南京航空航天大學
Nanjing University of Aeronautics and Astronautics

LTGC: Long-tail Recognition via Leveraging LLMs-driven Generated Content

Qihao Zhao[1,2,*], Yalun Dai[3,*], Hao Li[4], Wei Hu[1], Fan Zhang[1,†], Jun Liu[2]
[1]Beijing University of Chemical Technology, China
[2]Singapore University of Technology and Design, Singapore
[3]Nanyang Technological University, Singapore
[4]Northwestern Polytechnical University, China

CVPR 2024

Introduction

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

re-sampling

loss re-weighting

ensemble learning

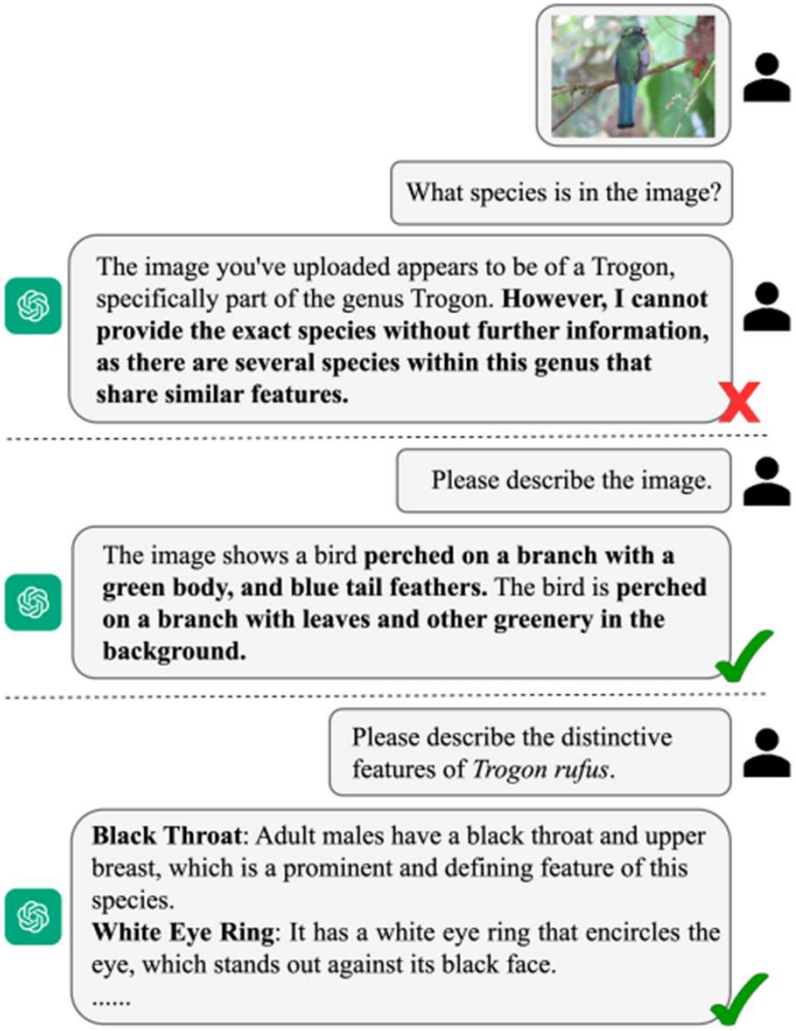
decoupling

contrastive learning

data augmentation

transfer learning methods based on CLIP

LLMs and LMMs

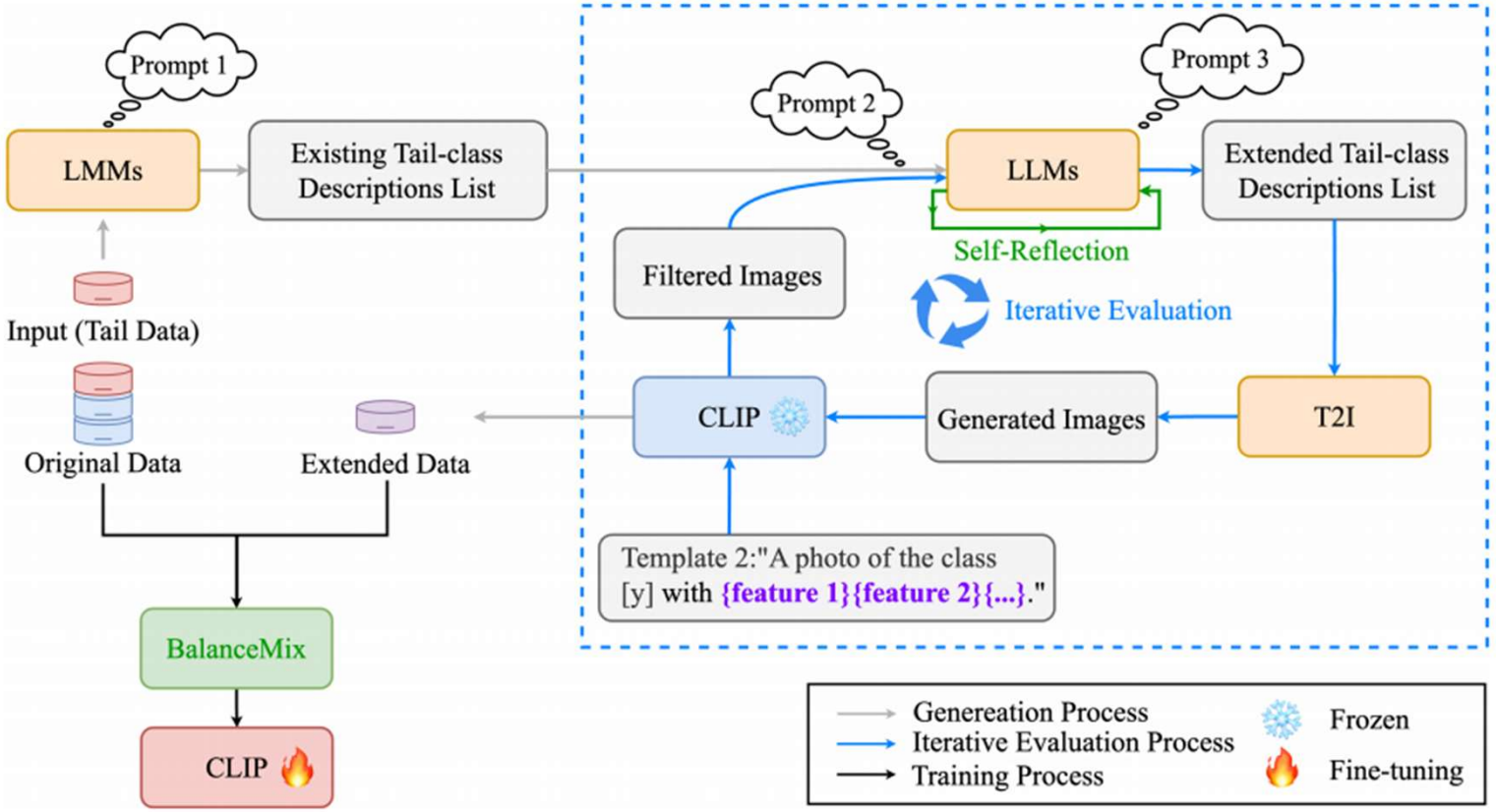


Due to the bias of imbalanced training data,LMMs perform poorly on some complex tasks.

Method

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Overall framework of LTGC



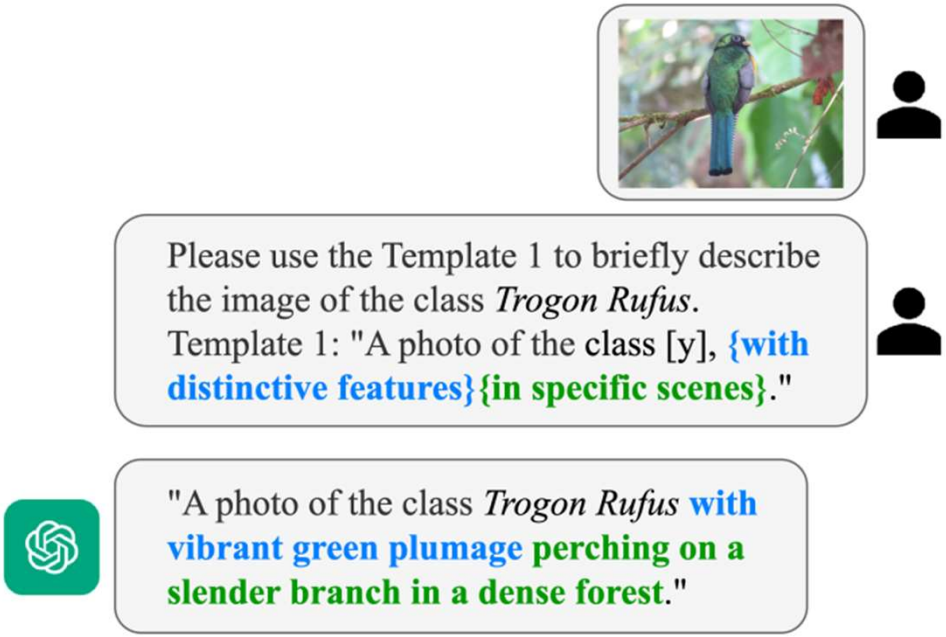
Method——Diverse Tail Images Generation

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Obtaining Existing Tail-class Descriptions List

Template 1： "A photo of the class [y], {with distinctive features}{in specific scenes}."

[Prompt 1]： "Please use the **Template 1** to briefly describe the image of the class [y]."

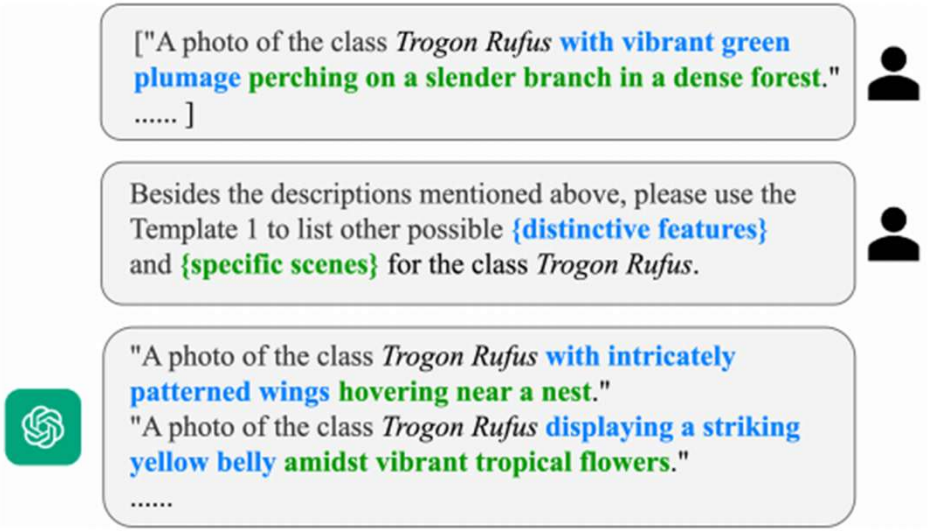


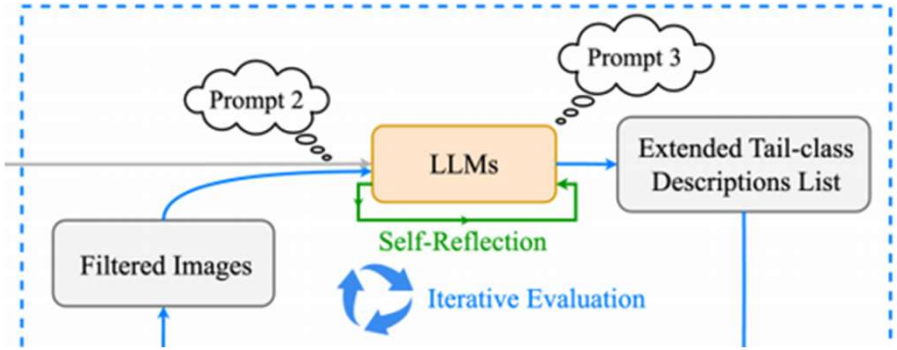
Method——Diverse Tail Images Generation

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Obtaining Extended Tail-class Descriptions List

1) Inputting the existing descriptions list into LLMs
2) Designing the **Prompt 2** to guide LLMs in generating the desired descriptions for images that are absent in the given tail class y: "Besides these descriptions mentioned above, please use the Template 1 to list other possible {distinctive features} and {specific scenes} for the class [y]."





self-reflection module:
(1) number-checking module: re-ask LLMs the **[Prompt 2]** question until a maximum number Ky of the tail class is achieved
(2) repetition-checking module: input the extended descriptions list and the following **[Prompt 3]** of each class y for LLMs' repetition checking

$$K_y = M_y + N_y$$

[Prompt 3]: "Please exclude any repetitive {distinctive features} and {specific scenes} for class [y] in this descriptions list."

Method——Diverse Tail Images Generation

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

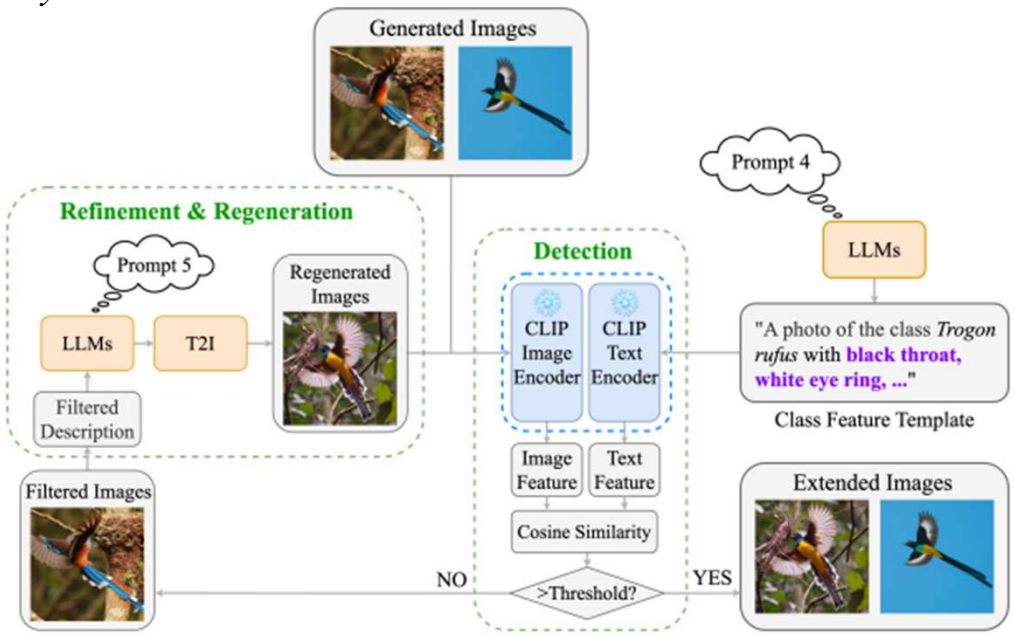
Transform Descriptions to Images

Employ T2I to generate images based on the descriptions list Ly:

$$i_n^y = \mathrm{T2I}(d_n^y), \text{where } n \in \{1, \ldots, N\}$$

lower-quality images

iterative evaluation module



(1) Detection：identify lower-quality images

$$\mathcal{S} = \mathrm{Encoder}_{\mathrm{vis}}(i_n^y) \cdot \mathrm{Encoder}_{\mathrm{text}}(C_y),$$

(2) Refinement：prompt LLMs to refine its corresponding description d_n^y

(3) Regeneration： regenerated the image i_n^y by the T2I model according to the improved textual description

[Prompt 4]: "Please use **Template 2** to summarize the most distinctive features of class [y]}."

Template 2: "A photo of the class [y] with {feature 1}{feature 2}{...}."

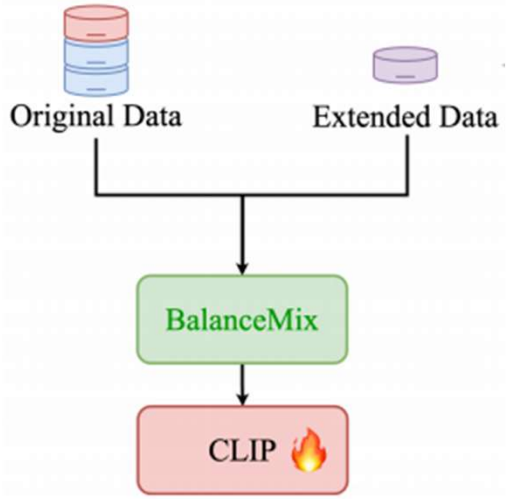
[Prompt 5]:"This description d_n^y doesn't seem to be representative of the class [y].Could you refine it to enhance the distinctive features of class [y]?"

Method—— BalanceMix

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

How to efficiently use these generated data and original data to perform long-tailed recognition well?

BalanceMix



First define the original data and generated data as D_o and D_g

Then BalanceMix balance-sample an image x_i from D_o and sample an image x_j from D_g

Meanwhile, it mixes the images x_i and x_j and their corresponding labels:

$$\tilde{x} = \lambda \odot x_i + (\mathbf{1} - \lambda) \odot x_j,$$
$$\tilde{y} = \lambda \odot y_i + (\mathbf{1} - \lambda) \odot y_j,$$

Fine-tune the CLIP's vision encoder with **LORA** on all mixed data pairs $(\tilde{x}, \tilde{y})$ for efficient long-tail recognition

Experiments

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Evaluation Setup:

Many-shot(more than 100 images)
Medium-shot (20 to 100 images),
Few-shot (less than 20 images)

Method Implementation:

For LMM: GPT-4V(ision) version of ChatGPT.
For LLM: GPT-4 version of ChatGPT.
For T2I: DAll-E .
For the pre-trained CLIP: use ViT-B/32 for its visual encoder and the transformer architecture for its text encoder.

In LLM's self-reflection module, we set the maximum number Ky to 100, 300, and 800 for iNaturalist 2018, ImageNet-LT, and Place-LT, respectively

In the iterative evaluation module, the threshold μ is set at 0.8 for ImageNet-LT and Place-LT, and at 0.6 for iNaturalist.

Experiments

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Table 1. Comparison with SOTA methods on ImageNet-LT and Places-LT.

Dataset	ImageNet-LT		Places-LT	
	Few	All	Few	All
CLIP Zero [36]	58.6	59.8	40.1	38.0
CLIP Finetune [36]	34.5	60.5	22.7	39.7
VL-LTR [36]	59.3	77.2	42.0	50.1
RAC [28]	-	-	41.8	47.2
LPT [14]	-	-	46.9	50.1
LTGC(Ours)	**70.5**	**80.6**	**52.1**	**54.1**

Table 2. Comparison with SOTA methods on iNaturalist 2018.

Method	Many	Medium	Few	All
Softmax	74.7	66.3	60.0	64.7
LADE [16]	64.4	47.7	34.3	52.3
RIDE [44]	71.5	70.0	71.6	71.8
PaCo [11]	69.5	73.4	73.0	73.0
MDCS [57]	76.5	75.5	75.2	75.6
CLIP Zero [36]	6.1	3.3	2.9	3.4
CLIP Finetune [36]	76.6	74.1	70.2	72.6
VL-LTR [36]	-	-	-	76.8
RAC [28]	75.9	80.5	81.0	80.2
LPT [14]	-	-	79.3	76.1
LTGC(Ours)	**77.5**	**83.9**	**82.6**	**82.5**

Table 3. Comparison with different LMMs' methods on ImageNet-LT and iNaturalist2018.

Method	ImageNet-LT	iNaturalist 2018
LENS [3]	69.5	17.4
MiniGPT4 [60]	60.4	20.9
MiniGPT4-v2 [7]	68.5	27.1
GPT-4	72.1	64.3
Ours	**80.6**	**82.5**

Table 4. Evaluation on the effectiveness of the iterative evaluation.

Method	ImageNet-LT	iNaturalist 2018
w/o iterative evaluation	55.8	64.9
Detection and exclusion	71.5	77.4
Ours	80.6	82.5

Table 5. Evaluation on the effectiveness of the BalanceMix.

Method	ImageNet-LT	iNaturalist
w/o BalanceMix	58.3	69.5
Balanced sample [47]	63.9	73.8
Mixup [54]	73.4	75.2
Ours	80.6	82.5

Experiments

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

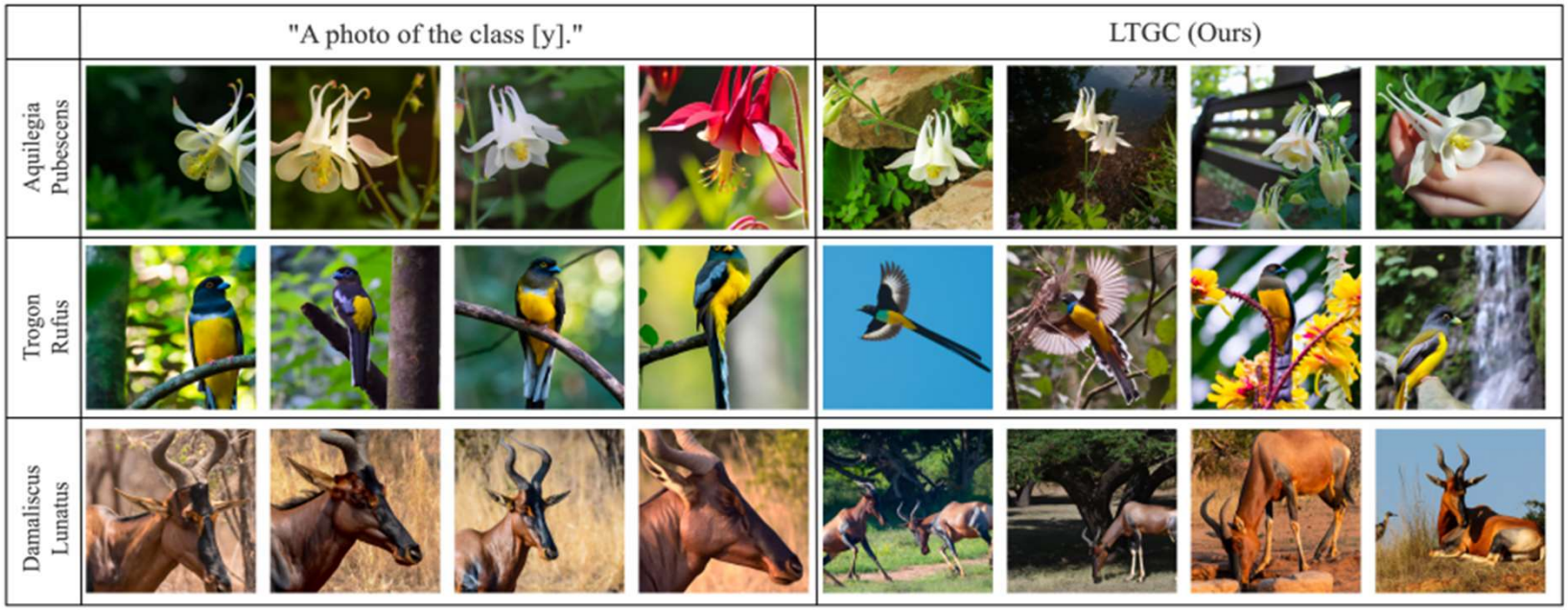


Figure 6. **The visualization of generated images: The template "A photo of the class [y]" and LTGC.** Each row represents a different class. The four images on the left are generated using the simple template "A photo of the class [y]," which results in images with uniform poses and plain backgrounds. The four images on the right are from the proposed LTGC and demonstrate the diversity of classes.

Experiments

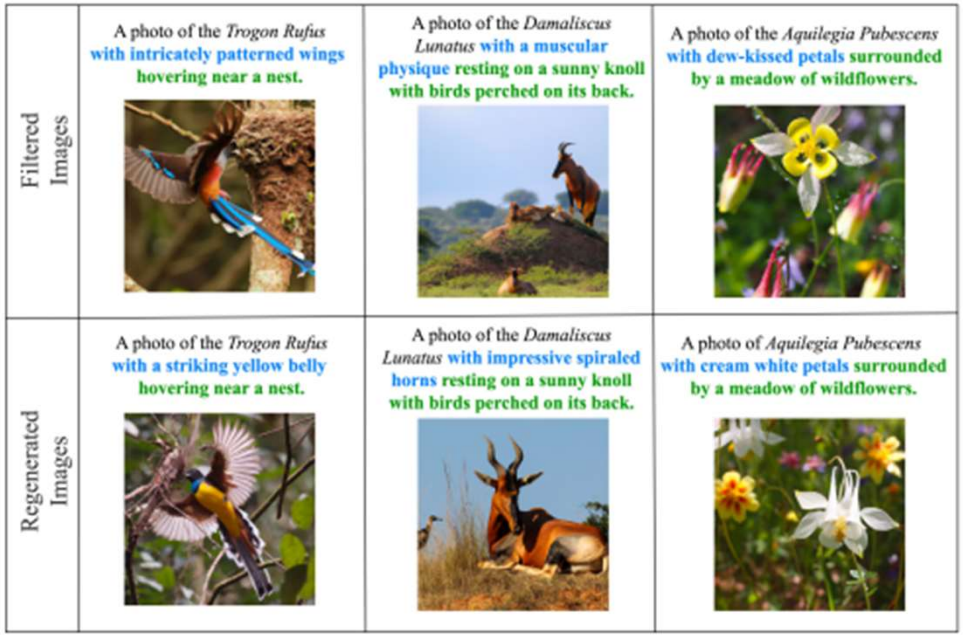


Figure 7. **The visualization of the images generated before and after passing the iterative evaluation module.** The top row displays images that were filtered out, while the bottom row shows images regenerated by T2I after refining their corresponding descriptions. More visualizations are in Appendix.

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Thanks