Appendix

A. MIE Algorithm Flow

Algorithm 1 MIE

```
Require: Large vision-language model $f_\theta$; Clean image $\boldsymbol{x}$; Query $\boldsymbol{q}$
Require: Pertubation bound $\epsilon$; Iteration steps $S$; Learning rate $\alpha$
Ensure: Adversarial image $\boldsymbol{x}^{adv}$
Initialise $\boldsymbol{x}^{adv} = \boldsymbol{x}$
Enabel gradients for variable $\boldsymbol{x}^{adv}$
for each step in 1, 2, ..., S do
    $p_i, \mathcal{A}_i, h_i = f_\theta(\boldsymbol{x}^{adv}, q)$
    Calculate $\mathcal{L}_1$ using Equation 2
    Calculate $\mathcal{L}_2$ using Equation 3
    Calculate $\mathcal{L}_3$ using Equation 4
    $\mathcal{L} = \lambda_1\mathcal{L}_1 + \lambda_2\mathcal{L}_2 + \lambda_3\mathcal{L}_3$
    $\boldsymbol{x}^{adv} = \boldsymbol{x}^{adv} + \alpha \times sign(\nabla_{\boldsymbol{x}^{adv}}\mathcal{L})$
    $\delta = \text{Clip}(\boldsymbol{x}^{adv} - \boldsymbol{x}, -\epsilon, \epsilon)$
    $\boldsymbol{x}^{adv} = \text{CLIP}(\boldsymbol{x} + \epsilon)$
end for
return $\boldsymbol{x}^{adv}$
```

B. Experimental Setup

For the models used in our experiments, we provide a more detailed description below.

For the BLIP model, we utilize the version **blip-image-captioning-base**, which is capable of performing image captioning tasks without requiring additional textual instructions. BLIP effectively leverages noisy web data by bootstrapping captions: a captioner generates synthetic captions, while a filter removes the noisy ones. For the BLIP-2 model, we employ the version **blip2-opt-2.7b**, which utilizes OPT as its language model. BLIP-2 introduces the use of a Q-former to efficiently extract image features and reduce the number of visual tokens. For the InstructBLIP model, we utilize the version **instructblip-vicuna-7b**, which employs Vicuna as its language model. InstructBLIP utilizes 26 public datasets for instruction tuning, thereby attaining strong instruction-following capabilities. For the LLaVA1.5 model, we use the version **llava-1.5-7b-hf**. For the LLaVA1.6 model, we use the version **llava-v1.6-vicuna-7b**. LLaVA is currently the most commonly used baseline for vision-language models and is widely studied. For the CLIP model, we use **clip-vit-base-patch32**. **MiniGPT-4** aligns a frozen visual encoder from BLIP-2 with a frozen LLM, Vicuna, using just one projection layer. All the mentioned models can be accessed on the **Hugging Face** official website.

In the phase of manual evaluation of attack success rates, our focus is on whether the models generate hallucinations, ensuring factual correctness. Therefore, personal biases are minimized to a great extent. We primarily involve two individuals independently conducting statistical calculations to determine their success rates.

C. Changes in Internal Features

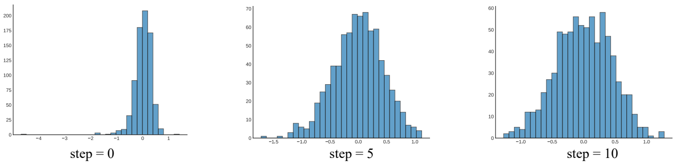


Fig. 5: Visualization of internal features at different attack steps.

We also visualize changes in internal features. The model's representation learning capability is severely compromised due to the intrusion. As shown in Figure 5, the internal features learned by the model from clean images exhibit a relatively concentrated distribution, which becomes dispersed following an attack. Previous white-box attack methods often target attacks based solely on model output. We believe this approach is insufficient, particularly for multimodal large models with deep neural network layers, which hampers the effective generation of adversarial examples [26]. Therefore, by simultaneously considering both the model's internal features and its output results, we can optimize and generate more robust visual adversarial examples.

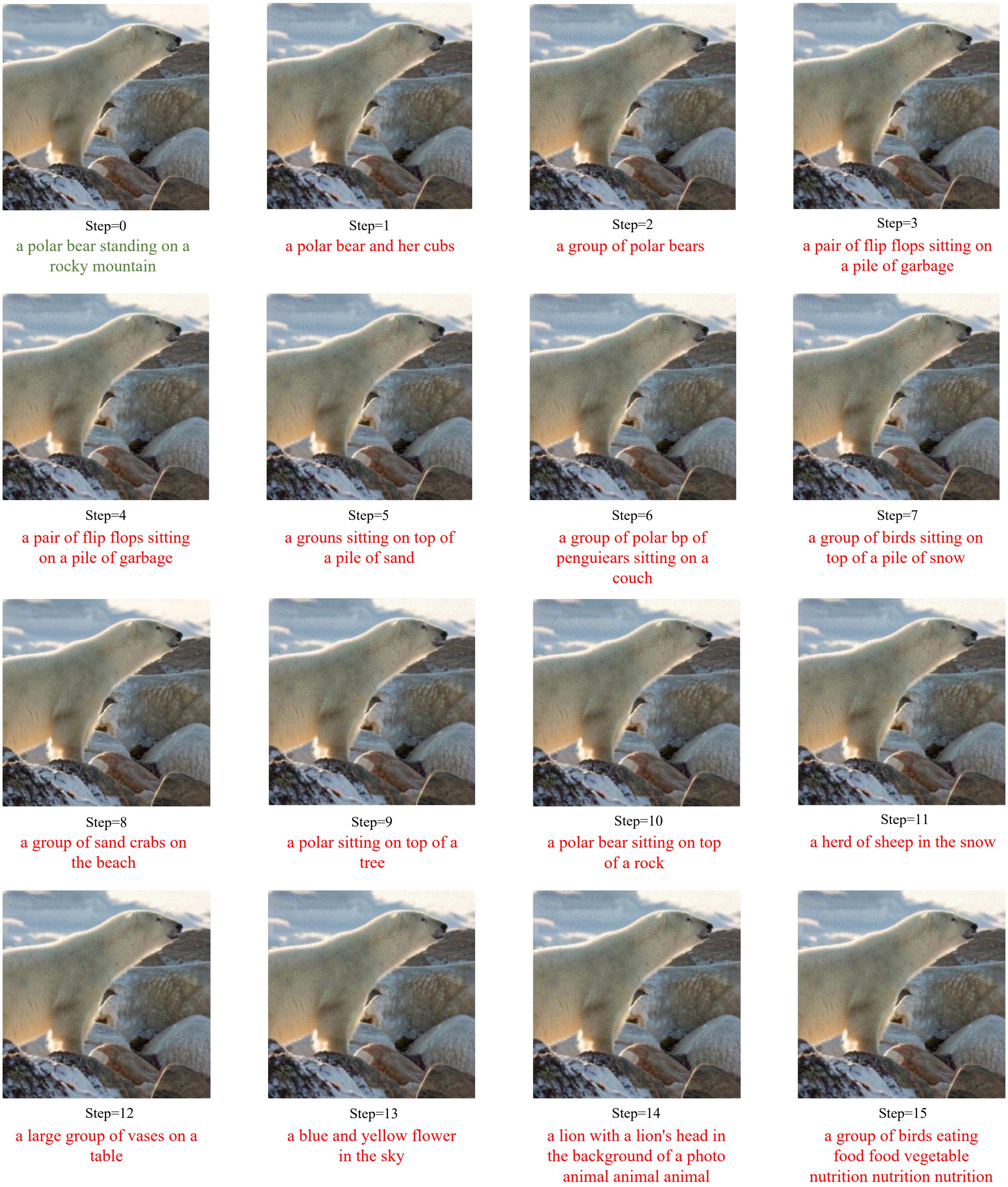
D. More Cases

In this section, we demonstrate the efficacy of the proposed MIE method through a comprehensive set of examples.

To begin, we illustrate through a case study that as the number of iterations increases, the effectiveness of MIE attacks gradually strengthens, as shown in Figure 6. Initially, the model merely generates hallucinations related to the image. Eventually, the model produces content that completely lacks syntactic coherence, indicating a loss of its fundamental language capabilities. From the perspective of deceiving model users, generating grammatically logical but hallucinatory content is harder to detect, which can be achieved by reducing the number of iterations. Conversely, from the standpoint of the model's intrinsic robustness, producing grammatically incorrect errors indicates significant damage to the model's cognitive abilities, which can be achieved by increasing the number of iterations. A large number of specific cases and statistical results can demonstrate the effectiveness of MIE in evaluating the adversarial robustness of VLMs.

As demonstrated in the accompanying figures below, our model can generate adversarial examples with minimal noise, rendering them visually indistinguishable from the original images to the human eye. A fundamental assumption in adversarial example research is that the perturbation should be imperceptible to the human visual system. However, many existing adversarial attacks against vision-language models deviate from this principle to some extent, introducing noticeable noise to the original image that compromises its visual fidelity.

As demonstrated in the provided examples, our non-targeted attack adheres strictly to the criterion of imperceptibility to the human visual system. This characteristic enhances its practical utility in real-world applications, where targeted attacks may introduce visually noticeable artifacts.



Step=0
a polar bear standing on a rocky mountain

Step=1
a polar bear and her cubs

Step=2
a group of polar bears

Step=3
a pair of flip flops sitting on a pile of garbage

Step=4
a pair of flip flops sitting on a pile of garbage

Step=5
a grouns sitting on top of a pile of sand

Step=6
a group of polar bp of penguiears sitting on a couch

Step=7
a group of birds sitting on top of a pile of snow

Step=8
a group of sand crabs on the beach

Step=9
a polar sitting on top of a tree

Step=10
a polar bear sitting on top of a rock

Step=11
a herd of sheep in the snow

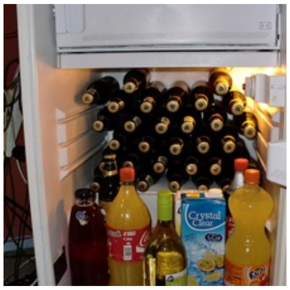
Step=12
a large group of vases on a table

Step=13
a blue and yellow flower in the sky

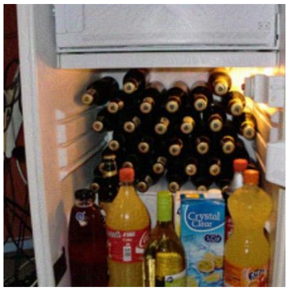
Step=14
a lion with a lion's head in the background of a photo animal animal animal

Step=15
a group of birds eating food food vegetable nutrition nutrition nutrition

Fig. 6: Example showcase of step-by-step attacks. From a semantic perspective, the model's understanding of the images progressively deteriorates, culminating in incoherent sentences.

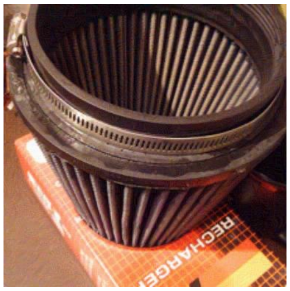


clean image
a refrigerator with a bunch of soda bottles in it



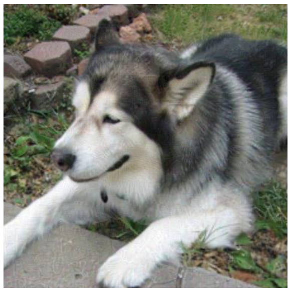
adversarial image
a sign that reads the best beer in the world

clean image
a black air filter on top of a box



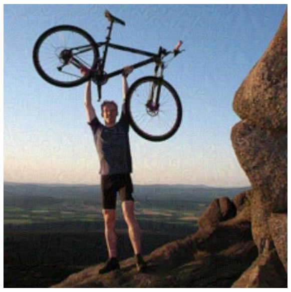
adversarial image
the new e - z - e - z - z - e - z - z - z

clean image
a dog laying on the ground



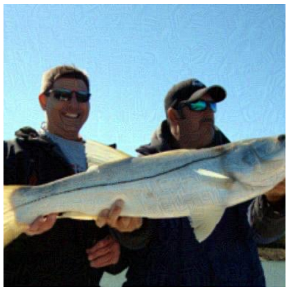
adversarial image
a set of drill drills and a drill bit

clean image
a man holding a bike on top of a mountain



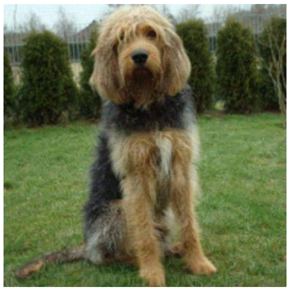
adversarial image
a poster for the new movie, the big bad

clean image
two men holding a fish on a boat



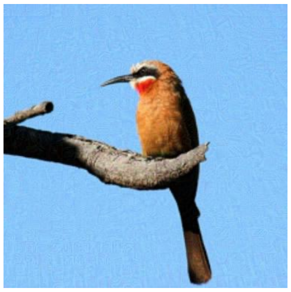
adversarial image
a white and black cake with a white and black cake top

clean image
a shaggy dog sitting on the grass



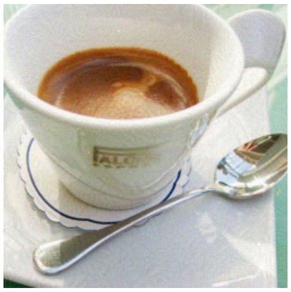
adversarial image
a sign that reads 'no smoking'

clean image
a bird sitting on a branch



adversarial image
a poster with a photo of a man in a suit and tie

clean image
a cup of coffee



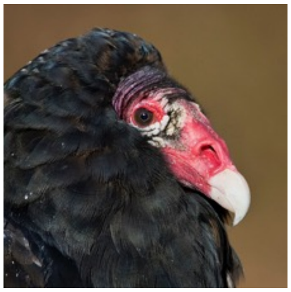
adversarial image
a white shirt with the words 'i'm' on it

clean image
a man wearing a gas mask

adversarial image
the back support system

clean image
a man with a guitar on stage

adversarial image
the new logo for the new game, the new game



clean image
a black bird with a red beak and a white head

adversarial image
a group of people with their heads in a circle

clean image
the parliament building in townsville

adversarial image
a truck with the words 'the best way to get a job'

clean image
a room with a view of a garden

adversarial image
a - e - e - e - e - e - e - e - e

clean image
a bunch of trains

adversarial image
steel roof flashings for volvo v40

clean image
a drum and a drum stick sitting on a table

adversarial image
a bottle of wine with a label for the wine

clean image
a yellow van parked in a field

adversarial image
a man with a laptop and a laptop on the grass