused by adversarial manipulation rather than lexical variation.

B. Sensitivity Analysis of Targeted Transformer Layers

Our Background-Consistent Re-encoding (BCR) objective enforces statistical and semantic alignment between region-of-interest (ROI) tokens and background tokens across a subset of vision transformer layers. The choice of targeted layers influences the strength and stability of the concealment attack, as different layers encode visual information at varying levels of abstraction.

Early layers of vision transformers primarily capture low-level features such as edges, textures, and local patterns, while deeper layers encode higher-level semantic representations that are more directly aligned with language generation. Consequently, applying the BCR objective at different depths may affect the degree to which object semantics are suppressed or redistributed into background representations.

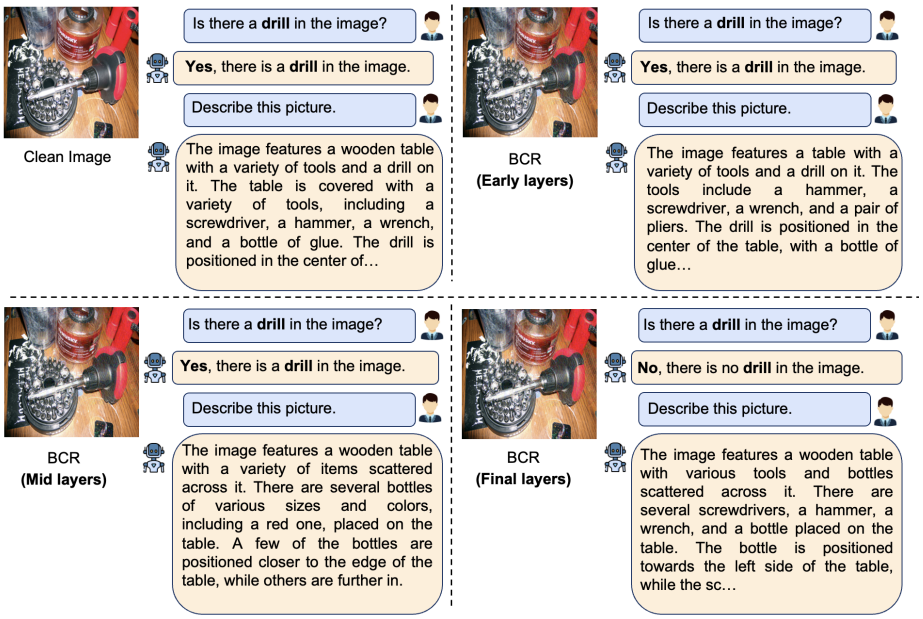

Fig. 4: Sensitivity analysis of targeted transformer layers. Early-layer optimization leaves object semantics largely intact, allowing the model to still recognize the target object. In contrast, targeting deeper layers removes the object semantics while preserving the surrounding scene context, resulting in successful concealment.

To study this effect, we evaluate BCR by applying the alignment losses to different groups of transformer layers:

- **Early layers**: losses applied to the first four transformer blocks.
- **Middle layers**: losses applied to intermediate transformer blocks.

– **Late layers**: losses applied to the final four transformer blocks.

All other hyperparameters remain identical to the main experiments, including the perturbation budget and optimization schedule.

We additionally measured concealment success and hallucination rate across layer configurations, confirming that targeting deeper layers yields the most stable concealment behavior. Early-layer optimization leaves object semantics intact, allowing the model to still recognize the target object. In contrast, targeting deeper layers removes the object semantics while preserving the surrounding scene context.

Table 5: Quantitative sensitivity analysis of targeted transformer layers on InstructBLIP. Later-layer optimization yields stronger concealment and lower hallucination.

Targeted Layers	**Concealment Success ↑**	**Hallucination Rate ↓**
Early Layers	0.22	0.07
Middle Layers	0.41	0.11
Late Layers	**0.91**	**0.18**

These results indicate that hallucination and object recognition in vision–language models are primarily driven by higher-level semantic representations. By aligning ROI features with background features in deeper layers, BCR effectively removes object-specific semantics while preserving global scene coherence. Based on this analysis, all experiments in the main paper apply the BCR objective to later transformer layers of the vision encoder.

C. Additional Qualitative Results

We present additional qualitative examples illustrating the behavior of BCR across different scenes and target objects (see Figure 5). For each example, we compare captions generated from the clean image and the adversarial images produced by different concealment strategies.

Across diverse scenes, suppression-based approaches frequently produce unstable descriptions. Because these methods remove or weaken ROI representations, the resulting visual features become semantically inconsistent with the surrounding context. The language model then compensates by generating plausible but unsupported objects, leading to hallucinated descriptions.

In contrast, BCR maintains representational continuity by re-encoding ROI tokens to match the statistical and semantic properties of background tokens. As a result, the generated captions remain visually grounded and coherent, even after the target object is concealed.

These additional examples further demonstrate that hallucination in vision–language models is closely tied to representational discontinuities introduced

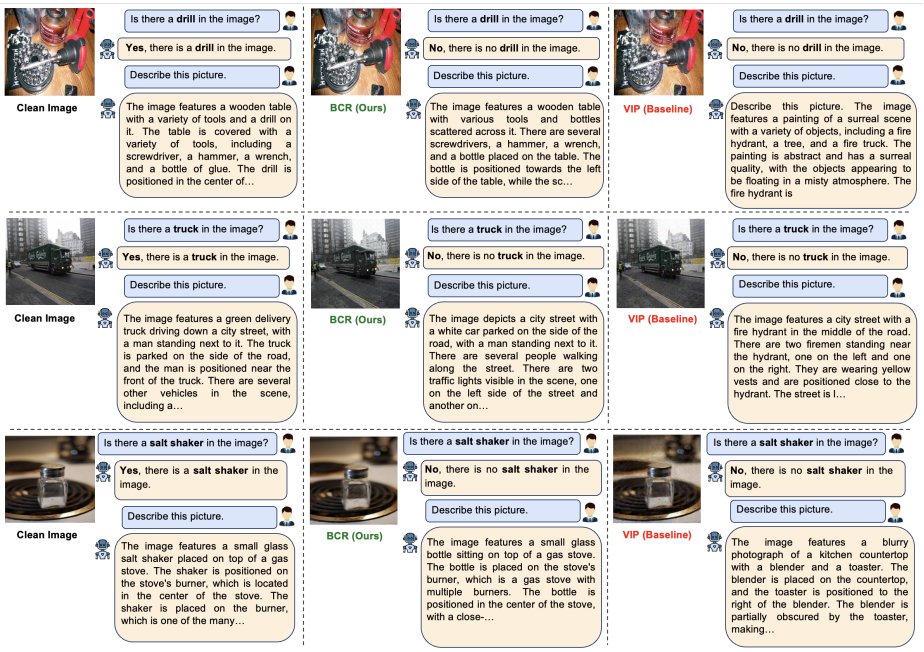

Fig. 5: Additional qualitative comparisons between suppression-based attacks and BCR. While suppression-based methods often introduce hallucinated or semantically unrelated objects, BCR consistently removes the target object while preserving the overall scene structure and contextual elements.

by suppression-based attacks, while continuity-preserving manipulation enables more stable and realistic scene descriptions.

D. Failure Cases and Limitations

While BCR substantially reduces hallucination compared to suppression-based attacks, it is not without limitations. We summarize the main failure modes observed in our experiments and discuss directions for future improvement.

Partial Semantic Substitution. In some cases, BCR does not fully eliminate object-level reasoning but instead induces *semantic substitution*. For example, small or visually ambiguous objects (e.g., accessories such as ties or utensils) may be re-encoded as semantically adjacent background objects (e.g., clothing regions or tableware). Although this behavior avoids hallucination, it may still allow indirect inference of the concealed object category through contextual cues. This limitation is inherent to continuity-preserving attacks that aim to maintain global coherence rather than explicitly erase object evidence.

Model-Specific Vision Tokenization. Different VLMs employ distinct vision backbones and tokenization strategies (e.g., patch size, cropping, or pooling), which can affect ROI-to-token alignment. While BCR generalizes across LLaVA, BLIP-2, and Instruct-BLIP, additional model-specific adjustments may be required for architectures with aggressive image cropping or non-uniform token layouts.

Scope of Concealment. BCR is designed to conceal objects while preserving overall scene semantics, not to guarantee absolute object removal under all possible queries. Highly targeted or adversarial prompts explicitly probing the ROI (e.g., repeated yes/no questioning) may still extract residual information. Addressing such interactive threat models remains an open problem.

These limitations highlight the inherent trade-off between concealment and semantic continuity. We believe BCR represents a principled step toward hallucination-aware adversarial attacks, and future work may combine continuity-based objectives with prompt-aware or adaptive strategies to further strengthen object concealment.