Pandora's Box: Towards Building Universal Attackers against Real-World Large Vision-Language Models

Daizong Liu[1], Mingyu Yang[2], Xiaoye Qu[2], Pan Zhou[2], Xiang Fang[3], Keke Tang[4], Yao Wan[2], Lichao Sun[5]
[1]Peking University [2]Huazhong University of Science & Technology [3]Nanyang Technological University
[4]Guangzhou University [5]Lehigh University

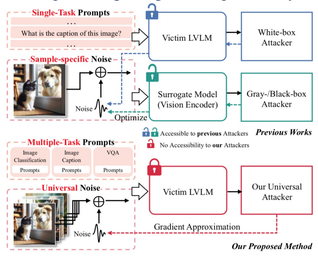
ABSTRACT

Large Vision-Language Models (LVLMs) have demonstrated remarkable capabilities across a wide range of multimodal understanding tasks. Nevertheless, these models are susceptible to adversarial examples.

- **Previous challenges:** In real-world applications, existing LVLM attackers generally rely on the detailed prior knowledge of the model to generate effective perturbations. Moreover, these attacks are task-specific, leading to significant costs for designing perturbation.
- **Our motivation:** Motivated by the research gap and practical demands, in this paper, we make the first attempt to build a universal attacker against real-world LVLMs, focusing on two critical aspects: (i) restricting access to only the LVLM inputs and outputs. (ii) devising a universal adversarial patch, which is task-agnostic and can deceive any LVLM-driven task when applied to various inputs.
- **Our method:** We start by initializing the location and the pattern of the adversarial patch through random sampling, guided by the semantic distance between their output and the target label. Subsequently, we maintain a consistent patch location while refining the pattern to enhance semantic resemblance to the target. In particular, our approach incorporates a diverse set of LVLM task inputs as query samples to approximate the patch gradient, capitalizing on the importance of distinct inputs. In this way, the optimized patch is universally adversarial against different tasks and prompts, leveraging solely gradient estimates queried from the model.

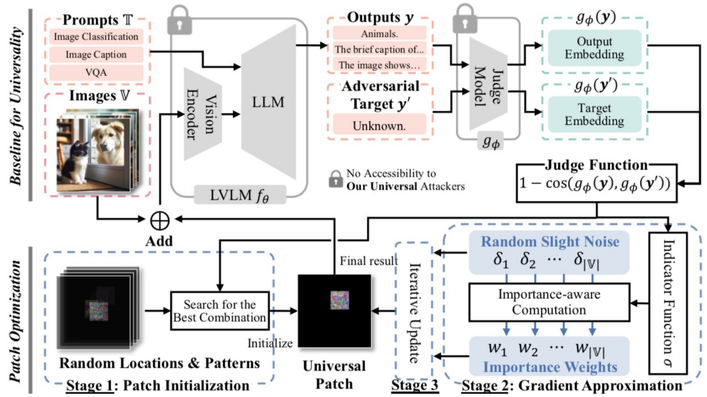
MOTIVATION

Motivation: Although the existing LVLM attackers demonstrate significant performance against LVLMs, we argue that they fail to consider the essential characteristics of attack practicality and universality among various realistic downstream multimodal tasks: (1) Existing white-, gray- and black-box methods severely rely on the prior model knowledge, making the attacks less practical since most real-world LVLM applications will not disclose their model details with users. Under such circumstances, the attackers can only query LVLMs to obtain corresponding output results, making it challenging to steer the adversarial perturbations in the correct optimization direction during the gradient estimation process. (2) LVLMs demonstrate impressive versatility in addressing diverse vision-language tasks through varying prompts. However, the current attackers targeting LVLMs can only produce adversarial examples to deceive a particular task within a singular process. Consequently, to compromise different downstream tasks, they must generate distinct adversarial perturbations, which incur significant time and resource expenditure. Therefore, it is efficient and effective to design a universal perturbation for all samples across different tasks. Upon applying this universal perturbation to any input sample, regardless of the task, it has the capability to mislead the LVLM into predicting a target label specified by the attacker.



METHODS

Overview: Overview of the proposed universal adversarial attack against real-world LVLM models. To make the perturbation universally adversarial to multiple LVLM downstream tasks, we design a special patch-wise perturbation pattern by first initializing it on a fixed location of various image inputs and then optimizing it against images of different tasks. To update the adversarial patch by solely querying the LVLM, we introduce a language-based judge model to evaluate the LVLM output and design a novel importance-aware gradient approximation strategy to adaptively estimate gradients and adjust weights on gradient directions for optimizing the perturbations on input samples.



- Universal adversarial objective

$$\mathcal{J}(\boldsymbol{y}, \boldsymbol{y}') = 1 - \cos(g_\phi(\boldsymbol{y}), g_\phi(\boldsymbol{y}')).$$

$$\min_{\boldsymbol{\Delta}} \frac{1}{K} \sum_{k=1}^{K} \mathcal{J}(\boldsymbol{y}_k, \boldsymbol{y}').$$

$$\min_{\boldsymbol{\Delta}} \frac{1}{|\mathbb{V}|K} \sum_{i=1}^{|\mathbb{V}|} \sum_{k=1}^{K} 1 - \cos(g_\phi(f_\theta(\boldsymbol{v}_i', \boldsymbol{t}_k)), g_\phi(\boldsymbol{y}'))$$

- Importance-aware gradient approximation

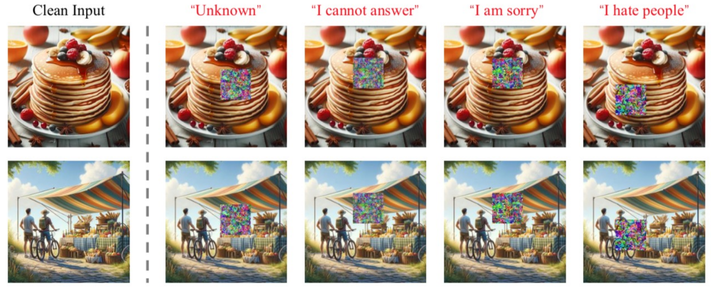
$$\sigma_t = \operatorname{sgn}\Big(\frac{1}{K}\sum_{k=1}^{K} \cos(g_\phi(f_\theta(\boldsymbol{v}_i', \boldsymbol{t}_k)), g_\phi(\boldsymbol{y}')) - \tau\Big) \quad s.t. \quad \boldsymbol{v}' = (1-\boldsymbol{m}) \odot \boldsymbol{v} + \boldsymbol{m} \odot (\boldsymbol{\Delta} + \boldsymbol{\delta}_t).$$

$$w_t = \begin{cases} \exp(\Delta\sigma_t)/\gamma, & \text{if} \quad \Delta\sigma_t > 0 \quad \text{and} \quad \text{avg}(\sigma_t) = 1 \\ \exp(\Delta\sigma_t), & \text{if} \quad \Delta\sigma_t > 0 \quad \text{and} \quad \text{avg}(\sigma_t) \neq 1 \\ \ln(\Delta\sigma_t + 3), & \text{otherwise} \end{cases}$$

$$\nabla_{\boldsymbol{\delta}_t} = \begin{cases} \text{avg}(\sigma_t) \cdot \text{avg}(\boldsymbol{\delta}_t), & \text{if} \quad \text{avg}(\sigma_t) = 1 \quad \text{or} \quad \text{avg}(\sigma_t) = -1 \\ \text{avg}((\sigma_t - \text{avg}(\sigma_t)) \cdot \boldsymbol{\delta}_t), & \text{otherwise} \end{cases},$$

RESULTS

- Qualitative results



- Quantitative comparison

LVLM Model	Attack Method	ImageClassification	ImageCaption	VQA	Overall
Dataset: MS-COCO					
LLaVA	Clean image	0.385	0.479	0.436	0.433
	w/o importance	0.703	0.679	0.711	0.698
	Full attack	**0.850**	**0.812**	**0.828**	**0.830**
MiniGPT-4	Clean image	0.438	0.451	0.463	0.450
	w/o importance	0.713	0.670	0.719	0.701
	Full attack	**0.847**	**0.826**	**0.851**	**0.841**
Flamingo	Clean image	0.475	0.468	0.492	0.478
	w/o importance	0.705	0.693	0.727	0.709
	Full attack	**0.862**	**0.803**	**0.839**	**0.835**
BLIP-2	Clean image	0.409	0.436	0.447	0.431
	w/o importance	0.724	0.682	0.716	0.707
	Full attack	**0.810**	**0.787**	**0.845**	**0.814**
Dataset: DALLE-3					
LLaVA	Clean image	0.407	0.453	0.517	0.459
	w/o importance	0.644	0.692	0.751	0.696
	Full attack	**0.824**	**0.806**	**0.879**	**0.837**
MiniGPT-4	Clean image	0.396	0.441	0.497	0.445
	w/o importance	0.682	0.738	0.714	0.711
	Full attack	**0.810**	**0.843**	**0.862**	**0.838**
Flamingo	Clean image	0.431	0.464	0.485	0.460
	w/o importance	0.719	0.746	0.742	0.735
	Full attack	**0.823**	**0.871**	**0.838**	**0.844**
BLIP-2	Clean image	0.368	0.425	0.466	0.419
	w/o importance	0.673	0.759	0.733	0.721
	Full attack	**0.795**	**0.837**	**0.840**	**0.824**

- Ablation study

Table 2: Attack performance on LLaVA model and DALLE-3 dataset with different target labels.

Adversarial Target	Attack Method	ImageClassification	ImageCaption	VQA	Overall
"Unknown"	w/o importance	0.644	0.692	0.751	0.696
	Full attack	**0.824**	**0.806**	**0.879**	**0.837**
"I cannot answer"	w/o importance	0.627	0.688	0.723	0.679
	Full attack	**0.816**	**0.835**	**0.862**	**0.844**
"I am sorry"	w/o importance	0.648	0.674	0.735	0.686
	Full attack	**0.845**	**0.813**	**0.868**	**0.842**
"I hate people"	w/o importance	0.593	0.639	0.664	0.632
	Full attack	**0.682**	**0.710**	**0.756**	**0.716**

Table 3: Comparison with existing LVLM attack: MF-Attack [33]. For a fair comparison, experiments are conducted on the same ImageNet-1k dataset [83] in the VQA task.

Method	Attack Type	LLaVA	BLIP-2	MiniGPT-4	Average
MF-Attack [33]	transfer-based black-box attack	0.590	0.681	0.668	0.646
Ours	universal and practical attack	**0.734**	**0.756**	**0.692**	**0.727**

Table 4: Comparison with existing LVLM attack: CroPA [31]. For a fair comparison, we follow CroPA to evaluate the same ASR metric on the same OpenFlamingo model and MS-COCO dataset.

Method	Attack Type	ImageClassification	ImageCaption	VQA	Overall
CroPA [31]	white-box attack	0.70	0.34	**0.92**	0.65
Ours	universal and practical attack	**0.73**	**0.51**	0.84	**0.69**

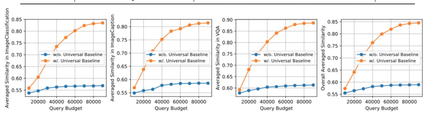


Figure 3: Analysis on the attack "Universality" on LLaVa model and DALLE-3 dataset.

CONCLUSIONS

In this paper, we propose to attack the real-world large vision-language models (LVLMs) in a practical but challenging setting, where the attacker can solely query the LVLM model. To make the perturbation universally adversarial to multiple LVLM-driven tasks, we design a universal adversarial patch with specific locations to perturb the visual inputs. By solely querying the model to estimate the gradient direction for optimizing the adversarial patch pattern, we develop a novel importance-aware gradient approximation strategy to adaptively estimate and adjust the weights on gradient directions for optimizing different samples. Experiments show the effectiveness of the proposed attack method.

ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China (NSFC) under grant No. 62476107.