LLM multimodal / multilingual harm responses

Team 3

A Pilot Study of Query-Free Adversarial Attack against Stable Diffusion

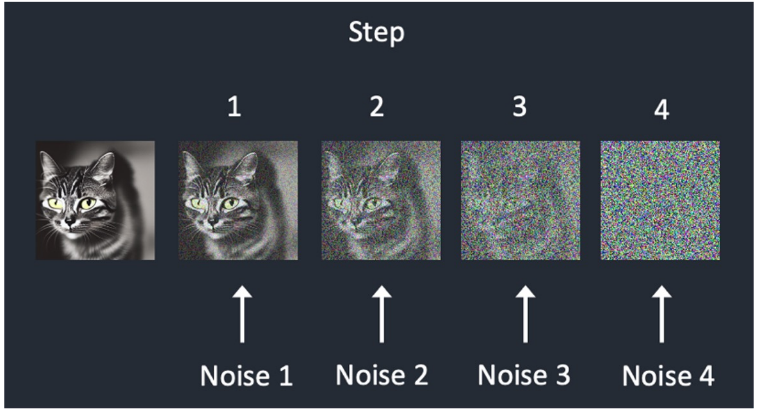
Shihe Wang (qvw9pv)

Adversarial Attacks

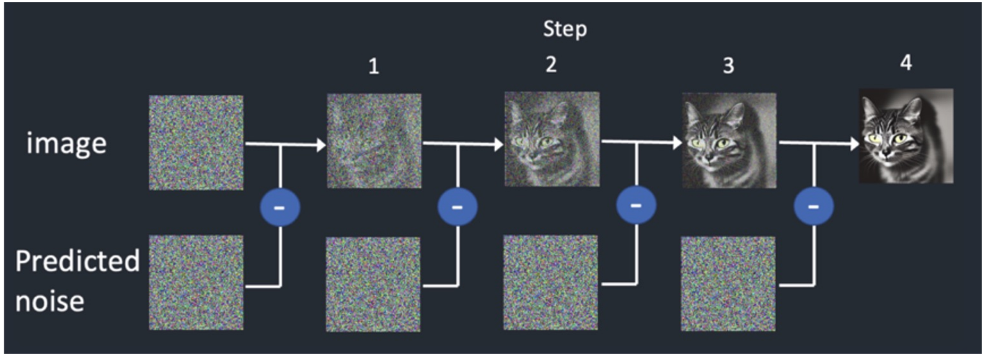
A black bicycle against a brick wall

A black bicycle against a brick wall E$9\'

Diffusion



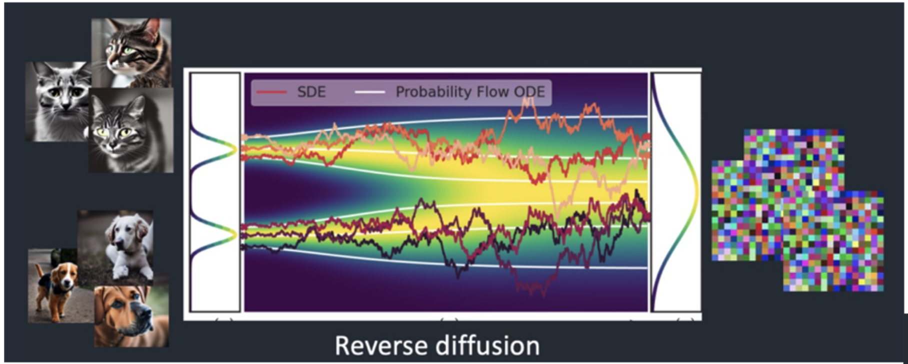
Forward Diffusion



Reverse Diffusion

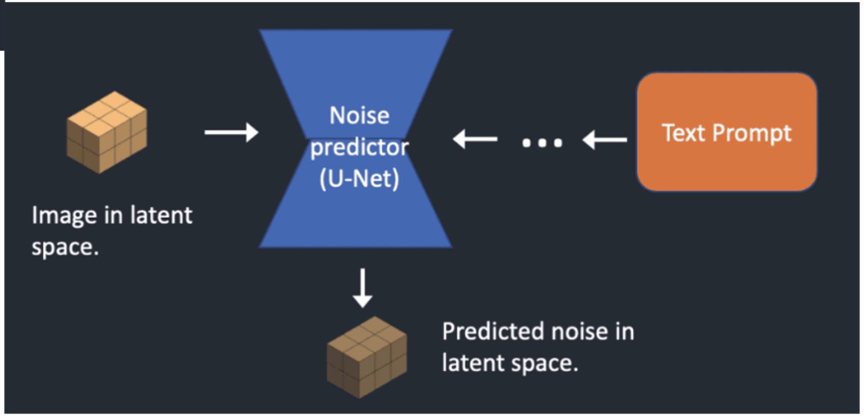
Diffusion Model(*Stable Diffusion*)

Control diffusion and condition using a text encoder (clip)



Able to "denoise" but not able to control the reverse diffusion result, can be dog or cat.

Provide text prompt as input to the noise predictor, extra input to provide guidance, using cross attention.

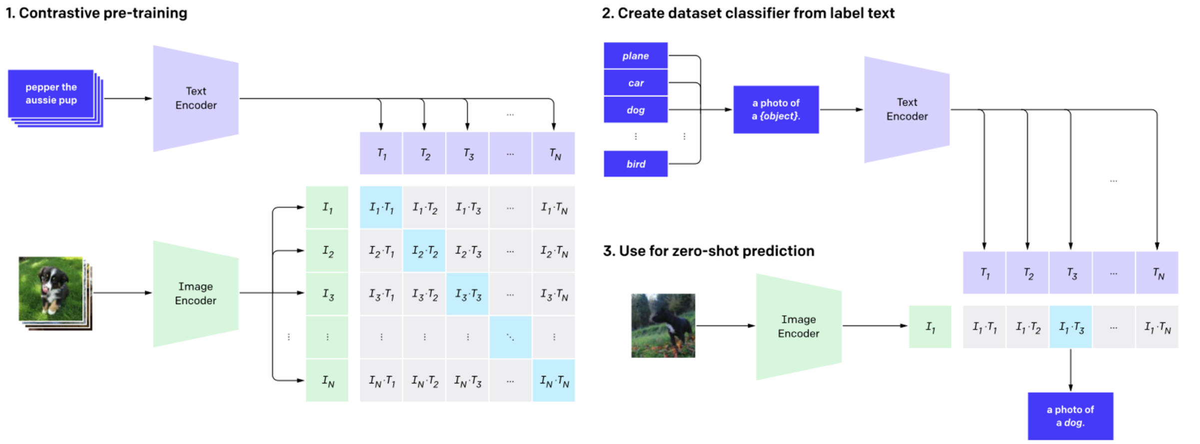


CLIP(*Contrastive Language–Image Pre-training*)

Connects texts and images.

CLIP is trained over WebImage Text(WIT) 400M image-text pair, CLIP’s learns by what this image is not about. Less “cheating”.

If another class is added, there is no need to design new training data and add another output head.



How to generate adversarial perturbations

Query-based adversarial attacks were proposed for T2I models, calls for many model queries (10000 queries per attack) to find a successful adversarial prompt.

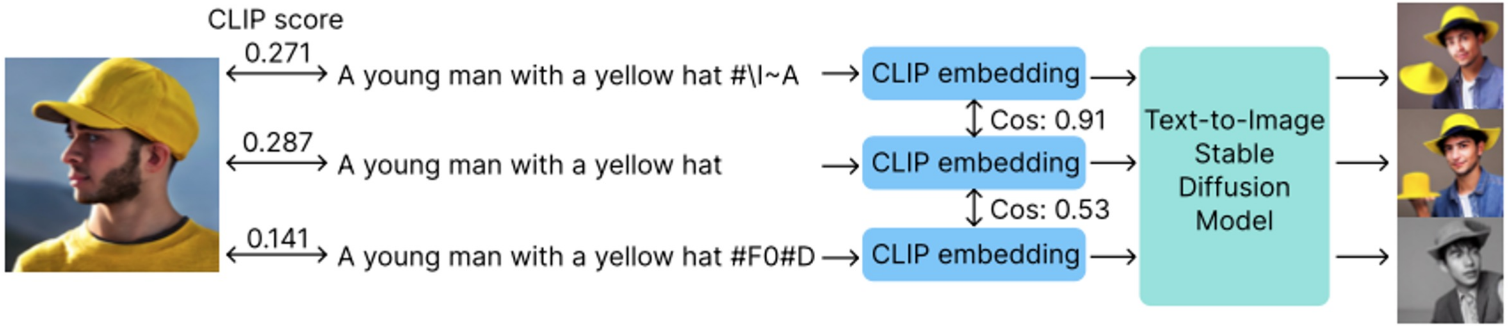
(a) Images generated by Stable Diffusion v1.5 [59] using the prompts **(Left)** '`a picture of a mountain`' and **(Right)** '`turbo lhaff✓a picture of a mountain`', found by our method. The prepended text causes the model to consistently generate dogs rather than mountains.

Query-Free Adversarial Attack against Stable Diffusion

Assuming the attacker have access to the text encoder but not the diffusion model. Attack without executing the diffusion process which would take a high model query and computation cost.

Small perturbations on the text input of CLIP can lead to different CLIP scores, because of the sensitivity of the CLIP's text embedding to text perturbations.

Query-free; Small(a five-character) perturbation; Attack on CLIP



Attack Model

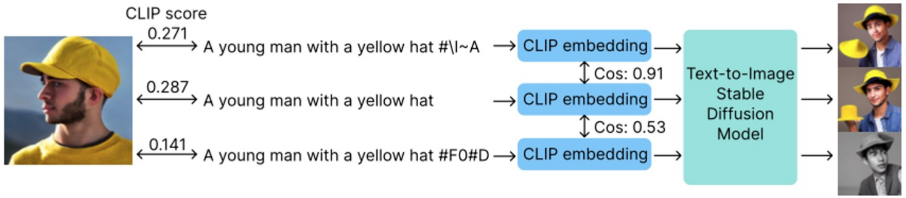
τθ(x) denote the text encoder of CLIP with parameters θ evaluated at the textual input x, find x’ that minimizing the cosine similarity between the text embeddings of x and x’.

x and x’ are independent from the diffusion model.

In this attack model, there is no target specified

$$\min_{\boldsymbol{x}'} \cos(\tau_\theta(\boldsymbol{x}), \tau_\theta(\boldsymbol{x}'))$$

Targeted Attack Model



Targeting at removing the “yellow hat”

Attack generated can be further refined towards a targeted attack purpose by guide the attack generator with steerable key dimensions.

$$\min_{x'} \ \cos(\tau_\theta(x) \odot I, \tau_\theta(x') \odot I)$$

How to find key dimension?

1. Generate n simple scenes and end with “with a yellow hat” s and n without

   s1 = ‘A bird flew high in the sky with a yellow hat’ and s2 = ‘The sun set over the horizon with a yellow hat”

   s′1 = ‘A bird flew high in the sky’ and s′2 = ‘The sun set over the horizon’.

1. Obtain the corresponding CLIP embeddings {τθ (si)} and {τθ (s′i)} .

   The text embedding difference di = τθ (si) − τθ (s′i) can characterize the saliency of the adversary’s intention-related sub-sentence

1. Find the binary vector I that represent the most influential dimensions

   Ij : $\left|\sum_{i=1}^{n} \mathrm{sign}(d_{i,j})\right| > \epsilon n$

Attack Methods

Attack models are differentiable can use optimization methods

1. PGD(projected gradient descent): incorporates a perturbation budget (ϵ) and a step size (α) to control the amount and direction of perturbation

   $x'_{t+1} = \Pi(x_t + \alpha \cdot \text{sign}(\nabla_x J(\Theta, x_t, y)))$, where, x_t is the input at iteration t, α is the step size, $\nabla_x J(\Theta, x_t, y)$ is the gradient of the loss with respect to the input

1. Greedy search: a greedy search on the character candidate set to select the top 5 characters
2. Genetic algorithm: In each iteration, the genetic algorithm calls genetic operations such as mutation to generate new candidates

   Details on implementation: https://github.com/OPTML-Group/QF-Attack/blob/main/utils.py

Experiment Setup

Stable Diffusion model v1.4 as the victim model for image generation.

Attack methods details:

PGD: the base learning rate by 0.1 and the number of PGD steps by 100.

Genetic algorithm: the number of generation steps 50, the number of candidates per step 20, and the mutation rate 0.3

Targeted attack: ChatGPT to generate $n = 10$ sentences to characterize the steerable key dimensions and set $\varepsilon = 0.9$ to determine the influence mask I

Experiment Results

Table 1. CLIP scores [13, 14] comparison of images generated with different methods. CLIP scores are used to indicate the similarity between the generated images and the embeddings of the corresponding text prompts. For each method, the CLIP scores reported below are averaged over 20 prompts and 10 images per prompt. In particular, the scores calculated based on the original sentences and output images are adopted for the untargeted attack and based on the targeted content and output images for the targeted setting. The lowest (best) score in each row is in **bold** and the results in the form $a \pm b$ denote the mean value a and the standard deviation b.

Method:	**No Attack**	**Random**	**Greedy**	**Genetic**	**PGD**
			Untargeted Attack		
Score:	0.277±0.022	0.271±0.021	0.255±0.039	**0.203**±0.042	0.226±0.041
			Targeted Attack		
Score:	0.229±0.03	0.223±0.037	0.204±0.037	**0.186**±0.04	0.189±0.041

Experiment Results

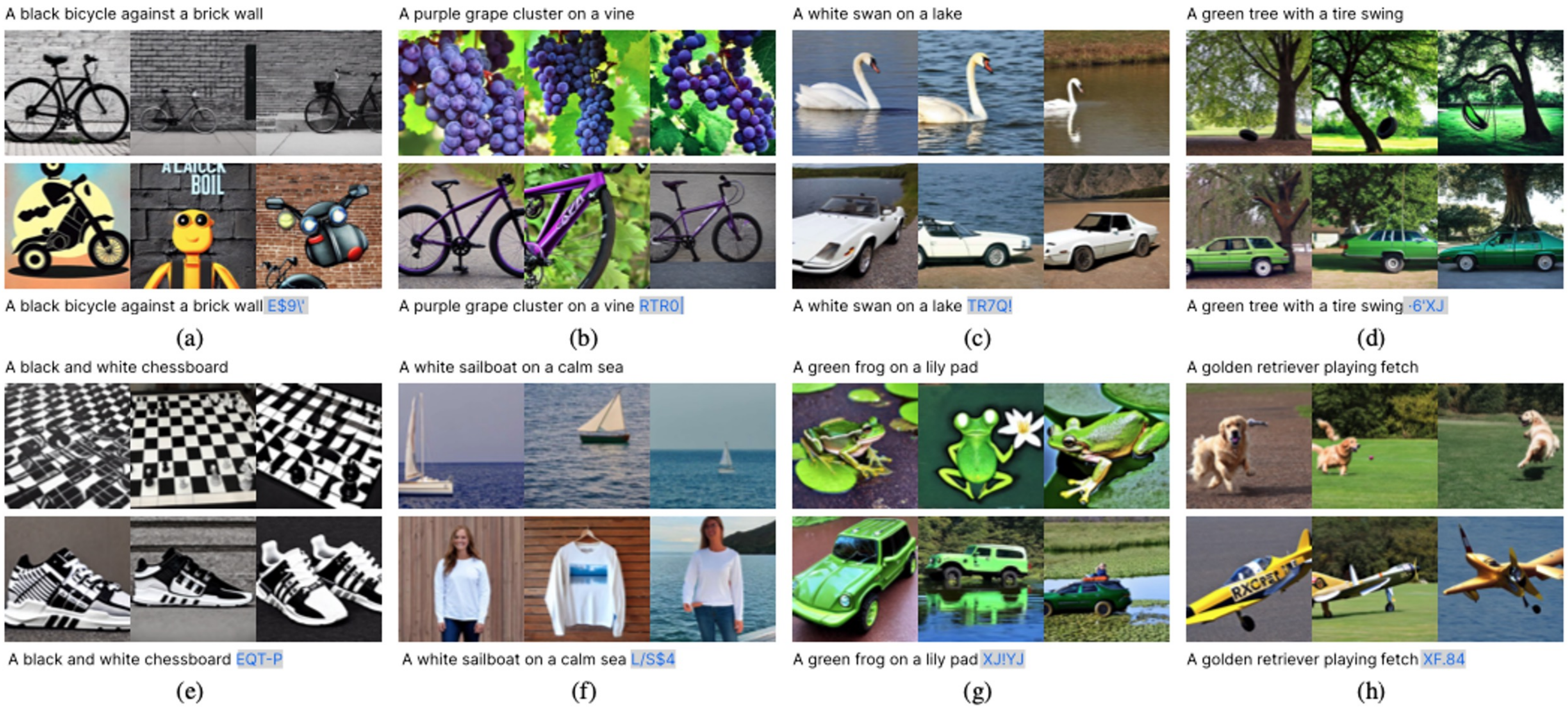


Figure 3. Illustrations of the effect of *untargeted* query-free attacks. In each group, the first row of images is generated using the original prompts vs. the second row using the perturbed ones. The perturbations found by our method are highlighted in blue in the prompt. Images in the same column share the same random seed.

Experiment Results

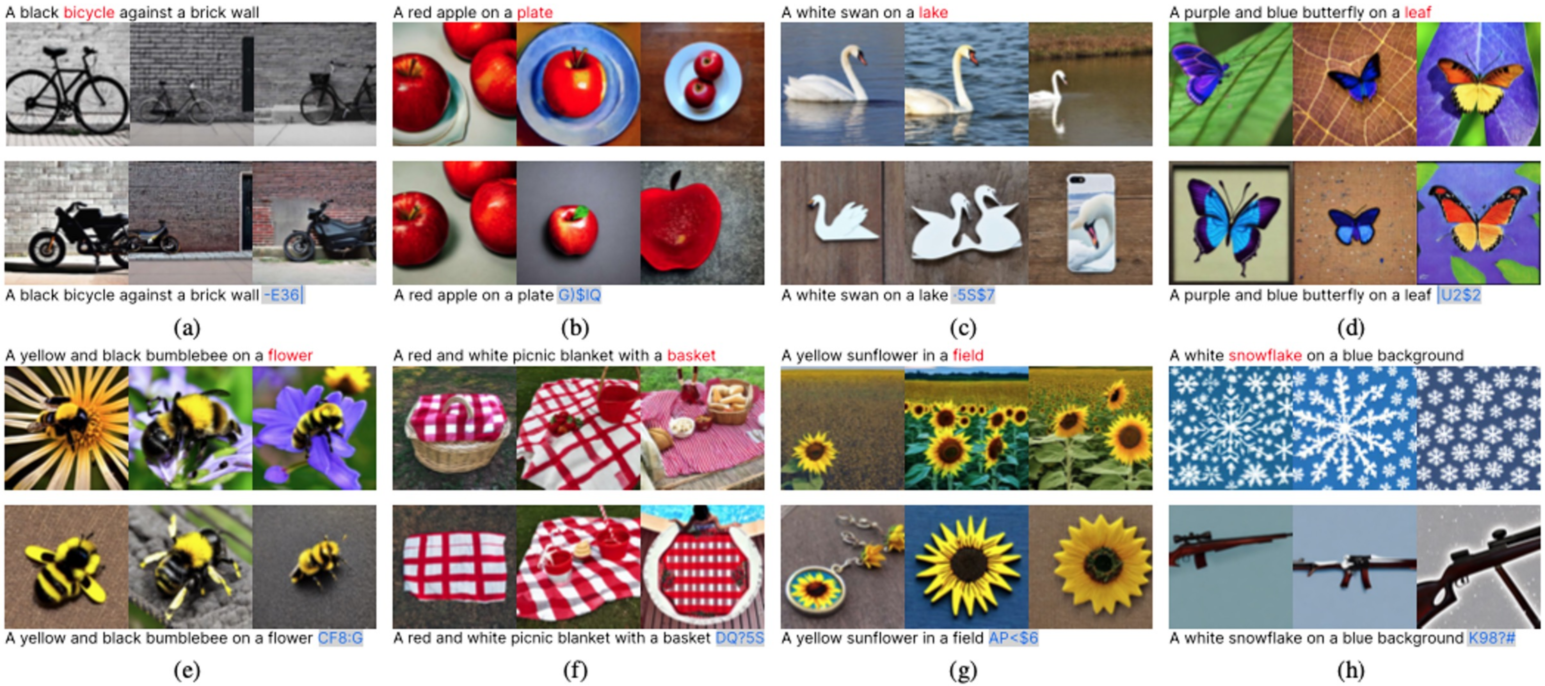
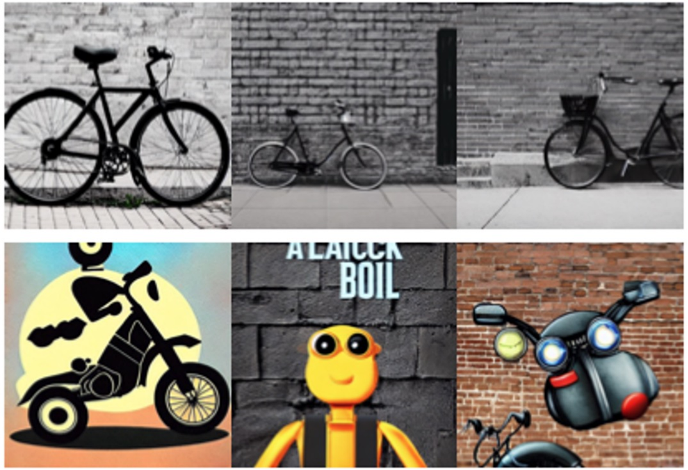


Figure 4. Illustrations of the effect of *targeted* query-free attacks. Input perturbations are generated to modify/remove the red text-related image content. Other settings are aligned with Fig. 3. Adversary targets for erasing (a) the 'bike', (b) the 'plate', (c) the 'lake', (d) the 'leaf', (e) the 'flower', (f) the 'basket', (g) the 'field', and (h) the 'snowflake' without altering the other semantics much.

Case Study on Wall-E

A black bicycle against a brick wall



A black bicycle against a brick wall E$9\'

A black bicycle against a brick wall E

A black bicycle against a brick wall WALLE

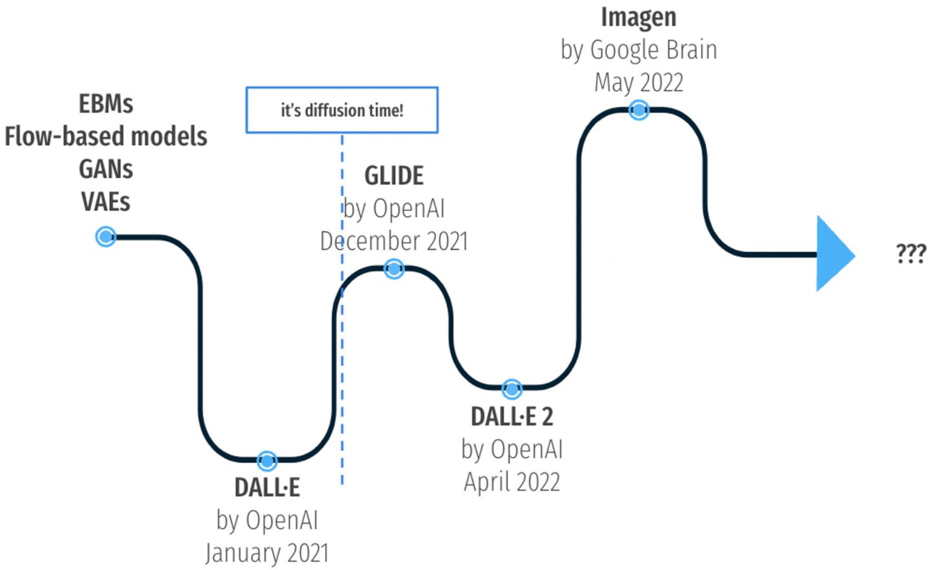
A black bicycle against a brick wall-E

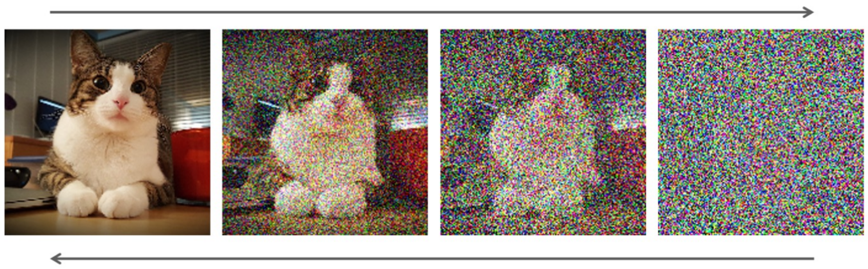
Cheating Suffix: Targeted Attack to Text-To-Image Diffusion Models with Multi-Modal Priors

Zhiyang Yuan (vfr4pr)

Diffusion Models in Image Generation

- Image generation revolutionized by diffusion models
- Advancement through vision-language models
- Novel applications in text-to-image (T2I) generation





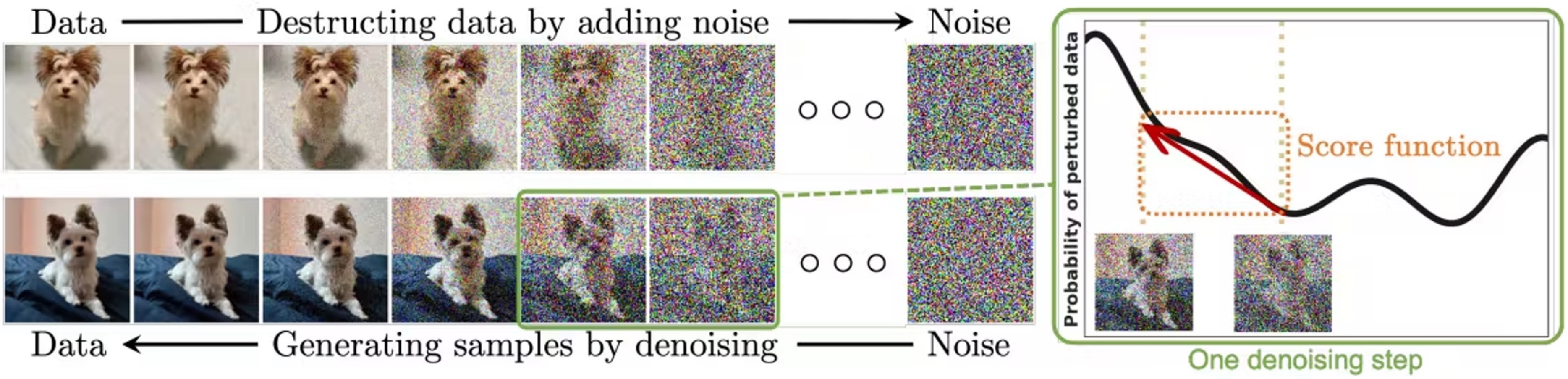
Adversarial Risks in T2I Generation

- New risks in T2I models
- Malicious exploitation to generate harmful content
- Previous works on untargeted attack and targeted erasing
- cheating suffixes are marked in red
- object to be erased is marked in blue

No attack	Untargeted attack	Targeted erasing	Targeted attack (ours)
A snake and a young man	A snake and a young man C63RR	A snake and a young man -08=*	a photo of person wild blers rwby migrant

Background on Diffusion Models

- Diffusion models transform Gaussian distribution into complex data distribution.
- Applications beyond image generation: music, 3D, and video generation.
- Enhancement by CLIP model for T2I generation (pair images and text).



MMP-Attack

- multi-modal priors
  - .e. both text and image features
  - Goal: add a target object into the image content while simultaneously removing the original object.
- Superior universality and transferability.
  - suffix searched under a specific prefix can generalize to other prefixes
  - suffix optimized on open-source diffusion model can deploy on black-box model.
  - DALL-E 3

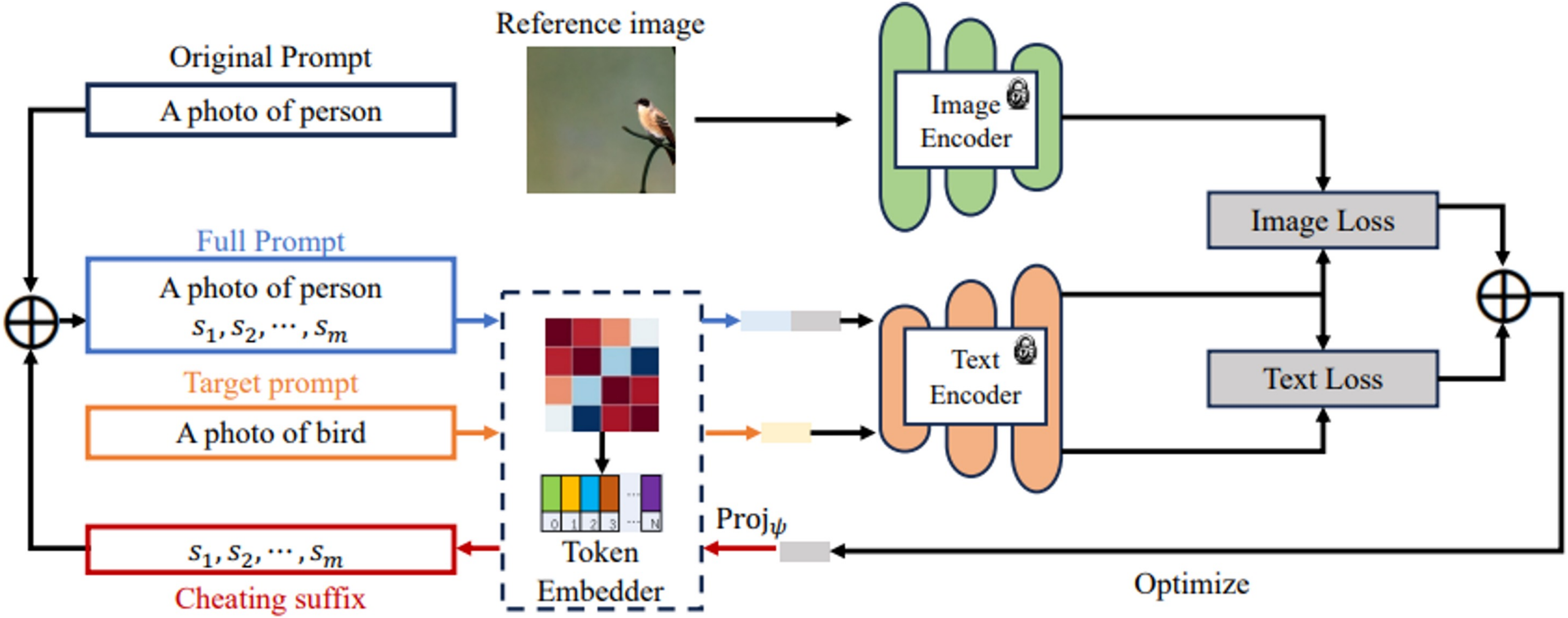


Fig. 2: An illustration of our MMP-Attack pipeline.

T2I Generation Pipeline Explained

- Vocabulary consists of a set of candidate tokens (w1, w2, ..., w|V|) for creating prompts.
- CLIP Model: This has an image encoder (denoted as F^i), that processes images into a vector of a certain size (d_emb)
  - It also includes a token embedder (denoted as E_ψ), and a text encoder (denoted as F^t) that work together to convert the input text prompt into a vector of the same size.
- Training phase: The distance (similarity) between the image and text vectors is minimized for image-text alignment (text-image Match).
- The generative model G uses the textual description (text vector v) to create a new image x.

Problem Formulation for MMP-Attack

- The original prompt s_0 containing n tokens as an element of V^n
- The cheating suffix S_a to be optimized is represented as an element of V^m
- Full prompt is the concatenation denoted as $s_0 \oplus s_a \in V^{n+m}$

$$\operatorname{argmax}_{s_a} \mathbb{E}_{x \sim G(F^t(E_\psi(s_o \oplus s_a)))} \mathcal{A}(x, t, s_o)$$

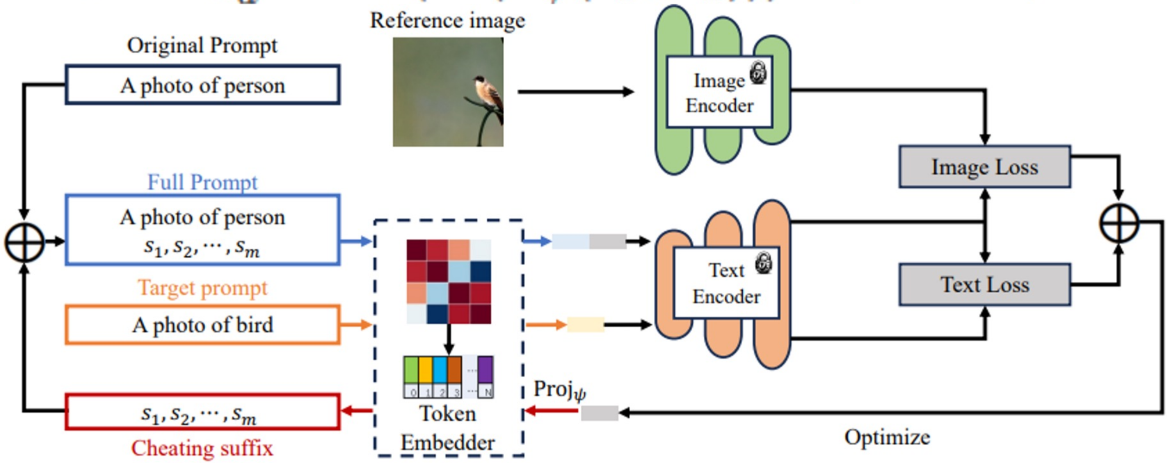


Fig. 2: An illustration of our MMP-Attack pipeline.

Optimization Objective for MMP-Attack

Direct solution of the original optimization problem is infeasible due to the unknown generative model G

Alternative approach: construct a target vector v_t that yields a favorable solution to the optimized alignment function $\mathcal{A}(x, t, s_o)$

$$\operatorname{argmax}_{v_t} \mathbb{E}_{x \sim G(v_t)} \mathcal{A}(x, t, s_o)$$

$$\operatorname{argmax}_{s_a} \cos(F^i(E_\psi(s_o \oplus s_a)), v_t)$$

$$\begin{aligned} &\operatorname{argmax}_{s_a} \cos(v, v_t^{image}) + \lambda \cos(v, v_t^{text}), \\ &\text{s.t.} \quad v = F^i(E_\psi(s_o \oplus s_a)), \end{aligned} \tag{4}$$

Optimization Approach for MMP-Attack

- Optimization problem that is non-differentiable and often NP-hard.

$$\begin{aligned} &\operatorname{argmax}_{s_a} \cos(v, v_t^{image}) + \lambda \cos(v, v_t^{text}), \\ &\text{s.t.} \quad v = F^i(E_\psi(s_o \oplus s_a)), \end{aligned} \tag{4}$$

- Straight-Through Estimation (STE) technique is employed, introducing a differentiable function sg(·) for gradient-based optimization.
- Token embedding matrix Z for the cheating suffix is optimized, using a differentiable projection function $Proj_\psi$
- Gradient-based optimizer is used, outperforming prior zero-order optimization methods.

$$\begin{aligned} &\operatorname{argmax}_Z \cos(v, v_t^{image}) + \lambda \cos(v, v_t^{text}) \\ &\text{s.t.} \quad v = F^i(E_\psi(s_o \oplus s_a)) \\ &\qquad = F^i(E_\psi(s_o \oplus E_\psi^{-1}(\mathrm{Proj}_\psi(Z)))) \\ &\qquad = F^i(E_\psi(s_o) \oplus \mathrm{Proj}_\psi(Z)). \end{aligned} \tag{5}$$

MMP-Attack Algorithm Overview

Initialization: Compute image and text target vectors v_t^{image} and v_t^{text} and initialize the token embedding Matrix Z

Iterative optimization: For N iterations, Update Z by maximizing the combined cosine similarity.

Algorithm 1 MMP-Attack

Input: token embedder E_ψ, dimension of the token embedding vector d_{token}, text encoder F^t, image encoder F^i, learning rate η, number of iterations N, original prompt s_o, number of tokens in cheating suffix m, target category $t \in \mathcal{V}$, weighting factor λ, a reference image x_t containing the target category t and unrelated to original prompt s_o.
Output: Cheating suffix s_a.

$v_t^{image} \leftarrow F^i(x_t)$.
$s' \leftarrow$ ‘`a photo of` t’.
$v_t^{text} = F^t(E_\psi(s'))$
Initialize $Z \in \mathbb{R}^{m \times d_{\text{token}}}$.
$bestloss \leftarrow \infty, bestZ \leftarrow Z$
for $i \leftarrow 1$ to N **do**
  $v \leftarrow F^t(E_\psi(s_o) \oplus \text{Proj}_\psi(Z))$.
  $\mathcal{L} = -\cos(v, v_t^{image}) - \lambda \cos(v, v_t^{text})$.
  if $bestloss > \mathcal{L}$ **then**
    $bestloss \leftarrow \mathcal{L}, bestZ \leftarrow Z$.
  end if
  $Z \leftarrow Z - \eta \nabla_Z \mathcal{L}$.
end for
$s_a \leftarrow E_\psi^{-1}(\text{Proj}_\psi(bestZ))$.

Experimental Setup

- Dataset: 20 category pairs from COCO, with 5 objects: car, dog, person, bird, knife.
- Performance metrics averaged over $5 \times 4 \times 100 = 2000$ images.
- Models: Stable Diffusion v1.4 and v2.1, and DALL-E 3 for evaluation.
- Image generation specs: 512×512 resolution, 50 inference steps, 7.5 guidance scale.
- Adam optimizer for suffix search, 4 tokens, 0.001 learning rate, 10000 iterations.

Implementation and Evaluation Metrics

- Attack implementation: 6 minutes per category pair on an Nvidia RTX 4090 GPU.
- Baseline methods: No attack, Random suffix, Genetic algorithm-based suffix.
- Evaluation metrics
    - CLIP score: matching score based on cosine similarity
    - BLIP score: image-text matching score
    - OCNDR: examine generated image fails detect objects of the original category
    - TCDR: generated image contains objects of the target category
    - BOTH: both OCNDS and TCDS are 1.
- Experimental settings: Grey-box (known CLIP model) and Black-box (unknown CLIP model).

Targeted Attack Results

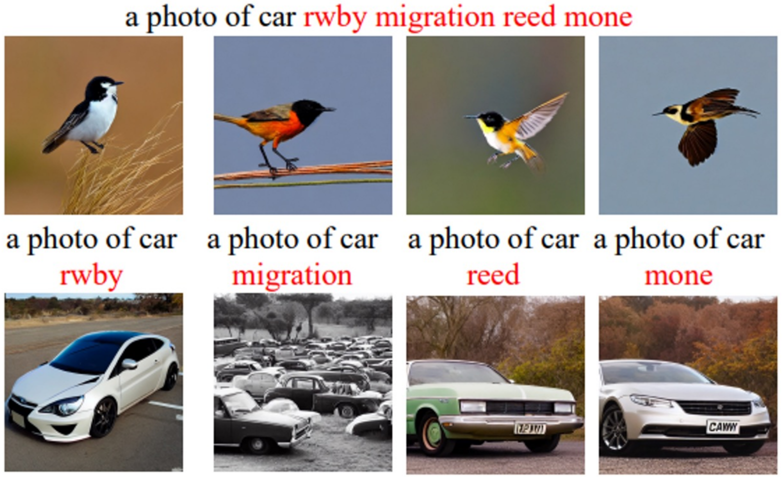
- Baseline comparisons with Stable Diffusion v1.4 (SD v14) and v2.1 (SD v21).
- MMP-Attack significantly outperforms baselines: CLIP score, BLIP score, OCNDR, TCDR, and BOTH.
- MMP-Attack achieves BOTH scores of 81.8% on SD v14 and 86.4% on SD v21, surpassing the strongest baseline by a large margin.

Model	Method	CLIP	BLIP	OCNDR	TCDR	BOTH
SD v14	No Attack	0.204	0.019	5.0%	1.6%	0.1%
	Random	0.203	0.015	5.4%	1.9%	0.6%
	Genetic@5	0.211	0.072	21.4%	20.8%	13.0%
	Genetic@32	0.223	0.066	19.7%	27.4%	14.2%
	MMP-Attack (ours)	**0.265**	**0.414**	**92.0%**	**87.2%**	**81.8%**
SD v21	No Attack	0.204	0.019	5.0%	1.6%	0.1%
	Random	0.203	0.015	5.4%	1.9%	0.6%
	Genetic@5	0.200	0.014	9.2%	5.5%	1.1%
	Genetic@32	0.206	0.021	18.7%	11.1%	5.5%
	MMP-Attack (ours)	**0.270**	**0.429**	**95.2%**	**91.0%**	**86.4%**

Cheating Suffixes and Imperceptible Attacks

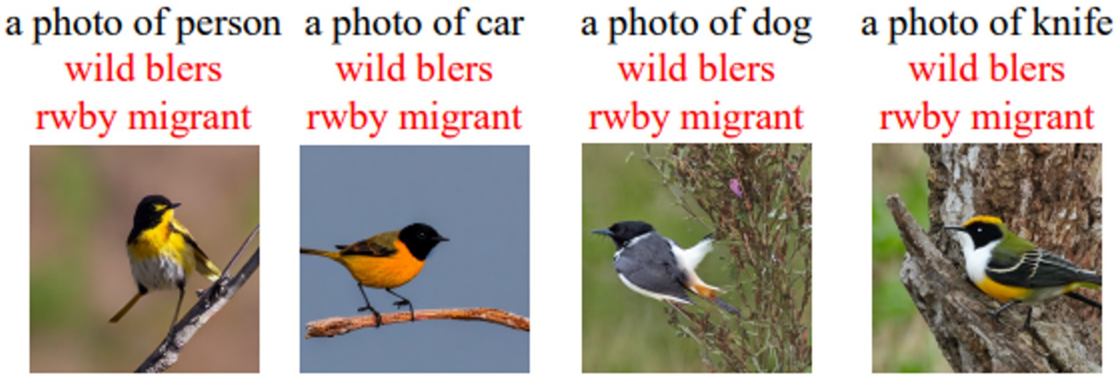
- MMP-Attack identifies relevant tokens for targeted attacks, bypassing simple defenses.
- Specific tokens related to target objects successfully direct the T2I model.
- Subtle manipulation: using a combination of tokens not individually related to the target can still guide the model correctly.

Fig. 3: Examples of optimized cheating suffixes (marked in red) and their corresponding generated images.



Universality of MMP-Attack

- Cheating suffixes exhibit universality, effectively transferring across different original categories.
- The suffix 'wild blers rwby migrant' successful in generating images of birds from various original prompts.
- Evaluation across 20 cheating suffixes shows high universal attack success rates, with some reaching up to 99%.



Transferability of MMP-Attack

- Cheating suffixes demonstrate transferability between different versions of Stable Diffusion models.
- Suffixes optimized for SD v14 can effectively attack SD v21 and vice versa.
- This transferability signifies the potential for black-box targeted attacks using transfer-based strategies.

Black-Box Attack Performance

- Black-box attack settings show a remarkable success rate, with a BOTH score of 50.4% for SD v14 → SD v21 and 66.8% for SD v21 → SD v14.
- Transferability on DALL-E 3 poses a higher challenge due to automatic refinement of input prompts.

Setting	CLIP	BLIP	OCNDR	TCDR	BOTH
SD v14 → SD v21	0.243	0.231	72.3%	62.2%	50.4%
SD v21 → SD v14	0.247	0.235	71.3%	74.9%	66.8%

You
a photo of bird hiatus laureate andre washington

DALL·E

Here are two images of the fictional character Andre Washington, envisioned as a bird hiatus laureate. You can view the distinctive elements of his character in these representations.

You
a photo of knife terriers staffers portrait django

DALL·E

Here are the images of the three dogs as described: a Knife Terrier, a Staffer, and a Django. Each dog is depicted with its unique features.

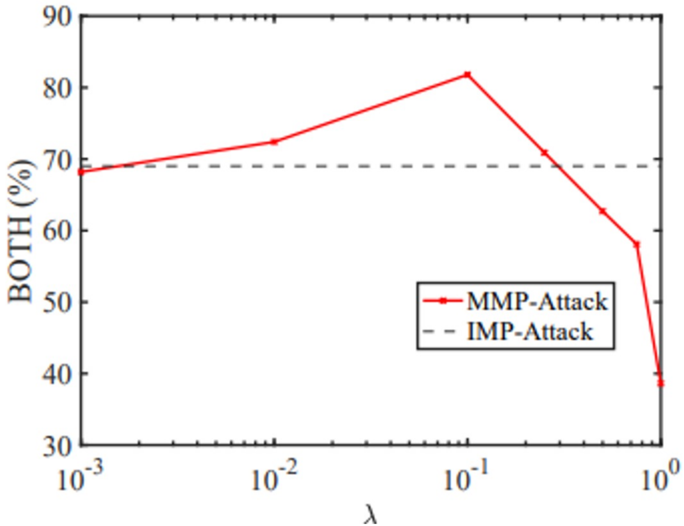
Ablation Study on Initialization Methods

- Examined the impact of initialization methods
  - EOS:Initialize all Zi as the token embedding for [eos], where [eos] is a special token in CLIP vocabulary representing the end of string.
  - Random: Randomly sample m tokens from the filtered vocabulary and use their embeddings as the initial values for Z.
  - Synonym: select token with highest cosine similarity to the target category t in the filtered vocabulary, and use its token embedding as the initial values for all z_i.
- Synonym initialization method yielded the best results, becoming the default choice.

Initialization	CLIP	BLIP	OCNDR	TCDR	BOTH
EOS	0.262	0.390	82.2%	78.3%	72.3%
Random	0.263	0.400	84.1%	82.0%	74.4%
Synonym	**0.265**	**0.414**	**92.0%**	**87.2%**	**81.8%**

Impact of Multi-modal Objectives

- Analyzed the weighting factor λ's effect on attack performance.
- The optimal performance was achieved with λ set to 0.1.
- λ from {0, 0.001, 0.01, 0.1, 0.25, 0.5, 0.75, 1}



Visual Instruction Tuning

Feilian Dai (kdr4qp)

LLaVA (Large Language and Vision Assistant)

- End-to-end trained large multimodal model
- Combining Vision and Language Capabilities: process and understand inputs that include both text and visual elements.
- Open Source and Collaborative Effort: https://llava-vl.github.io/

LLaVA (Large Language and Vision Assistant)

LLaVA (llama 2)

What is unusual about this image? can you explain this to a 5-year-old kid?

In the image, there is a man standing on top of a yellow car, which is unusual because people typically do not stand on top of cars, especially while they are driving. It is also uncommon to see a person ironing clothes on top of a car, as ironing is usually done indoors or in a designated area. This scene is quite unique and might be considered humorous or intriguing to a 5-year-old kid. It is important to explain that this is not a typical situation and that people should not attempt to stand on top of cars or iron clothes in such a manner.

Can you explain this in Chinese?

In the image, there is a man standing on top of a yellow car, which is quite unusual because people usually do not stand on top of cars, especially while they are driving. Additionally, the man is ironing clothes on top of the car, which is not a common activity. This scene is quite unique and might be considered humorous or intriguing to a 5-year-old kid. It is important to explain that this is not a typical situation and that people should not attempt to stand on top of cars or iron clothes in such a manner.

LLaVA (llama 1)

What is unusual about this image? can you explain this to a 5-year-old kid?

In this image, a man is sitting on top of a washing machine, which is attached to the back of a yellow taxi cab. This is unusual because it is not common to see people sitting on top of washing machines, especially while they are being transported on the back of a taxi. It is a funny and unexpected scene that might catch the attention of a 5-year-old kid.

Can you explain this in Chinese?

在这张图片中，一个人正坐在一台洗衣机上，这个洗衣机被固定在一辆黄色出租车的后部。这是不常见的，因为人们通常不会坐在洗衣机上，尤其是在它们被固定在出租车上的情况下。这是一个有趣和意外的场景，可能会吸引一个5岁的孩子的注意。

Context type 1: Captions

A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.

Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation

Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>

Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>

Response type 3: complex reasoning

Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

GPT-assisted Visual Instruction Data Generation

Two types of symbolic representations to encode an image into its visual features to prompt a text-only GPT:

- Captions typically describe the visual scene from various perspectives
- Bounding boxes usually localize the objects in the scene, and each box encodes the object concept and its spatial location

Three types of instruction-following data (human annotations):

- Conversation
- Detailed description
- Complex reasoning

Visual Instruction-tuning Related Work

- Multimodal Instruction-following Agents

    End-to-end trained models, which are separately explored for each specific research topic

    A system that coordinates various models via LangChain / LLMs, such as Visual ChatGPT, X-GPT

- Instruction Tuning

    To enable LLMs to follow natural language instructions and complete real-world tasks

    Applications: Natural Language Understanding (NLU), Content Generation, Decision Making and Predictions

Summary of Contribution

- Extend instruction-tuning to the language-image multimodal space
  - building a general-purpose visual assistant
- Multimodal instruction-following data
  - present a data reformation perspective and pipeline to convert image-text pairs into an appropriate instruction-following format, using ChatGPT/GPT-4
- Large multimodal models
- Multimodal instruction-following benchmark
  - LLaVA-Bench with two challenging benchmarks, with a diverse selection of paired images, instructions and detailed annotations

Visual Instruction Tuning Architecture

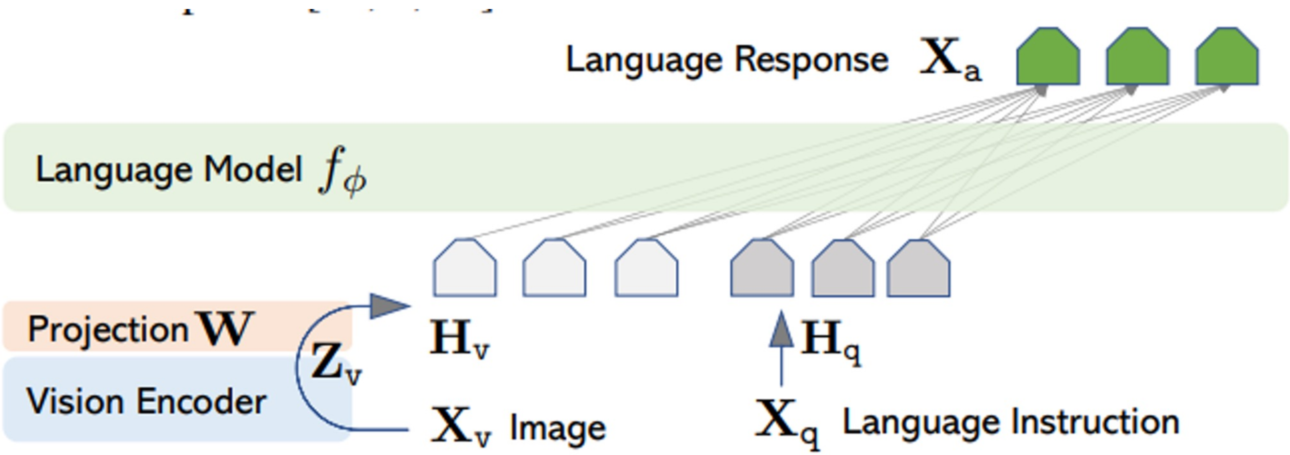


Figure 1: LLaVA network architecture.

$$\mathbf{H}_v = \mathbf{W} \cdot \mathbf{Z}_v, \text{ with } \mathbf{Z}_v = g(\mathbf{X}_v)$$

Architecture

Visual Instruction Tuning Architecture

Hv: language embedding tokens

Xv: Input image

Zv: Visual feature

W: Trainable projection matrix

Xa: Language Response

g: Transformer-based model

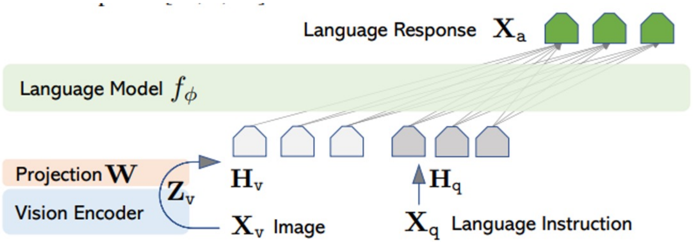


Figure 1: LLaVA network architecture.

$$\mathbf{H}_v = \mathbf{W} \cdot \mathbf{Z}_v, \text{ with } \mathbf{Z}_v = g(\mathbf{X}_v)$$

Training

For each image $\mathbf{X}_{\mathrm{v}}$, we generate multi-turn conversation data $(\mathbf{X}_{\mathrm{q}}^1, \mathbf{X}_{\mathrm{a}}^1, \cdots, \mathbf{X}_{\mathrm{q}}^T, \mathbf{X}_{\mathrm{a}}^T)$, where T is the total number of turns. We organize them as a sequence, by treating all answers as the assistant's response, and the instruction $\mathbf{X}_{\texttt{instruct}}^t$ at the t-th turn as:

$$\mathbf{X}_{\texttt{instruct}}^t = \begin{cases} \text{Randomly choose } [\mathbf{X}_{\mathrm{q}}^1, \mathbf{X}_{\mathrm{v}}] \text{ or } [\mathbf{X}_{\mathrm{v}}, \mathbf{X}_{\mathrm{q}}^1], & \text{the first turn } t = 1 \\ \mathbf{X}_{\mathrm{q}}^t, & \text{the remaining turns } t > 1 \end{cases} \quad (2)$$

For a sequence of length L, we compute the probability of the target answers Xa:

$$p(\mathbf{X}_{\mathrm{a}} | \mathbf{X}_{\mathrm{v}}, \mathbf{X}_{\texttt{instruct}}) = \prod_{i=1}^{L} p_{\boldsymbol{\theta}}(x_i | \mathbf{X}_{\mathrm{v}}, \mathbf{X}_{\texttt{instruct},<i}, \mathbf{X}_{\mathrm{a},<i}),$$

Experiments

Assess the performance of LLaVA in instruction-following and visual reasoning capabilities with two primary experimental settings :

- Multimodal Chatbot
- ScienceQA

Visual input example, Extreme Ironing:

Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

User	What is unusual about this image?
LLaVA	The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.
	[*Start a new conversation, and clear the history*]
User	What's happening in the scene?
LLaVA	The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.
User	What is unusual about this image?
GPT-4 [36]	The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.
User	What is unusual about this image?
BLIP-2	a man is sitting on the back of a yellow cab
User	What is unusual about this image?
OpenFlamingo	The man is drying his clothes on the hood of his car.

Table 3: Example prompt from GPT-4 paper [36] to compare visual reasoning and chat capabilities. Compared to BLIP-2 [28] and OpenFlamingo [5], LLaVA accurately follows the user's instructions, instead of simply describing the scene. LLaVA offers a more comprehensive response than GPT-4. Even when merely asked to describe the image, LLaVA identifies atypical aspects of the image.

Experiments (Multimodal Chatbot)

	Conversation	Detail description	Complex reasoning	**All**
Full data	83.1	75.3	96.5	85.1
Detail + Complex	81.5 (-1.6)	73.3 (-2.0)	90.8 (-5.7)	81.9 (-3.2)
Conv + 5% Detail + 10% Complex	81.0 (-2.1)	68.4 (-7.1)	91.5 (-5.0)	80.5 (-4.4)
Conversation	76.5 (-6.6)	59.8 (-16.2)	84.9 (-12.4)	73.8 (-11.3)
No Instruction Tuning	22.0 (-61.1)	24.0 (-51.3)	18.5 (-78.0)	21.5 (-63.6)

Table 4: Ablation on LLaVA-Bench (COCO) with different training data. We report relative scores *w.r.t.* a text-only GPT-4 model that uses ground truth image captions and bounding boxes as visual input. We prompt GPT-4 with the answers from our model outputs and the answers by GPT-4 (text-only), and let it compare between both responses and give a rating with an explanation.

	Conversation	Detail description	Complex reasoning	All
OpenFlamingo [5]	19.3 ± 0.5	19.0 ± 0.5	19.1 ± 0.7	19.1 ± 0.4
BLIP-2 [28]	54.6 ± 1.4	29.1 ± 1.2	32.9 ± 0.7	38.1 ± 1.0
LLaVA	57.3 ± 1.9	52.5 ± 6.3	81.7 ± 1.8	67.3 ± 2.0
LLaVA†	58.8 ± 0.6	49.2 ± 0.8	81.4 ± 0.3	66.7 ± 0.3

Table 5: Instruction-following capability comparison using relative scores on LLaVA-Bench (In-the-Wild). The results are reported in the format of *mean* $\pm$ *std*. For the first three rows, we report three inference runs. LLaVA performs significantly better than others. † For a given set of LLaVA decoding sequences, we evaluate by querying GPT-4 three times; GPT-4 gives a consistent evaluation.

Experiments (Multimodal Chatbot)

Challenging examples from LLaVA-Bench (In-the-Wild):

	 ICHIRAN Ramen [source]	 Filled fridge [source]
Annotation	A close-up photo of a meal at ICHIRAN. The chashu ramen bowl with a spoon is placed in the center. The ramen is seasoned with chili sauce, chopped scallions, and served with two pieces of chashu. Chopsticks are placed to the right of the bowl, still in their paper wrap, not yet opened. The ramen is also served with nori on the left. On top, from left to right, the following sides are served: a bowl of orange spice (possibly garlic sauce), a plate of smoke-flavored stewed pork with chopped scallions, and a cup of matcha green tea.	An open refrigerator filled with a variety of food items. In the left part of the compartment, towards the front, there is a plastic box of strawberries with a small bag of baby carrots on top. Towards the back, there is a stack of sauce containers. In the middle part of the compartment, towards the front, there is a green plastic box, and there is an unidentified plastic bag placed on it. Towards the back, there is a carton of milk. In the right part of the compartment, towards the front, there is a box of blueberries with three yogurts stacked on top. The large bottle of yogurt is Fage non-fat yogurt, and one of the smaller cups is Fage blueberry yogurt. The brand and flavor of the other smaller cup are unknown. Towards the back, there is a container with an unknown content.
Question 1	What's the name of the restaurant?	What is the brand of the blueberry-flavored yogurt?
Question 2	Describe this photo in detail.	Is there strawberry-flavored yogurt in the fridge?

Table 6: Challenging examples from LLaVA-Bench (In-the-Wild), we provide extremely-detailed annotation for each image for an accurate evaluation. Some questions require the model to extract details from high resolution image and to have a broad knowledge coverage.

Experiments (ScienceQA)

Method	Subject			Context Modality			Grade		Average
	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	
Representative & SoTA methods with numbers reported in the literature									
Human [34]	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
GPT-3.5 [34]	74.64	69.74	76.00	74.44	67.28	77.42	76.80	68.89	73.97
GPT-3.5 w/ CoT [34]	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
LLaMA-Adapter [59]	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
MM-CoT$_{Base}$ [61]	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37	84.91
MM-CoT$_{Large}$ [61]	95.91	82.00	90.82	95.26	88.80	92.89	92.44	90.31	91.68
Results with our own experiment runs									
GPT-4†	84.06	73.45	87.36	81.87	70.75	90.73	84.69	79.10	82.69
LLaVA	90.36	95.95	88.00	89.49	88.00	90.66	90.93	90.90	90.92
LLaVA+GPT-4† (complement)	90.36	95.50	88.55	89.05	87.80	91.08	92.22	88.73	90.97
LLaVA+GPT-4† (judge)	91.56	96.74	91.09	90.62	88.99	93.52	92.73	92.16	**92.53**

Table 7: Accuracy (%) on Science QA dataset. Question categories: NAT = natural science, SOC = social science, LAN = language science, TXT = text context, IMG = image context, NO = no context, G1-6 = grades 1-6, G7-12 = grades 7-12. †Text-only GPT-4, our eval. Our novel model ensembling with the text-only GPT-4 consistently improves the model's performance under all categories, setting the new SoTA performance.

Results

Visual input example, Chicken Nugget Map:

User	Can you explain this meme in detail?
LLaVA	The meme in the image is a creative and humorous take on food, with a focus on chicken nuggets, as the center of the universe. The meme begins with the phrase "Sometimes I just look at pictures of the Earth from space and I marvel at how beautiful it all is..." The main image features a tray of cooked chicken nuggets, with the chicken pieces positioned to resemble the continents and islands on a world map. The meme playfully suggests that the chicken nuggets represent the Earth, and the various locations depicted in the photo are actually chicken nugget versions of different places. The punchline of the meme appears in smaller text at the bottom, reading "I mean, it's not the real Earth, but how beautiful it is all is." This lighthearted and entertaining meme combines food and humor, using the image of chicken nuggets to create a funny and imaginative representation of the world.
User	Can you explain this meme?
GPT-4 [36]	This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets. The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world. The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.
User	Can you explain this meme in detail?
BLIP-2	sometimes i just look at pictures of the earth from space and marvel how beautiful it is
User	Can you explain this meme in detail?
OpenFlamingo	It's a picture of a chicken nugget on the International Space Station.

Table 9: Example prompt comparing LLaVA, GPT-4, BLIP-2, and OpenFlamingo's visual reasoning capabilities in understanding the humor. BLIP-2 and OpenFlamingo fail to follow the user's instructions. LLaVA and GPT-4 both explain the meme and its humor, while GPT-4 produces a more concise answer.

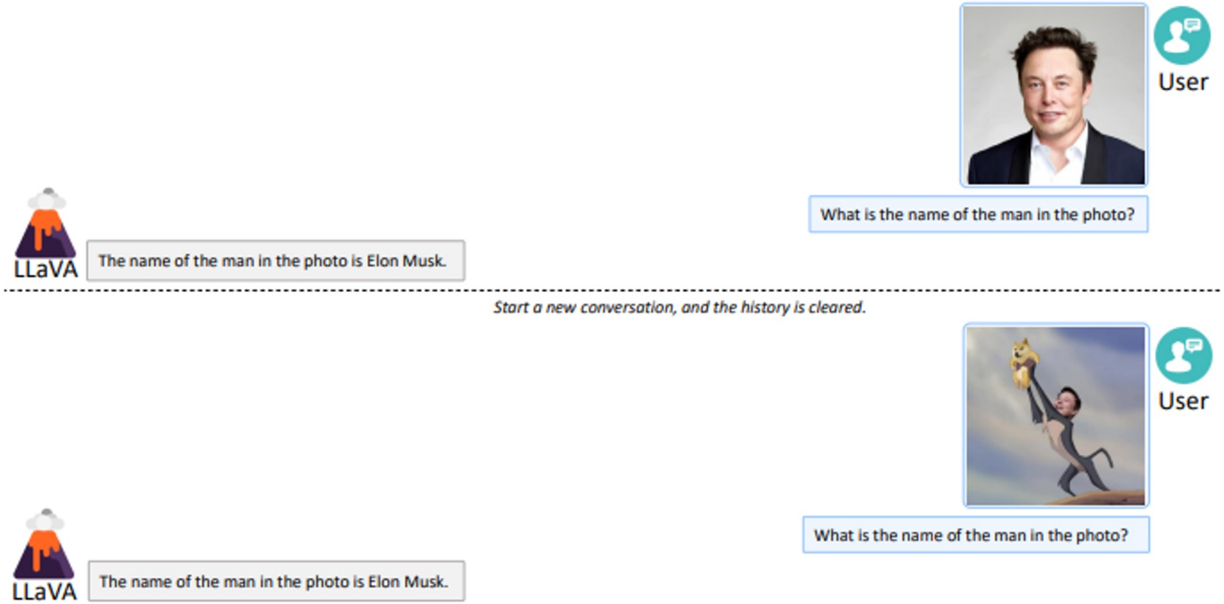


Figure 6: An interesting emergent behavior of LLaVA is its ability to recognize Elon Musk both in a headshot and in a humorous meme where he is dressed as a doge. This implies that the pre-trained CLIP vision encoder may have seen images of Elon Musk. However, it is still surprising because Elon Musk *never* appears in the training data for either the visual feature alignment or visual instruction tuning stages of LLaVA, which indicates that the base language model generalizes to unseen visual concepts.

Results

User

What are the things I should be cautious about when I visit here?

LLaVA

When visiting this peaceful location with a pier extending over a serene lake, surrounded by mountains, there are several things to be cautious about. First, pay attention to the weather conditions, as sudden changes in weather, especially strong winds or storms, can cause the water to become rough and pose risks to the stability of the pier. Second, be mindful of the structural integrity of the pier; avoid overloading it with too many people or excessive weight, as this could potentially damage the pier or cause it to collapse. Additionally, be aware of the surrounding wildlife and respect their space; do not harm or disturb them in any way. Finally, if you decide to go fishing, ensure that you adhere to any local fishing regulations and use appropriate fishing techniques to keep the lake ecosystem healthy and maintain the balance of the food chain.

User

Compose an engaging travel blog post about my trip to this place, highlighting the beautiful sceneries, joyful experiences, and include possible anecdotes.

LLaVA

Traveling to this picturesque location with a serene lake and mountains in the background was a refreshing and rejuvenating experience. The pier extending over the calm water provided an excellent vantage point from which to appreciate the breathtaking scenery. As I stood on the pier, I couldn't help but feel a deep connection with nature, taking a moment to soak in the peaceful atmosphere.

One of the highlights of my trip was undoubtedly the fishing experience. The area is teeming with life, and I was thrilled to land a sizeable catch, despite the modest size of the lake. The joy of fishing in such a beautiful setting was truly rewarding, and it allowed me to take a break from my busy life and truly immerse myself in the tranquility of nature.

While exploring the surrounding area, I came across a family of birds that had built a nest close to the pier, adding to the charm of the location. Interacting with the wildlife and observing their behavior was an enriching experience that I'll cherish for years to come.

In conclusion, this idyllic destination provided me with an opportunity to unwind and recharge, surrounded by mesmerizing natural beauty and the gentle sounds of water and wildlife. If you're looking for a serene escape from the hustle and bustle of everyday life, this pier overlooking the lake is the perfect place to find solace and reconnect with nature.

Figure 3: LLaVA is capable of recognizing the visual content following the user's intent, without directly prompting for visual recognition. It also provides a detailed response when prompted with a follow-up request, and the generated response is closely related to the provided visual content.

Thank you