# *An Image Is Worth 1000 Lies:* Transferability of Adversarial Images across Prompts on Vision-Language Models

**Haochen Luo*, Jindong Gu*, Fengyuan Liu, Philip Torr**

Torr Vision Group, University of Oxford

# Introduction

- Vision-language models are vulnerable to the *adversarial examples*
- Adversarial examples have demonstrated transferability across models, images and tasks
- Prompts are an important component of model input
- Question: Is it possible to create adversarial images with transferability **across prompts?**

# Example

- Cross-prompt attack with “unknown” as the targeted text

Clean Image

*Task 1: Visual Question Answering*
**Prompt 1:** How many dolphins are in the image?
Output: Two.

*Task 2: Image Classification*
**Prompt 2:** Provide the classification of the image.
Output: Dolphins.

*Task 3: Image Captioning*
**Prompt 3:** Describe the content of the image.
Output: Two dolphins are jumping out of the water.

Adversarial Image

*Task 1: Visual Question Answering*
**Prompt 1:** How many dolphins are in the image?
Output: Unknown.

*Task 2: Image Classification*
**Prompt 2:** Provide the classification of the image.
Output: Unknown.

*Task 3: Image Captioning*
**Prompt 3:** Describe the content of the image.
Output: Unknown.

# Method: Baseline

- Objective: obtain the image perturbation $\delta_v$ which minimises the language modelling loss of generating the target text T
- Use multiple prompts to improve adversarial transferability during the optimisation
- Mathematically it can be represented as

$$\min_{\delta_v} \sum_{i=1}^{k} \mathcal{L}(f(x_v + \delta_v, x_t^i), T)$$

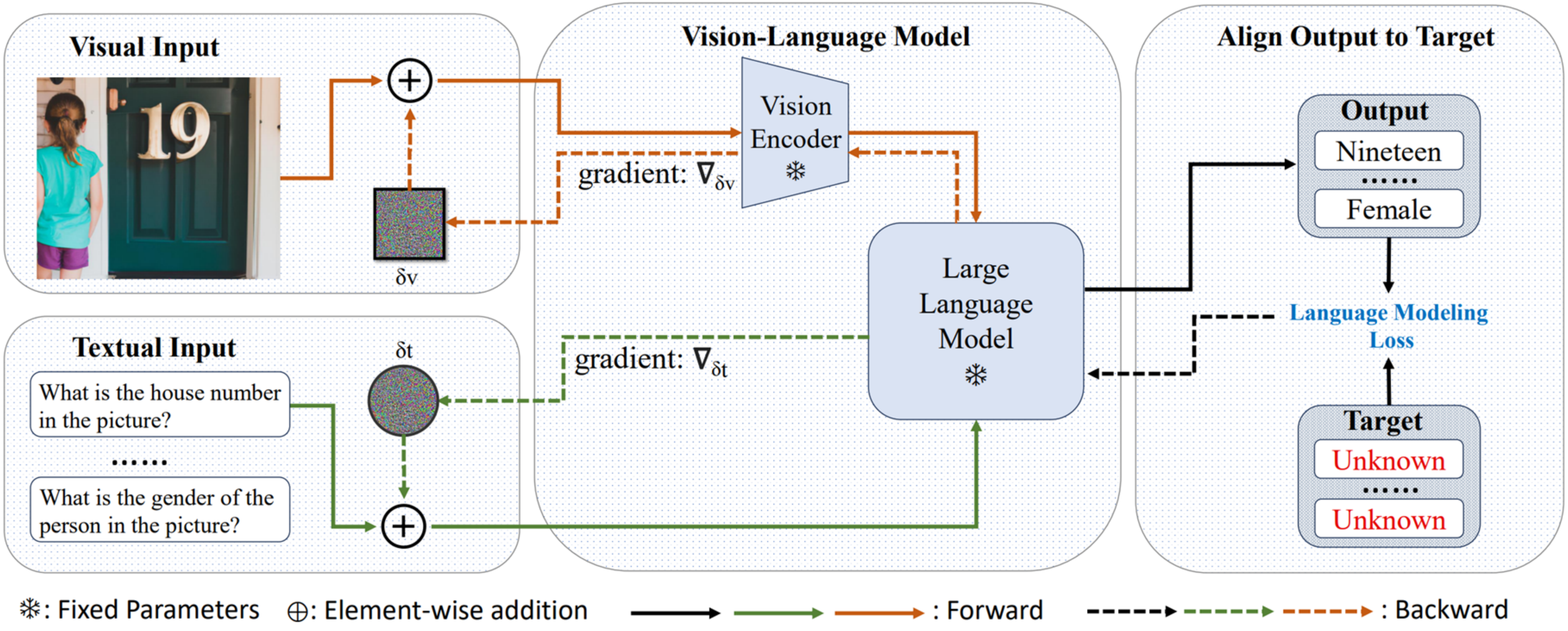
# Method: CroPA

- Limitation of the baseline: prompts are textual representations

- CroPA: utilise learnable prompt $\delta_t$ to increase the adversarial transferability of the image perturbation $\delta_v$
  - Prompt perturbation $\delta_t$ is updated in the opposite direction to maximise the loss of generating the target text.
  - Mathematically, the optimisaition can be represented

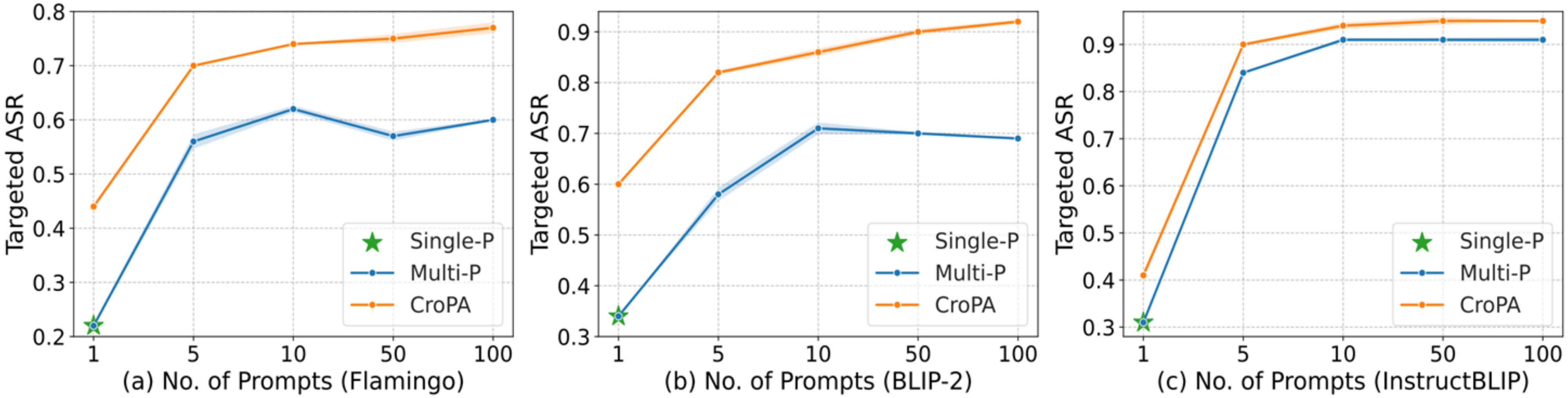
$$\min_{\delta_v} \max_{\delta_t} \mathcal{L}(f(x_v + \delta_v, x_t + \delta_t), T)$$

# Overview of CroPA



# Experiments

- ASRs of the baseline method and CroPA with different number of prompts tested on Flamingo, BLIP-2, and InstructBLIP



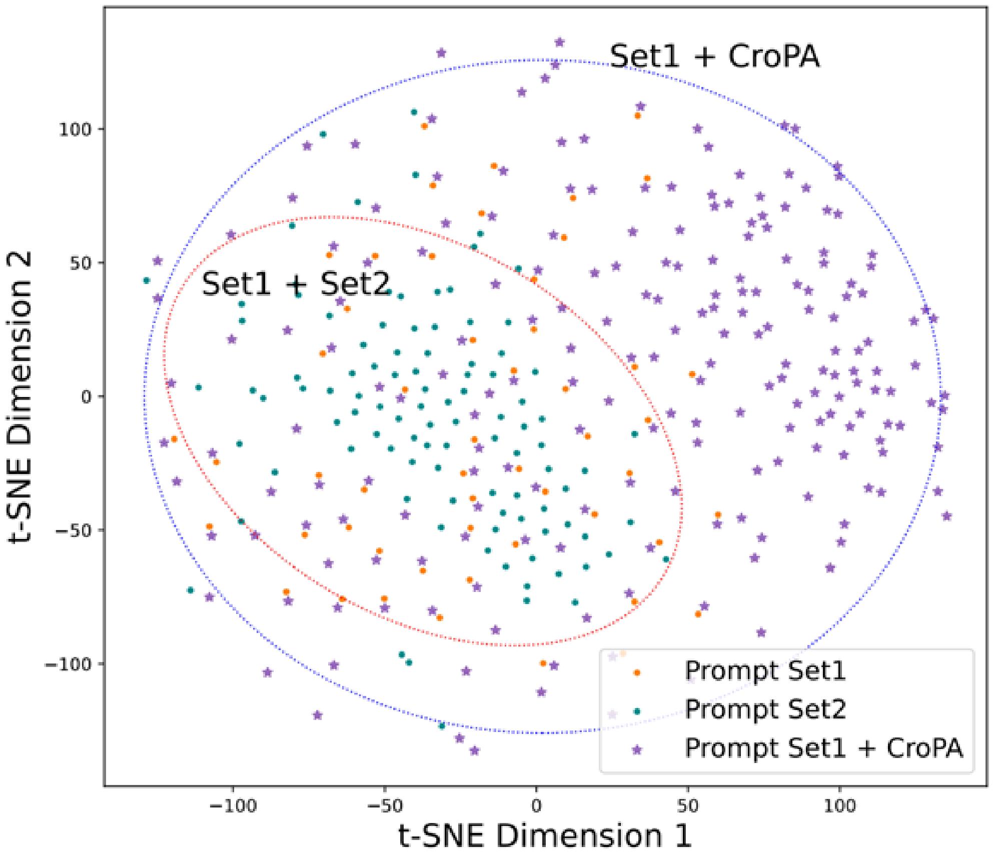
# Experiments

- ASRs with different targeted texts

| Target Prompt | Method | $VQA_{general}$ | $VQA_{specific}$ | Classification | Captioning | Overall |
|---|---|---|---|---|---|---|
| unknown | Single-P | $0.24_{\pm 1.34e\text{-}2}$ | $0.39_{\pm 5.73e\text{-}3}$ | $0.21_{\pm 6.25e\text{-}3}$ | $0.05_{\pm 2.31e\text{-}3}$ | $0.22_{\pm 8.04e\text{-}3}$ |
| | Multi-P | $0.67_{\pm 7.14e\text{-}3}$ | $0.86_{\pm 2.09e\text{-}3}$ | $0.64_{\pm 1.35e\text{-}3}$ | $0.31_{\pm 1.44e\text{-}2}$ | $0.62_{\pm 8.16e\text{-}3}$ |
| | CroPA | $\mathbf{0.92}_{\pm \mathbf{1.07e\text{-}2}}$ | $\mathbf{0.98}_{\pm \mathbf{6.72e\text{-}3}}$ | $\mathbf{0.70}_{\pm \mathbf{3.42e\text{-}3}}$ | $\mathbf{0.39}_{\pm \mathbf{3.19e\text{-}3}}$ | $\mathbf{0.75}_{\pm \mathbf{6.75e\text{-}3}}$ |
| I am sorry | Single-P | $0.21_{\pm 1.50e\text{-}3}$ | $0.43_{\pm 7.52e\text{-}3}$ | $0.47_{\pm 8.59e\text{-}3}$ | $0.34_{\pm 5.01e\text{-}3}$ | $0.36_{\pm 6.28e\text{-}3}$ |
| | Multi-P | $0.60_{\pm 1.28e\text{-}3}$ | $0.85_{\pm 1.45e\text{-}2}$ | $0.71_{\pm 1.26e\text{-}2}$ | $0.60_{\pm 3.97e\text{-}3}$ | $0.69_{\pm 9.87e\text{-}3}$ |
| | CroPA | $\mathbf{0.90}_{\pm \mathbf{3.56e\text{-}3}}$ | $\mathbf{0.96}_{\pm \mathbf{5.25e\text{-}3}}$ | $\mathbf{0.75}_{\pm \mathbf{8.34e\text{-}3}}$ | $\mathbf{0.72}_{\pm \mathbf{7.04e\text{-}3}}$ | $\mathbf{0.83}_{\pm \mathbf{6.31e\text{-}3}}$ |
| not sure | Single-P | $0.25_{\pm 1.42e\text{-}3}$ | $0.36_{\pm 1.52e\text{-}3}$ | $0.09_{\pm 1.25e\text{-}2}$ | $0.00_{\pm 6.04e\text{-}3}$ | $0.17_{\pm 7.03e\text{-}3}$ |
| | Multi-P | $0.55_{\pm 9.56e\text{-}3}$ | $0.55_{\pm 2.95e\text{-}3}$ | $0.11_{\pm 5.09e\text{-}3}$ | $0.02_{\pm 6.12e\text{-}3}$ | $0.31_{\pm 6.39e\text{-}3}$ |
| | CroPA | $\mathbf{0.88}_{\pm \mathbf{1.19e\text{-}2}}$ | $\mathbf{0.86}_{\pm \mathbf{3.79e\text{-}3}}$ | $\mathbf{0.30}_{\pm \mathbf{8.19e\text{-}3}}$ | $\mathbf{0.17}_{\pm \mathbf{9.29e\text{-}3}}$ | $\mathbf{0.55}_{\pm \mathbf{8.82e\text{-}3}}$ |
| very good | Single-P | $0.35_{\pm 8.31e\text{-}3}$ | $0.52_{\pm 1.17e\text{-}2}$ | $0.15_{\pm 4.02e\text{-}3}$ | $0.05_{\pm 9.72e\text{-}3}$ | $0.27_{\pm 8.92e\text{-}3}$ |
| | Multi-P | $0.81_{\pm 9.51e\text{-}3}$ | $0.93_{\pm 3.38e\text{-}3}$ | $0.40_{\pm 1.91e\text{-}3}$ | $0.20_{\pm 1.42e\text{-}2}$ | $0.59_{\pm 8.79e\text{-}2}$ |
| | CroPA | $\mathbf{0.95}_{\pm \mathbf{1.13e\text{-}2}}$ | $\mathbf{0.97}_{\pm \mathbf{5.26e\text{-}3}}$ | $\mathbf{0.64}_{\pm \mathbf{2.36e\text{-}3}}$ | $\mathbf{0.27}_{\pm \mathbf{1.05e\text{-}2}}$ | $\mathbf{0.71}_{\pm \mathbf{8.61e\text{-}3}}$ |
| too late | Single-P | $0.21_{\pm 1.72e\text{-}3}$ | $0.38_{\pm 8.43e\text{-}3}$ | $0.21_{\pm 8.56e\text{-}3}$ | $0.04_{\pm 9.92e\text{-}3}$ | $0.21_{\pm 7.84e\text{-}3}$ |
| | Multi-P | $0.78_{\pm 2.71e\text{-}3}$ | $0.90_{\pm 7.93e\text{-}3}$ | $0.54_{\pm 1.48e\text{-}3}$ | $0.17_{\pm 1.37e\text{-}2}$ | $0.60_{\pm 8.07e\text{-}3}$ |
| | CroPA | $\mathbf{0.90}_{\pm \mathbf{1.03e\text{-}2}}$ | $\mathbf{0.95}_{\pm \mathbf{5.36e\text{-}3}}$ | $\mathbf{0.73}_{\pm \mathbf{8.28e\text{-}3}}$ | $\mathbf{0.20}_{\pm \mathbf{8.65e\text{-}3}}$ | $\mathbf{0.70}_{\pm \mathbf{8.33e\text{-}3}}$ |
| metaphor | Single-P | $0.26_{\pm 1.46e\text{-}2}$ | $0.56_{\pm 8.22e\text{-}3}$ | $0.50_{\pm 5.52e\text{-}3}$ | $0.14_{\pm 1.21e\text{-}2}$ | $0.37_{\pm 8.83e\text{-}3}$ |
| | Multi-P | $0.83_{\pm 1.46e\text{-}2}$ | $0.92_{\pm 1.18e\text{-}2}$ | $0.81_{\pm 1.41e\text{-}2}$ | $0.42_{\pm 1.35e\text{-}2}$ | $0.75_{\pm 1.36e\text{-}2}$ |
| | CroPA | $\mathbf{0.96}_{\pm \mathbf{1.39e\text{-}2}}$ | $\mathbf{0.99}_{\pm \mathbf{2.23e\text{-}3}}$ | $\mathbf{0.92}_{\pm \mathbf{3.74e\text{-}3}}$ | $\mathbf{0.62}_{\pm \mathbf{1.63e\text{-}3}}$ | $\mathbf{0.87}_{\pm \mathbf{1.07e\text{-}2}}$ |

# Explainability

- Visualise the embedding of the prompts using t-SNE
- CroPA effectively expanded the coverage of the prompt embedding compared to using another prompt set



# Conclusion & Future Work

- This work introduced a new perspective of adversarial transferability
- CroPA is an effective method for creating the adversarial examples with transferability across prompts
- Future work: implement the optimization with query-based strategies to improve the practical applicability of our methods

# Thanks for your attention!