Appendix

A. Experiments Settings

A.1. Baseline methods

- **TIM:** TIM adopts a translation operation that shifts the benign example by i and i pixels along the two dimensions, respectively. TIM uses a kernel matrix in gradient calculation to replace the translation. In our experiments, we chose the Gaussian kernel as $\tilde{W}_{i,j} = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{i^2+j^2}{2\sigma^2}\right)$ and $W_{i,j} = \frac{\tilde{W}_{i,j}}{\sum_{i,j} \tilde{W}_{i,j}}$.
- **SIM:** The scale-invariant method (SIM) scales every pixel by a set of levels and uses these scaled images for gradient calculation. In our experiments, we choose the number of scale samples $m = 5$ and the scale factor $\gamma_i = 1/2^i$.
- **Admix:** Admix randomly mixes the benign examples with images from other categories and scales the mixed examples in different scales. We set the scale copies $m_1 = 5$ and scale factor $\gamma_i = 1/2^i$ and random sample images $m_2 = 3$ and mixup strength as 0.2.
- **DEM:** DEM provided an ensemble version of diversity invariant methods, which uses five transformed copies for gradient calculation. In our experiments, we set the diversity list to [340, 380, 420, 460, 500].
- **Masked:** Maskblock separates the images into several blocks and sequentially masks every block in the benign examples. Thus, the number of transformed copies is equal to the number of blocks. We set the number of blocks to 16 in our experiments.
- **IDE:** IDE conducts input dropout on a being example at different rates and gets multiple transformed examples to form an ensemble attack. In our experiments, we choose the dropout rate to be 0.0, 0.1, 0.2, 0.3, 0.4, and the weight factor as equal.
- S^2**IM:** S^2IM provides a frequency domain perspective of input transformation, which utilizes DCT and IDCT techniques in transformation. In our experiments, we set the tuning factor ρ = 0.5 and the standard deviation σ the same with perturbation scale ϵ and the number of spectrum transformations $N = 20$.
- **BSR:** BSR splits the input image into several blocks and then randomly shuffles and rotates these blocks. In our experiments, we split the image into $2x2$ blocks with the maximum rotation angle 24% and calculate the gradients on $N = 20$ transformed images.
- **SIA:** SIA decomposed the images into several blocks and transformed each block with an input transformation choosing from seven transformation candidates [2]. We followed the suggested settings in the paper and chose splitting number $s = 3$, number of transformed images for gradient calculation $N = 20$.
- **AutoMA:** AutoMA targeted finding a strong model augmentation policy to boost adversarial transferability. Following the setting in the paper, we trained the augmentation policy search network on 1000 images from ImageNet [36] validation set, which does not overlap with the benign example set. We adopt the transformation number $m = 5$ and set the ten operation types and their corresponding magnitude the same as the original paper.
- **ATTA:** ATTA uses a two-layer network to mimic the transformation function. The benign examples are first passed through this transformation network and then sent for calculating the adversarial perturbations. We use the data from ImageNet [36] training partition to train the transformation network. We trained different transformation networks according to the surrogate models. For the training hyperparameters, we follow the settings from the authors.
- **AITL:** AITL introduces selecting input transformations by different benign examples. AITL trains three networks to predict the input transformations for every image. We adopt the 20 image transformations in the same paper and use the pre-train model weights from the authors to initialize the above networks. We set the number of iterations during optimizing the image transformation feature to 1, the corresponding step size to 15, and the number of image transformation operations to 4.

A.2. Learning to Transform

We decomposed the existing methods and concluded their input transformation methods. We formulate the transformation candidates in 10 categories.

- **(1) Rotate**: Rotate refers to turning the image around a fixed point, usually its center, by a certain angle. The domain of angle is $[0, 360]$. We choose 10 angles from the domain, and the interval between the two angles is identical. Thus, we form 10 operations for the rotate category. The smallest rotation angle is $36°$, and the biggest rotation angle is $360°$.
- **(2) Scale**: the scale category comes from **SIM**. we form 10 operations in our experiments. Each operation differs in scale

[2] Vertical Shift, Horizontal Shift, Vertical Flip, Horizontal Flip, Rotate, Scale, Add noise, Resize, DCT, Dropout

factor $\gamma = 1/2^i, i \in [1, 2, ..., 10]$.

- **(3) Resize**: Resize refers to removing the margin part of examples and resizing the main body of the benign examples. We chose 10 resize rates for our experiments, which are 0,0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9 respectively.
- **(4) Pad**: the pad category comes from **DIM**. We choose to pad the bengin examples to different sizes where the size of the padded example will be $[size \times size]$. We chose 10 different sizes, which are 246.5, 257.6, 268.8, 280.0, 291.2, 302.4, 313.6, 324.8, 336.0, and 347.2.
- **(5) Mask**: The mask category comes from Masked, which separates the examples into several blocks and randomly blocks one of the blocks. We control the number of blocks and choose 4,9,16,25,36,49,64,81,100,121 in specific.
- **(6) Translate**: the translated category comes from **TIM**. We shift the benign examples into 10 levels, which are 10pixel, 20pixel, 30pixel, 40pixel,50pixel, 60pixel, 70pixel, 80pixel, 90pixel, 100pixel, along the x-axis and y-axis.
- **(8) Shuffle**: The shuffle category comes from **BSR**, which separates the examples into several blocks and randomly reorders these blocks. We control the number of blocks and choose 4,9,16,25,36,49,64,81,100,121 in specific.
- **(9) Spectrum**: the spectrum category comes from S^2**IM**, which adds noise in the spectrum domain of benign examples determined by strength ρ. We set ten different ρ as 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0.
- **(10) Mixup**: the mixup category comes from **Admix**. We choose two mixup strengths, 0.2 and 0.4, and five mixup numbers as 1, 2, 3, 4, 5. Thus, we form 10 operations by combining the two settings.

B. Numerical Results

Comparison with advanced methods: We include detailed results of the comparison with different baselines in Tab. 2, Tab. 3, Tab. 4, Tab. 5, Tab. 7, Tab. 6, Tab. 8, Tab. 9, Tab. 10, Tab. 11. For each table, we choose one model from ten models as the surrogate model and use the adversarial examples to attack all these ten models.

We show the attack success rate on adversarial examples crafted on ten different models corresponding to Fig. 5. Tab. 2 is the detailed results for Fig. 5(a). Tab. 3 is the detailed results for Fig. 5(b). Tab. 6 is the detailed results for Fig. 5(c). Tab. 5 is the detailed results for Fig. 5(d). Tab. 7 is the detailed results for Fig. 5(e). Tab. 4 is the detailed results for Fig. 5(f). Tab. 8 is the detailed results for Fig. 5(g). Tab. 9 is the detailed results for Fig. 5(h). Tab. 11 is the detailed results for Fig. 5(i). Tab. 10 is the detailed results for Fig. 5(j). The effectiveness of each attack varies significantly across different models. The L2T attack shows remarkably high effectiveness across all models, which outperforms all the other methods on all ten models.

Evaluation on the defense methods and cloud APIs: We include the detailed results across different defense methods and vision API in Tab. 12 corresponding to Fig. 7. The L2T attack, highlighted in gray, shows exceptionally high success rates across almost all defense methods and APIs, particularly against Bard and GPT-4V.

Ablation study on the number of iterations: We include the detailed results on the different iterations in Tab. 13 corresponding to Fig. 9. For most attacks, success rates increase as the number of iterations increases. This indicates that more iterations generally lead to more effective adversarial examples. After a certain number of iterations (around 20-30 for many attacks), the increase in success rate slows down or plateaus. For example, the L2T attack's success rate increases significantly up to about 30 iterations and then grows more slowly.

Ablation study on the number of samples: We include the detailed results on the different iterations in Tab. 15 corresponding to Fig. 8. This suggests that using more samples to generate adversarial examples can lead to more effective attacks.

Ablation study on the number of operations: We include the detailed results on the different iterations in Tab. 14 corresponding to Fig. 6. As the number of operations increases, there is a general trend of increasing success rates across most models. However, the increase is not significant after the number of operations exceeds 2.

C. Examples on attacking the Multi-modal Large Language Models

To show the scalability of L2T, we also conducted experiments on multi-modal large language models (MLLMs). As shown in Fig. 13Fig. 11, both GPT-4V and Bard can classify the benign example correctly into the "bee-eater". We use L2T to generate the adversarial examples against ResNet-18. As shown in Fig. 14Fig. 12, the Bard classified the adversarial example as a crocodile, and GPT-4V classified it as a dragonfly. It shows the vulnerability of MLLMs, posing great challenges in developing robust MLLMs.

Table 2. Attack success rate (%) across ten models on the adversarial examples crafted on ResNet-18 by different attack

Attack	Res-18	Res-101	NeXt-50	Dense-121	Inc-v3	Inc-v4	ViT	PiT	Visformer	Swin	Average
I-FGSM	100.0	30.3	28.5	36.2	25.9	20.6	7.2	8.9	11.6	16.8	28.6
MI-FGSM	100.0	66.6	71.1	77.7	54.8	50.6	18.6	25.5	35.3	42.7	54.3
Admix	100.0	89.6	90.5	94.6	80.3	77.3	31.8	38.5	56.0	60.4	71.9
BSR	100.0	95.8	96.6	98.1	88.9	90.2	46.1	58.7	77.7	77.6	83.0
DEM	100.0	95.5	95.8	98.1	92.2	90.4	46.9	45.0	67.7	64.3	79.6
DIM	100.0	84.6	87.8	93.6	77.6	73.3	31.1	37.7	53.1	56.8	69.6
SIA	100.0	96.5	97.1	98.6	90.0	89.2	44.4	56.8	74.3	76.0	82.3
IDE	99.9	66.0	68.4	75.5	56.3	51.3	18.8	23.4	34.2	40.9	53.5
Masked	100.0	71.6	76.2	80.5	58.7	54.7	20.1	26.1	37.4	44.4	57.0
SIM	100.0	83.0	85.9	90.7	74.0	69.3	26.2	35.2	48.4	52.4	66.5
S^2IM	100.0	90.4	92.6	94.1	83.8	80.4	32.9	41.6	56.2	62.4	73.4
TIM	100.0	58.7	67.4	72.4	52.1	48.6	18.3	17.4	26.8	34.6	49.6
ATTA	88.0	47.9	50.1	58.3	42.7	35.4	14.0	17.7	24.6	30.7	40.9
AutoMA	100	93.2	95.1	97.4	86.4	87.0	41	50.7	67.7	67.8	78.6
AITL	99.6	93.3	95.2	96.8	91.8	91.2	47.5	51.8	68.9	71.2	80.7
L2T (Ours)	100.0	99.3	99.2	99.6	96.9	97.4	63.7	71.1	86.6	86.0	90.0

Table 3. Attack success rate (%) across ten models on the adversarial examples crafted on ResNet-101 by different attack

Attack	Res-18	Res-101	NeXt-50	Dense-121	Inc-v3	Inc-v4	ViT	PiT	Visformer	Swin	Average
I-FGSM	36.6	100.0	35.4	33.2	25.8	20.6	8.0	10.3	13.0	16.3	29.9
MI-FGSM	72.6	100.0	73.8	71.7	54.1	49.6	22.7	27.2	34.5	38.3	54.4
Admix	94.6	100.0	94.0	94.6	82.9	78.0	38.2	46.9	57.9	60.3	74.7
BSR	97.4	100.0	97.9	97.8	89.2	90.9	56.4	67.4	80.6	81.1	85.9
DEM	97.6	100.0	96.8	97.5	91.7	89.5	52.2	51.9	66.8	68.4	81.2
DIM	86.0	99.9	89.9	89.3	75.1	74.5	38.5	45.6	56.8	57.3	71.3
SIA	98.1	100.0	97.9	98.0	87.8	89.4	48.9	58.9	75.0	74.3	82.8
IDE	78.5	96.4	72.8	73.6	59.9	56.6	23.8	25.6	34.7	43.0	56.5
Masked	80.9	100.0	80.9	80.2	58.8	54.5	25.0	30.4	40.2	43.2	59.4
SIM	86.8	100.0	88.0	89.2	74.9	68.7	33.1	39.1	50.1	51.7	68.2
S^2IM	95.9	100.0	94.8	94.7	88.3	84.3	45.7	51.7	62.3	67.1	78.5
TIM	69.3	100.0	72.8	67.2	50.9	47.8	23.2	23.2	30.7	36.8	52.2
ATTA	51.7	73.1	50.7	49.6	41.2	35.8	15.9	19.8	25.4	27.8	39.9
AutoMA	95.5	99.7	95.4	95.2	85.6	86.1	50.5	59.8	70.3	70.9	80.9
AITL	96.6	99.1	96.5	97.8	92.0	92.5	57.1	64.9	76.0	76.3	84.9
L2T (Ours)	99.3	100.0	99.2	99.5	97.1	96.8	72.3	77.9	88.9	88.1	91.9

Table 4. Attack success rate (%) across ten models on the adversarial examples crafted on DenseNet-121 by different attack

Attack	Res-18	Res-101	NeXt-50	Dense-121	Inc-v3	Inc-v4	ViT	PiT	Visformer	Swin	Average
I-FGSM	44.5	34.0	36.6	100.0	28.6	23.9	8.1	11.3	14.7	20.8	32.2
MI-FGSM	78.6	68.9	74.8	100.0	56.6	53.6	24.5	31.1	44.0	45.6	57.8
Admix	94.3	91.1	93.4	100.0	82.5	81.1	40.8	50.7	68.3	65.8	76.8
BSR	97.4	85.7	97.3	100.0	89.7	91.5	52.2	68.3	84.7	80.0	84.7
DEM	97.8	94.5	97.1	100.0	92.2	91.5	53.8	56.0	74.4	70.8	82.8
DIM	88.4	84.1	89.7	100.0	76.4	75.5	36.5	44.0	62.0	59.5	71.6
SIA	98.4	96.4	97.5	100.0	89.1	92.8	49.7	64.1	83.4	78.1	85.0
IDE	87.8	77.3	80.6	99.4	70.6	68.5	26.3	35.0	49.5	51.8	64.7
Masked	82.8	74.0	81.2	100.0	60.6	60.8	25.7	35.7	49.3	51.3	62.1
SIM	89.7	84.2	88.3	100.0	75.3	74.2	32.6	42.8	59.2	57.3	70.4
S^2IM	97.2	94.9	96.9	100.0	90.7	90.2	50.7	61.6	78.5	76.9	83.8
TIM	74.7	62.4	70.9	100.0	52.2	51.6	20.1	21.7	33.9	38.9	52.6
ATTA	54.8	45.6	49.7	79.4	42.2	36.8	15.3	20.6	28.3	32.3	40.5
AutoMA	95.3	93.8	95.2	99.9	85.4	86.9	46.5	59.6	73.0	71.3	80.7
AITL	97.1	94.3	96.0	99.5	91.3	92.6	53.7	61.5	76.0	74.6	83.7
L2T (Ours)	99.5	98.9	99.3	100.0	97.4	98.3	71.3	79.7	92.9	90.2	92.8

Table 5. Attack success rate (%) across ten models on the adversarial examples crafted on ResNeXt-50 by different attack

Attack	Res-18	Res-101	NeXt-50	Dense-121	Inc-v3	Inc-v4	ViT	PiT	Visformer	Swin	Average
I-FGSM	32.4	29.4	99.4	31.8	25.0	18.5	7.3	9.8	13.1	15.8	28.2
MI-FGSM	64.7	62.9	99.9	69.2	49.3	45.7	19.1	27.0	35.6	38.8	51.2
Admix	88.7	87.4	100.0	94.3	78.0	73.7	33.6	44.0	58.5	57.3	71.5
BSR	95.8	95.7	100.0	97.5	83.3	86.9	47.9	66.8	79.5	74.5	82.8
DEM	96.6	94.8	100.0	97.9	89.5	90.5	49.5	55.1	70.9	67.5	81.2
DIM	81.7	80.7	99.8	85.1	67.7	69.0	33.7	42.4	53.1	54.2	66.7
SIA	97.0	95.1	100.0	97.2	83.5	85.8	44.6	60.6	76.9	73.7	81.4
IDE	76.2	66.1	96.3	71.0	54.8	55.0	20.7	26.8	36.1	42.6	54.6
Masked	74.8	70.6	100.0	76.1	52.5	50.8	22.3	31.2	41.2	43.3	56.3
SIM	79.3	76.9	100.0	86.3	66.2	62.2	25.9	36.6	48.0	47.5	62.9
S^2IM	95.5	94.3	99.9	96.6	86.2	85.3	45.5	56.3	67.3	71.4	79.8
TIM	65.6	58.6	99.8	64.3	45.5	44.2	18.4	20.9	30.1	37.7	48.5
ATTA	43.1	39.8	66.9	42.9	34.3	29.9	14.0	17.5	22.9	25.1	33.6
AutoMA	89.6	91.0	99.7	93.4	78.4	80.8	42.3	57.7	67.7	66.9	76.8
AITL	94.0	92.4	98.9	96.6	88.7	88.9	47.5	59.8	72.5	70.1	80.9
L2T (Ours)	99.4	99.2	100.0	99.3	95.6	97.2	67.2	78.2	88.1	85.8	91.0

Table 6. Attack success rate (%) across ten models on the adversarial examples crafted on Inception-v3 by different attack

Attack	Res-18	Res-101	NeXt-50	Dense-121	Inc-v3	Inc-v4	ViT	PiT	Visformer	Swin	Average
I-FGSM	19.7	13.7	14.6	16.8	98.5	21.9	6.7	7.7	8.8	13.4	22.2
MI-FGSM	48.0	37.5	38.5	42.9	98.7	49.3	16.4	20.7	23.8	29.0	40.5
Admix	66.7	57.6	58.5	67.2	99.8	76.5	23.5	28.8	34.4	41.1	55.4
BSR	88.4	81.9	84.3	88.2	99.8	91.7	39.3	48.4	60.8	64.0	74.7
DEM	77.5	68.7	71.4	75.3	99.5	85.0	34.8	34.1	43.7	50.5	64.0
DIM	59.4	48.2	51.7	57.4	99.0	66.4	21.5	24.3	31.2	37.9	49.7
SIA	82.9	73.0	76.0	81.6	99.3	88.2	31.9	41.4	51.7	55.6	68.2
IDE	56.4	41.9	44.9	46.5	95.4	56.7	15.6	19.1	23.0	29.3	42.9
Masked	55.7	45.8	45.1	50.4	100.0	58.3	17.5	22.7	27.3	32.8	45.6
SIM	60.2	47.7	46.8	54.1	99.8	64.2	19.6	23.7	26.4	33.1	47.6
S^2IM	71.5	64.5	66.1	70.7	99.6	82.7	27.6	36.4	42.1	50.2	61.1
TIM	44.6	31.7	37.6	38.9	98.2	42.3	13.5	13.3	16.2	23.0	35.9
ATTA	31.0	21.0	22.1	23.8	50.9	28	10.4	11.6	13.3	19.2	23.1
AutoMA	65.6	58.0	62.2	65.6	98.5	76.1	27.1	32.6	38.8	44.2	56.7
AITL	77.1	69.9	72.2	79.6	98.9	85.8	34.3	38.9	46.6	53.4	65.7
L2T (Ours)	89.9	86.5	88.1	91.9	99.6	94.8	48.7	54.1	65.4	69.3	78.8

Table 7. Attack success rate (%) across ten models on the adversarial examples crafted on Inception-v4 by different attack

Attack	Res-18	Res-101	NeXt-50	Dense-121	Inc-v3	Inc-v4	ViT	PiT	Visformer	Swin	Average
I-FGSM	22.4	15.0	17.3	18.4	30.5	95.7	6.3	8.6	11.4	13.9	23.9
MI-FGSM	50.1	41.3	43.7	47.6	58.2	97.1	17.4	21.4	28.4	31.5	43.7
Admix	74.9	69.0	71.7	78.6	88.2	99.7	33.3	39.4	50.6	52.8	65.8
BSR	87.3	79.1	85.6	89.3	89.3	99.9	38.5	52.4	66.6	65.2	75.3
DEM	79.0	71.0	76.2	79.4	87.9	99.2	35.6	37.4	52.3	52.8	67.1
DIM	63.0	55.4	60.4	63.8	73.2	96.8	24.7	31.5	39.6	40.8	54.9
SIA	83.0	73.3	78.5	85.5	87.6	99.7	34.1	44.6	59.0	59.8	70.5
IDE	56.8	45.8	48.5	54.9	64.2	92.5	17.4	23.3	28.0	33.6	46.5
Masked	56.0	47.7	49.3	57.3	65.2	99.7	19.9	26.1	33.9	36.5	49.2
SIM	66.3	60.2	64.4	71.1	80.8	99.5	28.9	35.0	44.0	44.6	59.5
S^2IM	76.5	69.9	72.9	77.8	85.4	99.4	33.6	42.4	50.6	54.7	66.3
TIM	46.6	35.8	41.6	44.1	50.8	96.2	13.3	14.8	19.0	24.5	38.7
ATTA	32.6	24.1	25.6	28.4	36.2	46.2	11.3	13.3	17.0	20.0	25.5
AutoMA	71.8	63.8	69.4	75.1	84.1	97.9	32	39.5	50.3	49.8	63.4
AITL	81.1	75.3	79.4	86.1	90.8	99.3	41	47.3	59.5	59.2	71.9
L2T (Ours)	91.5	88.8	91.1	94.5	95.4	99.9	51.7	61.9	75.1	74.0	82.4

Table 8. Attack success rate (%) across ten models on the adversarial examples crafted on ViT by different attacks

Attack	Res-18	Res-101	NeXt-50	Dense-121	Inc-v3	Inc-v4	ViT	PiT	Visformer	Swin	Average
I-FGSM	26.3	19.8	21.7	23.6	23.4	20.6	99.7	20.0	20.6	33.1	30.9
MI-FGSM	52.9	44.7	48.3	51.3	45.6	42.2	99.7	44.6	45.7	60.6	53.6
Admix	64.9	59.8	61.2	64.1	62.1	57.3	99.2	60.6	62.2	74.4	66.6
BSR	83.6	83.8	86.2	87.8	79.9	81.8	99.7	90.3	90.4	89.6	87.3
DEM	76.6	78.5	80.8	81.8	79.6	79.0	99.9	82.1	81.7	81.0	82.1
DIM	63.2	60.7	62.5	65.3	61.1	59.8	98.7	66.5	64.1	71.4	67.3
SIA	82.0	79.9	82.0	83.4	75.2	78.1	99.7	85.4	85.8	88.4	84.0
IDE	67.1	60.8	64.2	66.3	62.5	59.7	99.3	56.8	58.8	72.6	66.8
Masked	55.6	47.5	50.9	54.8	49.3	44.5	99.8	49.2	49.7	65.6	56.7
SIM	60.8	53.0	55.6	60.8	55.1	51.7	99.3	53.7	56.4	68.4	61.5
S^{2}IM	67.8	63.2	65.6	69.4	68.3	65.5	99.9	66.7	67.3	78.3	71.2
TIM	49.1	42.3	46.3	47.1	40.3	37.6	98.9	34.5	37.7	46.5	48.0
ATTA	41.9	33.6	36.1	39.3	39.3	32.9	79.8	32.7	32.6	42.0	41.0
AutoMA	72.1	71.0	73.0	75.8	70.9	71.4	97.9	77.9	77.6	78.6	76.6
AITL	76.8	74.4	77.7	78.6	77.7	75.8	94.9	79.5	78.9	79.6	79.4
L2T (Ours)	89.7	87.3	88.7	89.6	87.4	86.8	98.2	90.6	90.8	92.3	90.1

Table 9. Attack success rate (%) across ten models on the adversarial examples crafted on PiT by different attacks

Attack	Res-18	Res-101	NeXt-50	Dense-121	Inc-v3	Inc-v4	ViT	PiT	Visformer	Swin	Average
I-FGSM	22.1	15.9	18.4	19.9	23.3	17.7	11.3	85.1	21.6	24.8	26.0
MI-FGSM	52.3	41.8	48.3	51.8	46.4	43.0	30.9	97.6	53.1	55.9	52.1
Admix	63.0	55.1	61.8	63.5	57.3	56.8	46.7	97.5	67.5	70.4	64.0
BSR	80.9	77.6	84.0	85.0	74.7	76.8	70.9	99.2	89.5	90.0	82.9
DEM	79.4	74.7	78.5	80.5	78.3	76.9	68.7	99.9	84.9	83.0	80.5
DIM	63.3	58.7	64.6	64.8	61.5	62.4	50.9	94.3	70.1	71.7	66.2
SIA	81.3	77.2	85.6	84.9	75.8	77.3	69.7	99.0	90.6	91.6	83.3
IDE	68.8	61.5	64.0	68.4	66.1	64.0	53.1	94.2	70.2	71.2	68.2
Masked	59.1	51.7	57.2	59.0	53.5	49.1	39.1	99.3	61.8	63.9	59.4
SIM	62.0	54.2	59.9	61.6	55.7	53.6	43.6	99.2	65.1	68.5	62.3
S^{2}IM	71.6	68.9	70.9	73.8	71.7	69.9	61.2	96.4	76.1	78.3	73.9
TIM	48.7	37.9	47.7	47.3	40.7	37.7	27.9	93.8	42.2	48.0	47.2
ATTA	44.4	32.1	38.1	40.3	39.7	35.4	23.7	71.6	37.6	40.2	40.3
AutoMA	71.1	67.9	74.8	76.2	69.8	67.5	62.8	96.6	80.4	81.2	74.8
AITL	79.6	79.0	82.5	83.5	81.2	80.1	74.6	93.5	86.7	86.4	82.7
L2T (Ours)	93.2	90.1	93.0	94.3	90.7	90.7	89.8	99.5	96.9	97.1	93.5

Table 10. Attack success rate (%) across ten models on the adversarial examples crafted on Visformer by different attacks

Attack	Res-18	Res-101	NeXt-50	Dense-121	Inc-v3	Inc-v4	ViT	PiT	Visformer	Swin	Average
I-FGSM	25.4	20.9	24.4	26.6	25.4	21.4	12.0	22.4	93.3	32.6	30.2
MI-FGSM	59.8	50.1	55.3	60.2	50.2	50.8	34.5	54.6	98.3	64.3	57.8
Admix	77.1	70.0	77.4	80.0	69.4	71.0	55.4	77.3	97.8	83.7	75.9
BSR	86.0	82.9	88.8	90.5	79.5	83.7	65.7	90.4	99.5	91.7	85.9
DEM	84.3	81.4	86.6	87.8	83.5	85.1	65.8	83.0	99.9	85.0	84.3
DIM	71.9	68.5	74.9	79.1	69.2	70.5	52.2	75.1	96.8	79.5	73.8
SIA	86.6	84.5	89.9	91.7	80.2	84.2	69.7	90.9	98.9	92.8	86.9
IDE	77.9	71.6	75.8	79.6	73.5	73.8	57.4	73.7	97.0	81.2	76.2
Masked	63.5	54.3	61.4	64.6	54.7	54.6	37.1	60.0	99.2	68.5	61.8
SIM	71.1	65.7	71.2	75.3	64.5	66.5	49.5	71.6	97.8	79.6	71.3
S^2IM	82.1	78.3	81.6	86.1	81.6	82.2	66.4	81.7	97.2	87.3	82.5
TIM	57.4	47.7	56.9	58.9	46.6	47.5	33.9	48.1	97.6	60.0	55.5
ATTA	50.0	39.5	45.7	49.5	41.5	41.8	26.8	42.8	85.9	51.8	47.5
AutoMA	79.3	78.0	85.4	86.7	77.3	80.9	66.8	85.4	98.2	87.8	82.6
AITL	87.2	85.0	88.4	89.3	84.1	87.0	76.6	88.7	96.5	90.5	87.3
L2T (Ours)	96.8	95.6	97.1	97.9	94.4	96.5	89.9	96.6	100.0	97.5	96.2

Table 11. Attack success rate (%) across ten models on the adversarial examples crafted on Swin by different attacks

Attack	Res-18	Res-101	NeXt-50	Dense-121	Inc-v3	Inc-v4	ViT	PiT	Visformer	Swin	Average
I-FGSM	14.3	10.8	9.9	13.2	17.5	11.6	5.9	8.1	10.8	72.3	17.4
MI-FGSM	44.9	32.6	36.6	39.9	37.1	31.7	22.5	32.0	40.1	98.8	41.6
Admix	56.0	41.6	47.2	51.7	45.0	41.6	31.4	43.8	53.7	99.2	51.1
BSR	86.9	79.1	86.3	87.3	76.4	78.6	65.6	88.8	92.0	99.3	84.0
DEM	79.4	75.6	78.3	80.0	76.5	77.2	61.5	79.1	81.4	100.0	78.9
DIM	70.9	64.8	70.4	72.0	66.8	67.3	52.3	73.4	76.4	98.0	71.2
SIA	82.7	74.5	79.3	84.2	70.5	72.1	59.3	82.5	88.7	99.1	79.3
IDE	67.3	54.8	59.1	63.9	61.4	56.8	43.8	54.2	61.9	98.4	62.2
Masked	46.5	33.4	39.7	43.8	39.7	33.2	26.7	35.0	44.8	99.5	44.2
SIM	53.0	38.3	44.6	48.2	42.2	40.4	29.9	39.9	49.5	99.2	48.5
S^2IM	83.4	75.6	80.1	83.9	77.9	79.2	67.8	80.8	85.7	99.1	81.4
TIM	58.7	46.9	58.0	58.9	48.1	46.2	33.5	45.0	51.7	99.0	54.6
ATTA	38.3	28.1	32.1	34.6	34.6	28.2	20.3	28.2	34.9	92.0	37.1
AutoMA	81.9	78.2	83.3	84.5	76.0	78.0	65.7	86.9	89.0	98.7	82.2
AITL	87.8	84.0	89.8	90.9	86.9	88.5	72.0	89.4	90.5	97.1	87.7
L2T (Ours)	94.4	91.9	94.2	95.9	90.7	93.1	85.9	94.5	96.3	99.6	93.6

Table 12. Attack success rate(%) on adversarial examples on ensemble attack across four defense methods and four vision API.

Attack	AT	HGD	NRP	RS	Google	Azure	GPT-4V	Bard
SIM	36.3	83.8	65.7	26.4	77.5	69.8	62.4	79.7
TIM	36.6	63.8	56.0	35.7	55.3	52.6	64.1	71.4
Admix	37.8	91.1	70.8	29.4	73.6	57.1	76.0	83.2
DEM	40.3	88.9	74.9	37.8	76.4	69.3	83.3	91.3
AutoMA	37.9	89.1	66.5	30.0	67.4	61.9	71.4	86.2
IDE	40.9	73.1	68.0	38.0	71.0	64.8	57.1	73.1
ATTA	30.3	49.9	47.8	18.4	49.0	47.9	39.4	75.9
Masked	32.6	72.9	49.6	21.1	57.3	52.7	72.0	84.3
AITL	44.3	91.1	79.9	42.1	79.4	65.2	79.6	90.2
S^2IM	41.1	90.6	80.1	37.0	67.0	65.1	86.2	93.6
BSR	38.7	92.6	63.4	29.7	74.4	55.8	82.5	95.1
SIA	37.6	91.5	63.1	28.9	77.5	69.1	89.6	94.2
L2T (Ours)	47.9	98.5	87.2	46.7	86.5	82.7	96.7	99.9

Table 13. Attack success rate(%) on adversarial examples crafts on ResNet-18 by different iterations.

Iteration	SIM	TIM	Admix	DEM	AutoMA	IDE	ATTA	Masked	AITL	S^2IM	BSR	SIA	L2T(Ours)
1	9.1	12.5	7.9	60.3	8.5	7.3	7.7	9.3	7.7	6.6	8.5	7.4	8.4
2	19.7	20.2	19.2	71.6	22.9	13.1	13.2	20.8	18.7	13.6	25.5	19.3	23.5
3	25.2	24.4	26.2	74.2	31.5	17.1	16.0	24.8	26.9	19.9	35.4	28.7	34.1
4	35.9	29.8	38.1	76.0	45.5	24.0	21.3	33.0	41.8	33.2	51.1	44.0	51.3
5	42.0	33.5	45.4	76.3	53.4	29.1	24.8	37.9	50.6	41.4	59.7	52.9	60.9
6	48.8	37.7	53.3	77.6	61.0	35.3	28.6	43.0	59.0	50.8	68.1	62.4	70.8
7	55.5	41.9	60.4	77.7	67.7	41.0	32.5	48.0	66.8	59.7	74.5	70.2	79.1
8	58.3	44.1	64.2	78.3	71.7	44.4	35.3	50.3	71.8	63.8	77.3	74.1	83.1
9	63.1	47.3	68.9	79.0	75.9	50.2	38.7	54.7	77.8	70.1	81.9	79.4	87.3
10	66.1	49.3	71.5	79.0	78.6	53.7	40.9	57.0	81.0	73.4	83.9	82.9	89.4
20	67.2	50.1	72.0	81.3	78.8	57.9	44.7	57.2	81.3	72.6	83.0	84.3	91.4
30	67.0	50.9	71.6	82.2	79.1	57.6	44.6	56.4	81.5	71.2	82.2	83.7	91.5
40	67.4	51.2	71.6	82.8	79.4	58.6	45.1	55.8	81.4	71.4	83.0	84.1	91.8
50	67.5	51.6	71.9	82.7	80.1	59.2	45.3	56.2	83.2	70.7	83.5	84.4	92.3
60	67.4	51.9	71.6	83.0	80.5	59.8	45.4	56.5	81.1	71.0	84.0	85.5	92.6
70	67.3	52.1	71.9	82.8	81.0	60.2	45.1	56.3	81.6	70.6	83.8	85.7	92.8
80	67.5	51.9	71.9	83.2	80.9	60.3	45.5	56.3	82.8	70.1	84.0	85.7	93.0
90	67.6	51.8	71.6	83.1	81.3	60.7	45.4	56.1	83.7	70.2	83.9	85.4	93.8
100	67.3	51.8	71.3	83.3	81.1	60.8	45.5	55.8	82.9	70.0	84.1	85.7	94.7

Table 14. Attack success rate (%) across ten models on adversarial examples crafted on ResNet-18 by different operation number

Operation Number	Res-18	Res-101	NeXT-50	Denset-121	Inc-v3	Inc-v4	ViT	PiT	Visformer	Swin	Average
1	100.0	96.7	96.9	98.3	90.7	89.9	46.6	56.5	74.6	76.1	82.6
2	100.0	99.3	99.2	99.6	96.9	97.4	63.7	71.1	86.6	86.0	90.0
3	100.0	99.4	99.5	99.6	98.2	98.6	63.2	76.0	89.1	89.5	91.2
4	100.0	99.6	99.6	99.8	98.5	99.4	64.1	77.1	90.1	90.0	91.8
5	100.0	99.6	99.7	99.8	98.6	99.5	64.9	77.8	90.5	90.3	92.0

Table 15. Attack success rate (%) across ten models on adversarial examples generated on Res-18 by different number of samples.

Sample Number	Res-18	Res-101	NeXt-50	Dense-121	Inc-v3	Inc-v4	ViT	PiT	Visformer	Swin	Average
1	100.0	90.6	92.3	95.3	85.5	82.5	38.9	46.4	61.0	64.9	75.7
2	100.0	95.4	95.7	98.0	91.3	90.0	47.9	55.9	72.7	74.1	82.1
3	100.0	96.7	97.1	98.6	93.1	93.4	51.6	59.4	78.6	77.7	84.6
4	100.0	97.3	98.3	98.9	94.4	94.0	55.3	62.7	79.0	80.7	86.1
5	100.0	98.3	98.3	99.4	95.4	95.1	57.4	65.7	82.6	83.1	87.5
6	100.0	99.1	98.7	99.6	96.0	96.5	59.3	67.2	83.1	82.2	88.2
7	100.0	99.3	98.4	99.6	96.1	96.3	61.2	67.9	85.0	83.5	88.7
8	100.0	99.1	98.9	99.6	97.2	96.0	59.5	68.9	84.4	85.1	88.9
9	100.0	99.2	99.2	99.5	97.0	96.4	62.3	70.5	86.3	86.3	89.7
10	100.0	99.3	99.2	99.6	96.9	97.4	63.7	71.1	86.6	86.0	90.0
11	100.0	99.2	99.0	99.7	96.5	97.2	64.7	72.7	87.1	86.5	90.3
12	100.0	99.1	98.8	99.8	96.7	96.6	63.8	72.7	86.6	86.0	90.0
13	100.0	99.3	99.0	99.7	96.0	97.5	65.4	72.1	87.6	86.7	90.3
14	100.0	99.4	99.4	99.6	96.9	97.2	65.4	73.8	88.5	89.2	90.9
15	100.0	99.2	99.5	99.6	97.3	97.5	65.4	73.0	88.1	86.8	90.6
16	100.0	99.3	99.4	99.7	97.4	97.6	67.2	74.7	88.6	87.8	91.2
17	100.0	99.4	99.3	99.7	97.9	98.1	66.4	73.0	89.1	87.9	91.1
18	100.0	99.2	99.3	99.5	97.2	97.3	66.7	74.5	89.3	88.1	91.0
19	100.0	99.3	99.2	99.6	97.4	97.9	66.1	73.9	88.4	87.9	91.1
20	100.0	99.3	99.6	99.7	96.6	97.5	66.4	74.2	88.8	89.3	91.1
21	100.0	99.4	99.4	99.5	97.0	98.2	66.1	75.0	89.0	87.8	91.1
22	100.0	99.3	99.6	99.7	97.0	97.8	67.8	75.0	89.3	88.8	91.4
23	100.0	99.4	99.3	99.6	97.0	98.0	68.3	74.2	89.6	88.9	91.4
24	100.0	99.5	99.4	99.7	97.6	97.9	67.4	75.4	89.6	89.7	91.6
25	100.0	99.3	99.5	99.5	97.4	98.1	67.3	75.1	88.8	88.4	91.3
26	100.0	99.3	99.4	99.6	97.3	98.5	68.1	76.1	89.6	88.9	91.7
27	100.0	99.4	99.4	99.8	97.6	97.7	67.7	76.3	90.0	89.7	91.8
28	100.0	99.3	99.2	99.8	97.6	98.0	68.4	76.8	90.3	89.6	91.9
29	100.0	99.3	99.4	99.6	97.5	98.4	67.8	75.5	89.5	89.8	91.7
30	100.0	99.4	99.6	99.6	97.6	98.4	68.3	76.1	90.3	88.7	91.8
31	100.0	99.5	99.5	99.6	97.5	98.4	68.2	76.2	89.7	90.4	91.9
32	100.0	99.5	99.5	99.5	98.0	98.4	68.6	75.9	90.2	89.5	91.9
33	100.0	99.3	99.5	99.7	97.6	98.4	68.0	76.6	90.2	90.1	91.9
34	100.0	99.5	99.5	99.8	97.9	98.2	69.3	76.7	90.4	90.2	92.2
35	100.0	99.5	99.4	99.8	98.0	98.8	69.9	76.6	90.3	90.2	92.2
36	100.0	99.4	99.6	99.8	97.7	98.2	70.1	76.9	90.0	90.1	92.2
37	100.0	99.6	99.6	99.8	97.6	98.2	68.8	76.9	90.6	90.6	92.2
38	100.0	99.4	99.5	99.8	97.6	98.3	69.5	76.0	91.3	89.8	92.1
39	100.0	99.4	99.4	99.5	97.3	98.1	70.5	77.8	90.6	90.2	92.3
40	100.0	99.3	99.6	99.8	97.9	98.6	67.7	76.1	90.4	90.0	91.9
41	100.0	99.5	99.6	99.7	97.6	98.5	69.0	77.4	90.4	90.8	92.2
42	100.0	99.5	99.6	99.8	97.6	98.4	69.7	76.5	90.7	90.2	92.2
43	100.0	99.5	99.3	99.7	98.0	98.8	70.1	77.2	91.3	89.7	92.4
44	100.0	99.5	99.6	99.8	98.2	98.3	69.5	76.6	90.3	89.8	92.2
45	100.0	99.6	99.6	99.8	97.7	98.4	69.7	77.2	90.6	90.4	92.3
46	100.0	99.5	99.7	99.8	97.7	98.5	69.6	77.1	91.6	90.4	92.4
47	100.0	99.7	99.8	99.8	97.9	98.9	69.9	77.0	91.4	90.9	92.5
48	100.0	99.5	99.5	99.7	97.6	98.4	69.5	76.9	90.9	91.3	92.3
49	100.0	99.6	99.6	99.8	97.8	98.7	69.9	76.9	91.3	90.8	92.2
50	100.0	99.5	99.5	99.8	98.2	98.6	69.7	77.4	91.5	91.4	92.6

You

What kind of thing is in the picture?

ChatGPT

The image shows a bird, specifically a kingfisher, in flight as it takes off from a branch. The bird is captured in a moment that beautifully displays its colorful plumage, with wings outstretched and tail feathers spread, which is characteristic of the agility and speed these birds possess, especially when darting out to catch their prey in the water.

Figure 11. The conversation with ChatGPT for the benign example

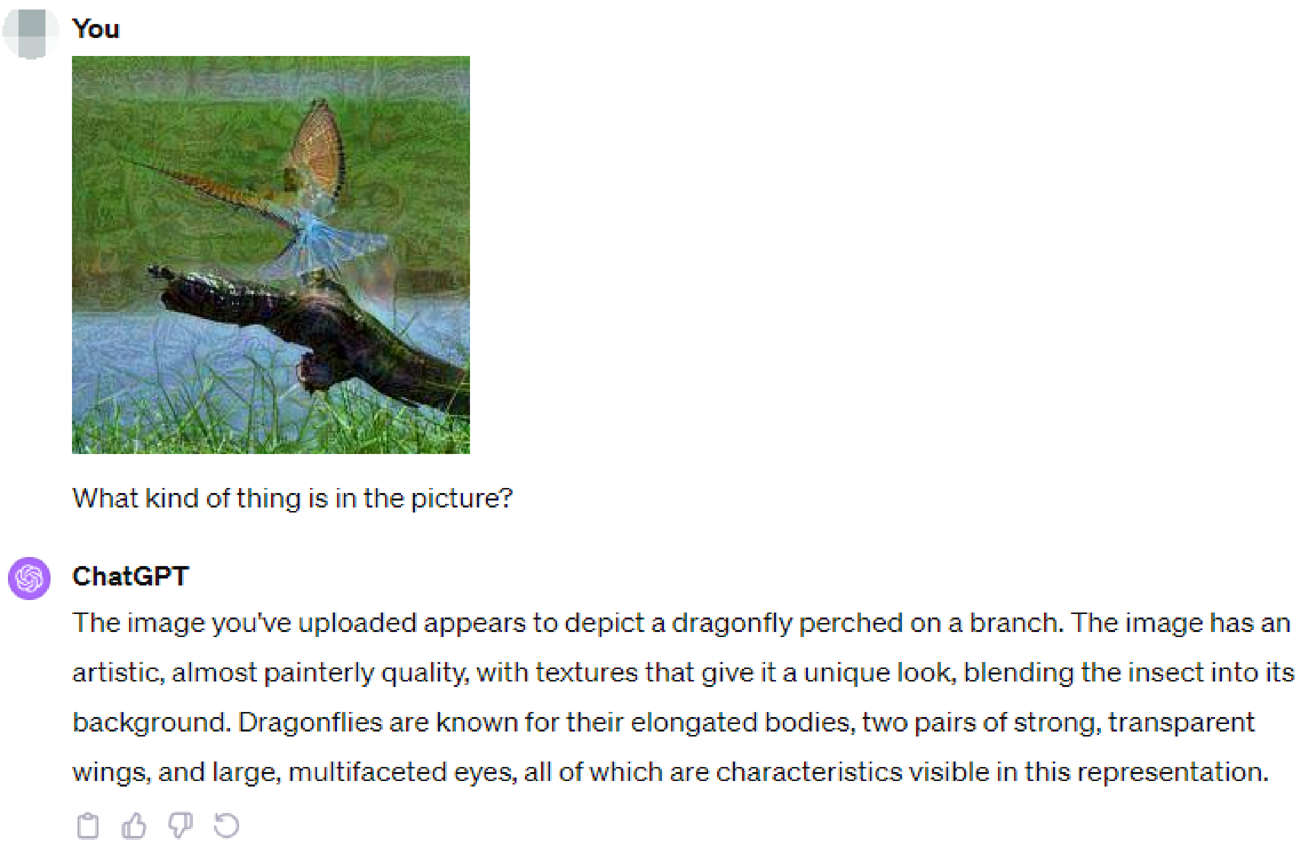

Figure 12. The conversation with ChatGPT for the adversarial example

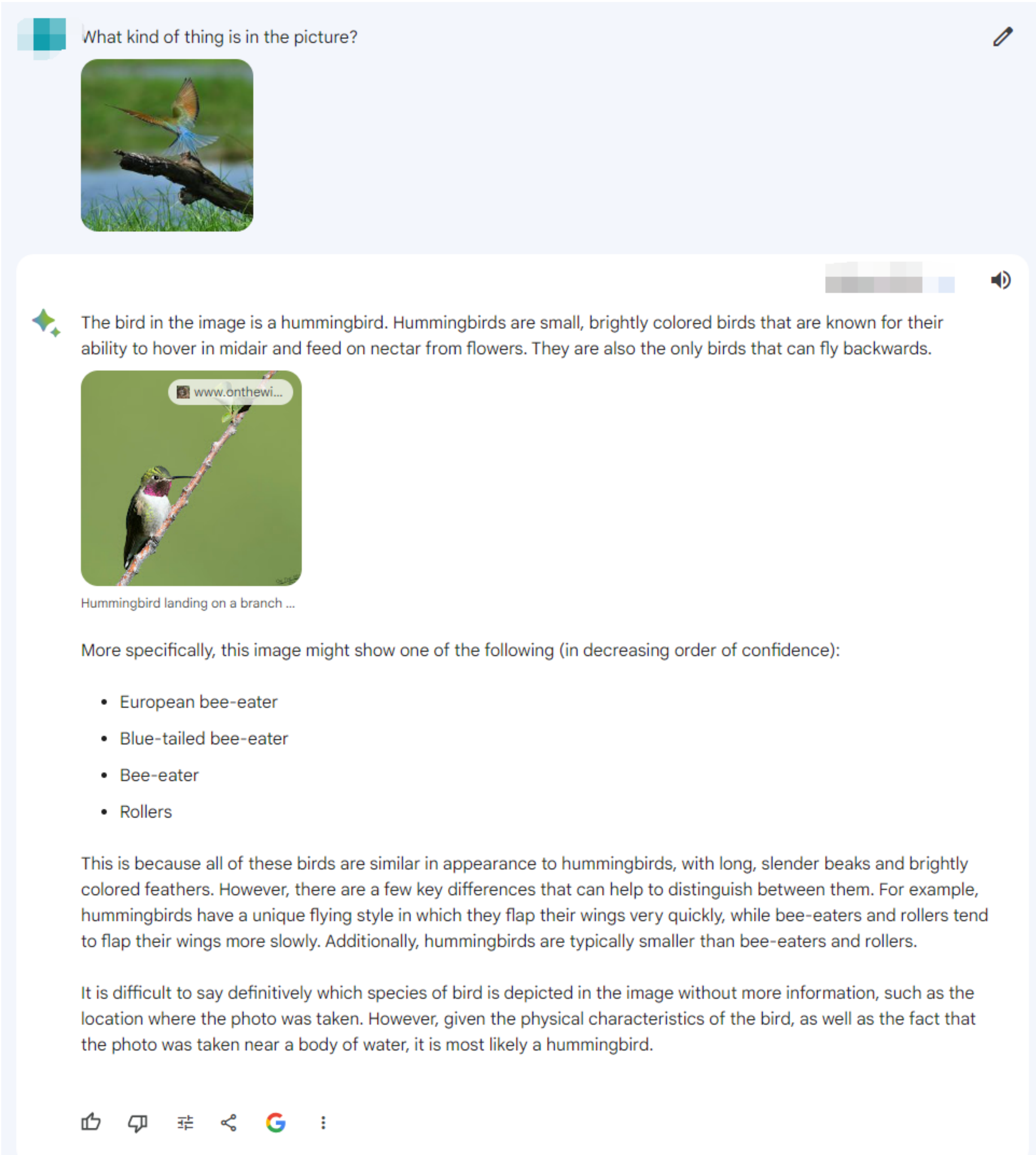

Figure 13. The conversation with Bard for the benign example

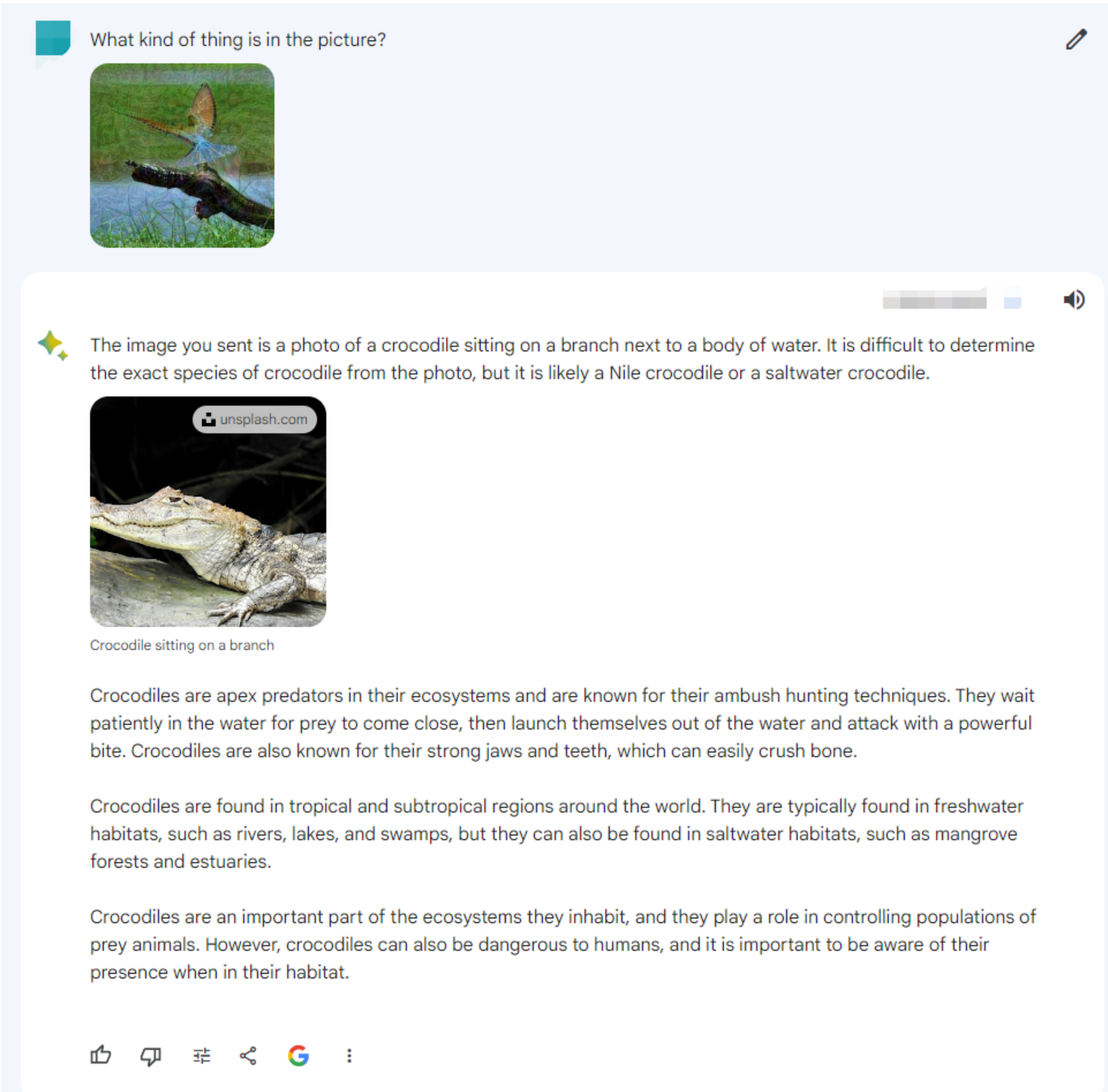

Figure 14. The conversation with Bard for the adversarial example