



AAAI-26 / IAAI-26 / EAAI-26
JANUARY 20-27, 2026 | SINGAPORE

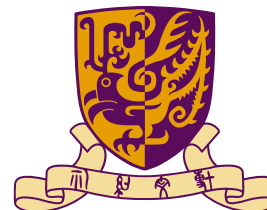
Robust-R1: Degradation-Aware Reasoning for Robust Visual Understanding

Jiaqi Tang*, Jianmin Chen*, Wei Wei†, Xiaogang Xu, Runtao Liu,
Xiangyu Wu, Qipeng Xie, Jiafei Wu, Lei Zhang, Qifeng Chen†

Oral Presentation



香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY



Background: Visual Degradation **Hurts** MLLMs

Question

What is next to each pizza pie?

Original
Image



Background: Visual Degradation **Hurts** MLLMs

Question

What is next to each pizza pie?

Original
Image



Encoder

Visual



Background: Visual Degradation **Hurts** MLLMs

Question

What is next to each pizza pie?

Original
Image



Encoder

Visual



Multimodal
Large
Language
Model

knife

Correct
Answer

Background: Visual Degradation Hurts MLLMs

Question: What is next to each pizza pie?



Original Image



Degradation



Degraded Image

Visual Encoder



Multimodal Large Language Model

knife

Correct Answer

Background: Visual Degradation **Hurts** MLLMs

Question: What is next to each pizza pie?

Original Image



Degradation

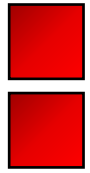
Degraded Image



Visual Encoder



Visual Encoder



Multimodal Large Language Model

knife

Correct Answer

Spoon?

Wrong Answer

Previous Solution: **Implicit** Adaptation Limits

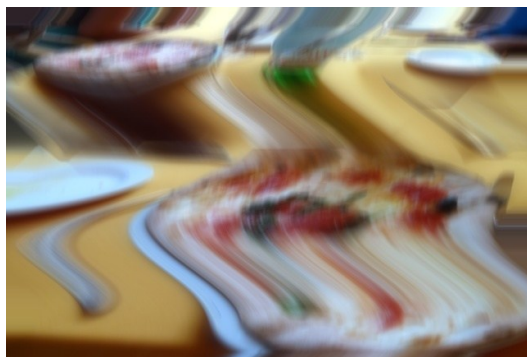
Question

What is next to each pizza pie?

Original
Image



Degradation



Degraded
Image

Visual
Encoder



Aligned



Visual
Encoder



Multimodal
Large
Language
Model

knife

Correct
Answer

Spoon?

Wrong
Answer

Previous Solution: **Implicit** Adaptation Limits

Question

What is next to each pizza pie?

Problem 1:
Limited
Interpretability

Original
Image



Visual
Encoder



Aligned



Visual
Encoder



Multimodal
Large
Language
Model



knife

Correct
Answer



Spoon?

Wrong
Answer

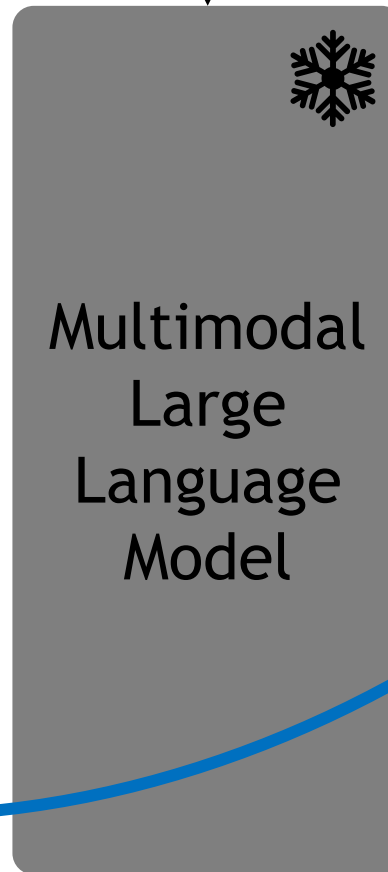
Previous Solution: **Implicit** Adaptation Limits

Question

What is next to each pizza pie?

Problem 2:
Isolated
Optimization

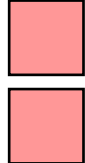
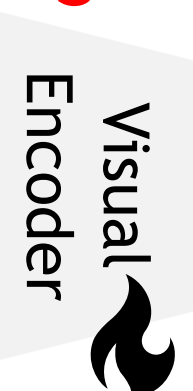
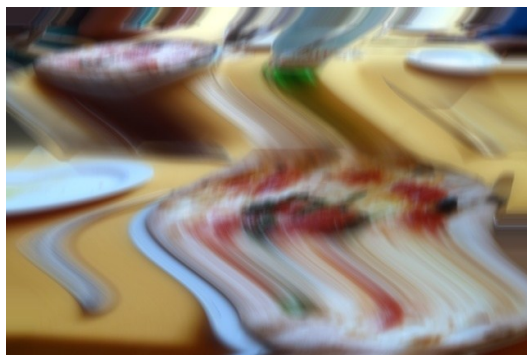
Original
Image



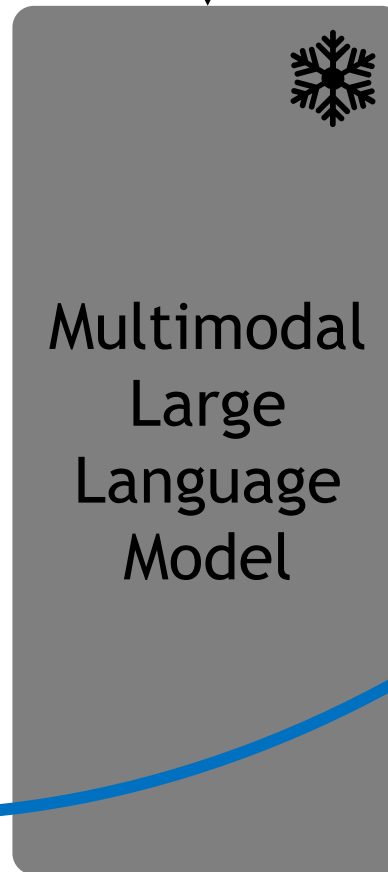
Correct
Answer

Degradation

Degraded
Image



Aligned



Wrong
Answer

Motivation: Explicitly Reasoning (ours)

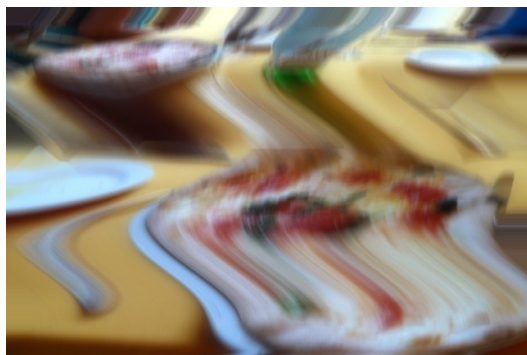
Question

What is next to each pizza pie?

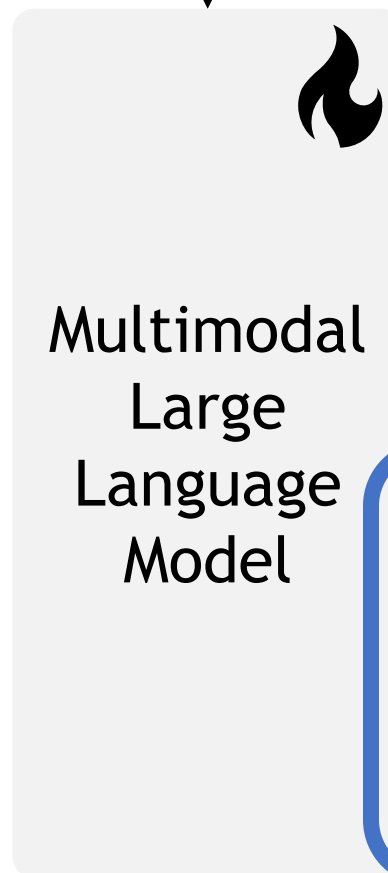
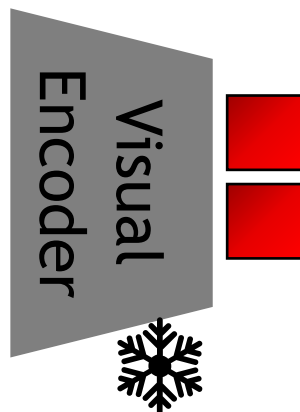
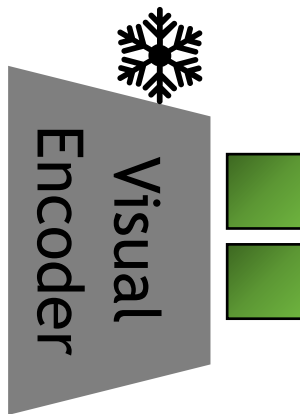
Original
Image



Degradation



Degraded
Image



Correct Answer



Correct Answer

Robust-R1: Degradation-Aware Reasoning

Input



What is next to
each pizza pie?

Robust-R1: Degradation-Aware Reasoning

Input



What is next to
each pizza pie?



MLLM
Thinking...

Robust-R1: Degradation-Aware Reasoning

To Obtain Correct Answer

Input



What is next to each pizza pie?

→ MLLM
Thinking... →

<TYPE>blur(0.48), turbulence(1.0)<TYPE_END>

<INFLUENCE>

High blur and atmospheric turbulence significantly obscure discernible details, making it difficult to identify objects clearly

<INFLUENCE_END>

<REASONING>

Despite severe degradation, certain hints in terms of shape patterns and defined forms can still be assessed to deduce

<REASONING_END>

<CONCLUSION>

..... the long, thin shape near the pizza aligns closely with what would represent a knife.

<CONCLUSION_END>

<ANSWER> **knife** <ANSWER_END>

Design of Reasoning Chain

<TYPE>blur(0.48), turbulence(1.0)<TYPE_END>

<INFLUENCE>

High blur and atmospheric turbulence significantly obscure discernible details, making it difficult to identify objects clearly

<INFLUENCE_END>

<REASONING>

Despite severe degradation, certain hints in terms of shape patterns and outlined forms can still be assessed to deduce

<REASONING_END>

<CONCLUSION>

..... the long, thin shape near the pizza aligns closely with what would represent a knife.

<CONCLUSION_END>

<ANSWER> knife <ANSWER_END>

1

Degradation Types & Intensities

2

Influence of Degradation

3

Reason to Original Semantic Info.

4

Conclusion

5

Answer

Design of Reasoning Chain

<TYPE>blur(0.48), turbulence(1.0)<TYPE_END>

<INFLUENCE>

High blur and atmospheric turbulence significantly obscure discernible details, making it difficult to identify objects clearly

<INFLUENCE_END>

<REASONING>

Despite severe degradation, certain hints in terms of shape patterns and outlined forms can still be assessed to deduce

<REASONING_END>

<CONCLUSION>

..... the long, thin shape near the pizza aligns closely with what would represent a knife.

<CONCLUSION_END>

<ANSWER> knife <ANSWER_END>

1

Degradation Types & Intensities

2

Influence of Degradation

3

Reasoning of Original Info.

4

Conclusion

5

Answer

Design of Reasoning Chain

<TYPE>blur(0.48), turbulence(1.0)<TYPE_END>

<INFLUENCE>

High blur and atmospheric turbulence significantly obscure discernible details, making it difficult to identify objects clearly

<INFLUENCE_END>

<REASONING>

Despite severe degradation, certain hints in terms of shape patterns and outlined forms can still be assessed to deduce

<REASONING_END>

<CONCLUSION>

..... the long, thin shape near the pizza aligns closely with what would represent a knife.

<CONCLUSION_END>

<ANSWER> knife <ANSWER_END>

1

Degradation Types & Intensities

2

Influence of Degradation

3

Reasoning of Original Info.

4

Conclusion

5

Answer

Design of Reasoning Chain

<TYPE>blur(0.48), turbulence(1.0)<TYPE_END>

<INFLUENCE>

High blur and atmospheric turbulence significantly obscure discernible details, making it difficult to identify objects clearly

<INFLUENCE_END>

<REASONING>

Despite severe degradation, certain hints in terms of shape patterns and outlined forms can still be assessed to deduce

<REASONING_END>

<CONCLUSION>

..... the long, thin shape near the pizza aligns closely with what would represent a knife.

<CONCLUSION_END>

<ANSWER> knife <ANSWER_END>

1

Degradation Types & Intensities

2

Influence of Degradation

3

Reasoning of Original Info.

4

Conclusion

5

Answer

Design of Reasoning Chain

<TYPE>blur(0.48), turbulence(1.0)<TYPE_END>

<INFLUENCE>

High blur and atmospheric turbulence significantly obscure discernible details, making it difficult to identify objects clearly

<INFLUENCE_END>

<REASONING>

Despite severe degradation, certain hints in terms of shape patterns and outlined forms can still be assessed to deduce

<REASONING_END>

<CONCLUSION>

..... the long, thin shape near the pizza aligns closely with what would represent a knife.

<CONCLUSION_END>

<ANSWER> knife <ANSWER_END>

1

Degradation Types & Intensities

2

Influence of Degradation

3

Reasoning of Original Info.

4

Conclusion

5

Answer

Design of Reasoning Chain

<TYPE>blur(0.48), turbulence(1.0)<TYPE_END>

<INFLUENCE>

High blur and atmospheric turbulence significantly obscure discernible details, making it difficult to identify objects clearly

<INFLUENCE_END>

<REASONING>

Despite severe degradation, certain hints in terms of shape patterns and outlined forms can still be assessed to deduce

<REASONING_END>

<CONCLUSION>

..... the long, thin shape near the pizza aligns closely with what would represent a knife.

<CONCLUSION_END>

<ANSWER> knife <ANSWER_END>

1

Degradation Types & Intensities

2

Influence of Degradation

3

Reasoning of Original Info.

4

Conclusion

5

Answer

Details of Our Solution

✓ Dataset

1. constructing a dataset of 11K samples
2. featuring realistic degradations across four critical stages
3. annotated with structured reasoning chains for degradation-aware reasoning.

✓ Methodology

1. Supervised Fine-Tuning (SFT): Acquiring Basic Reasoning Ability
2. Reinforcement Learning (RL): (1) Aligning Accurate Degradation Parameters and (2) Scaling to Suitable Reasoning Length

Details of Our Solution

✓ Dataset

1. constructing a dataset of 11K samples
2. featuring realistic degradations across four critical stages
3. annotated with structured reasoning chains for degradation-aware reasoning.

✓ Methodology

1. Supervised Fine-Tuning (SFT): Acquiring Basic Reasoning Ability
2. Reinforcement Learning (RL): (1) Aligning Accurate Degradation Parameters and (2) Scaling to Suitable Reasoning Length

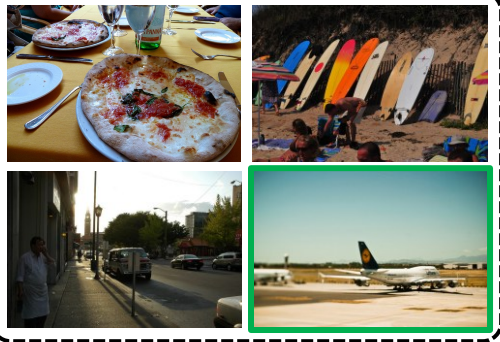
Data Construction - Pipeline

Original Images

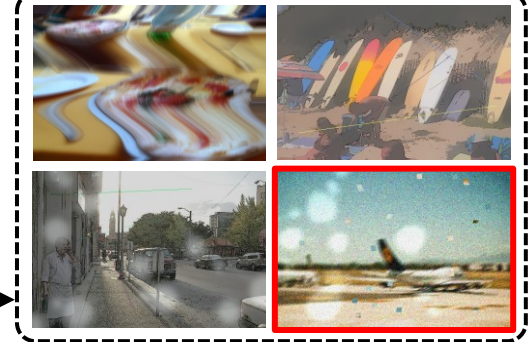


Data Construction - Pipeline

Original Images



Degraded Images



1. Realistic Degradations Synthesis

Data Construction - Pipeline

Original Images



Acquisition 

Degradation Intensity

Lens Blur	×
Lens Flare	0.31
Motion Blur	0.47
Dirty Lens	×
Saturation	×



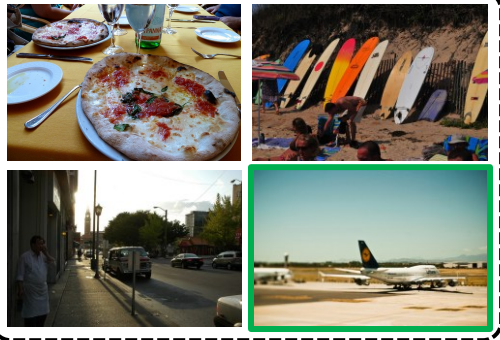
1. Realistic Degradations Synthesis

Degraded Images



Data Construction - Pipeline

Original Images



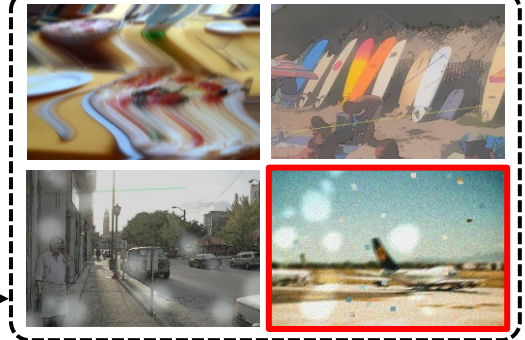
Acquisition 

Degradation Intensity	
Lens Blur	×
Lens Flare	0.31
Motion Blur	0.47
Dirty Lens	×
Saturation	×

Transmission 

Degradation Intensity	
Compression	×
Block Change	0.33
Shifting	0.14
Scan Lines	×

Degraded Images



1. Realistic Degradations Synthesis

Data Construction - Pipeline

Original Images



Acquisition

Degradation Intensity	
Lens Blur	×
Lens Flare	0.31
Motion Blur	0.47
Dirty Lens	×
Saturation	×

Transmission

Degradation Intensity	
Compression	×
Block Change	0.33
Shifting	0.14
Scan Lines	×

Environment

Degradation Intensity	
Darkness	×
Atmospheric Turbulence	0.24
Noise	0.5
Color Diffusion	×

1. Realistic Degradations Synthesis

Degraded Images



Data Construction - Pipeline

Original Images



Acquisition

Degradation Intensity	
Lens Blur	×
Lens Flare	0.31
Motion Blur	0.47
Dirty Lens	×
Saturation	×

Transmission

Degradation Intensity	
Compression	×
Block Change	0.33
Shifting	0.14
Scan Lines	×

Environment

Degradation Intensity	
Darkness	×
Atmospheric Turbulence	0.24
Noise	0.5
Color Diffusion	×

Postprocessing

Degradation Intensity	
Sharpness Change	0.04
Graffiti	×
Watermark Damage	×

Degraded Images



1. Realistic Degradations Synthesis

Data Construction - Pipeline

Original Images



Acquisition

Degradation Intensity

Lens Blur	×
Lens Flare	0.31
Motion Blur	0.47
Dirty Lens	×
Saturation	×

Transmission

Degradation Intensity

Compression	×
Block Change	0.33
Shifting	0.14
Scan Lines	×

Environment

Degradation Intensity

Darkness	×
Atmospheric Turbulence	0.24
Noise	0.5
Color Diffusion	×

Postprocessing

Degradation Intensity

Sharpness Change	0.04
Graffiti	×
Watermark Damage	×

1. Realistic Degradations Synthesis

Degraded Images



<TYPE>	
Lens Flare	0.31
Motion Blur	0.47
Block Change	0.33
Shifting	0.14
Atmospheric Turbulence	0.24
Noise	0.5
Sharpness Change	0.04
Overall	2.03
<TYPE_END>	

Data Construction - Pipeline

Original Images



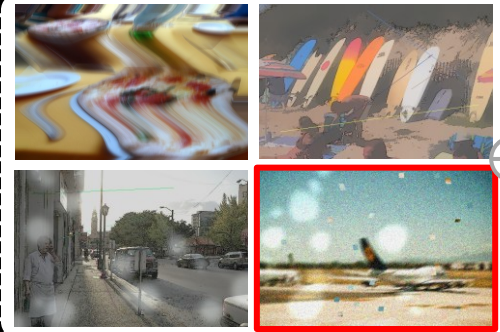
Acquisition	
Degradation Intensity	
Lens Blur	×
Lens Flare	0.31
Motion Blur	0.47
Dirty Lens	×
Saturation	×

Transmission	
Degradation Intensity	
Compression	×
Block Change	0.33
Shifting	0.14
Scan Lines	×

Environment	
Degradation Intensity	
Darkness	×
Atmospheric Turbulence	0.24
Noise	0.5
Color Diffusion	×

Postprocessing	
Degradation Intensity	
Sharpness Change	0.04
Graffiti	×
Watermark Damage	×

Degraded Images



<TYPE>	
Lens Flare	0.31
Motion Blur	0.47
Block Change	0.33
Shifting	0.14
Atmospheric Turbulence	0.24
Noise	0.5
Sharpness Change	0.04
Overall	2.03
<TYPE_END>	

2. Degradation-aware Reasoning Chain Generation

Data Construction - Pipeline

Original Images



Acquisition

Degradation Intensity	
Lens Blur	×
Lens Flare	0.31
Motion Blur	0.47
Dirty Lens	×
Saturation	×

Transmission

Degradation Intensity	
Compression	×
Block Change	0.33
Shifting	0.14
Scan Lines	×

Environment

Degradation Intensity	
Darkness	×
Atmospheric Turbulence	0.24
Noise	0.5
Color Diffusion	×

Postprocessing

Degradation Intensity	
Sharpness Change	0.04
Graffiti	×
Watermark Damage	×

Degraded Images



<TYPE>	
Lens Flare	0.31
Motion Blur	0.47
Block Change	0.33
Shifting	0.14
Atmospheric Turbulence	0.24
Noise	0.5
Sharpness Change	0.04
Overall	2.03
<TYPE_END>	

What is the plane on?

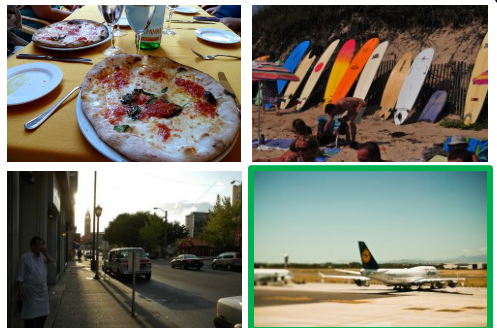
Runway

Original QAs

2. Degradation-aware Reasoning Chain Generation

Data Construction - Pipeline

Original Images



Acquisition

Degradation Intensity	
Lens Blur	×
Lens Flare	0.31
Motion Blur	0.47
Dirty Lens	×
Saturation	×

Transmission

Degradation Intensity	
Compression	×
Block Change	0.33
Shifting	0.14
Scan Lines	×

Environment

Degradation Intensity	
Darkness	×
Atmospheric Turbulence	0.24
Noise	0.5
Color Diffusion	×

Postprocessing

Degradation Intensity	
Sharpness Change	0.04
Graffiti	×
Watermark Damage	×

Degraded Images



<TYPE>	
Lens Flare	0.31
Motion Blur	0.47
Block Change	0.33
Shifting	0.14
Atmospheric Turbulence	0.24
Noise	0.5
Sharpness Change	0.04
Overall	2.03
<TYPE_END>	

What is the plane on?

Runway

OpenAI GPT 4o

EXPLAIN HOW THESE TYPES OF DEGRADATION WILL IMPACT YOUR ANSWERS WHEN YOU ARE ANSWER THIS QUESTION IN THE FORM OF PARAGRAPH

<INFLUENCE>
The gaussian noise and block exchange create a layer of random artifacts and irregular patterns, making it challenging to visually extract precise shapes and boundaries.
.....
Lens flare disrupts
Mean shift slightly alters the color consistency, potentially misleading the understanding of uniformity in surfaces like runways or grasslands.
<INFLUENCE_END>

Influence

2. Degradation-aware Reasoning Chain Generation

Data Construction - Pipeline

Original Images



Acquisition

Degradation Intensity	
Lens Blur	×
Lens Flare	0.31
Motion Blur	0.47
Dirty Lens	×
Saturation	×

Transmission

Degradation Intensity	
Compression	×
Block Change	0.33
Shifting	0.14
Scan Lines	×

Environment

Degradation Intensity	
Darkness	×
Atmospheric Turbulence	0.24
Noise	0.5
Color Diffusion	×

Postprocessing

Degradation Intensity	
Sharpness Change	0.04
Graffiti	×
Watermark Damage	×

Degraded Images



<TYPE>	
Lens Flare	0.31
Motion Blur	0.47
Block Change	0.33
Shifting	0.14
Atmospheric Turbulence	0.24
Noise	0.5
Sharpness Change	0.04
Overall	2.03
<TYPE_END>	

What is the plane on?



EXPLAIN HOW THESE TYPES OF DEGRADATION WILL IMPACT YOUR ANSWERS WHEN YOU ARE ANSWER THIS QUESTION IN THE FORM OF PARAGRAPH

<INFLUENCE>
The gaussian noise and block exchange create a layer of random artifacts and irregular patterns, making it challenging to visually extract precise shapes and boundaries.
Lens flare disrupts Mean shift slightly alters the color consistency, potentially misleading the understanding of uniformity in surfaces like runways or grasslands.
<INFLUENCE_END>



ANALYZE THE QUESTION AND PROVIDE REASONING IN THE FORM OF PARAGRAPH TO ANSWER

<REASONING>
The question involves identifying the surface on which the plane is situated. Despite the quality issues, certain details are discernible. The plane appears to be
The pixel variations, despite blurred edges due to motion blur and atmospheric disturbances, still allow some visible segments that align with a runway's characteristics.
<REASONING_END>

Original QAs $\Psi_{INFLUENCE}$ Influence $\Psi_{REASONING}$ Reasoning

2. Degradation-aware Reasoning Chain Generation

Data Construction - Pipeline

Original Images



Acquisition	
Degradation Intensity	
Lens Blur	×
Lens Flare	0.31
Motion Blur	0.47
Dirty Lens	×
Saturation	×

Transmission	
Degradation Intensity	
Compression	×
Block Change	0.33
Shifting	0.14
Scan Lines	×

Environment	
Degradation Intensity	
Darkness	×
Atmospheric Turbulence	0.24
Noise	0.5
Color Diffusion	×

Postprocessing	
Degradation Intensity	
Sharpness Change	0.04
Graffiti	×
Watermark Damage	×

Degraded Images



<TYPE>	
Lens Flare	0.31
Motion Blur	0.47
Block Change	0.33
Shifting	0.14
Atmospheric Turbulence	0.24
Noise	0.5
Sharpness Change	0.04
Overall	2.03
<TYPE_END>	

What is the plane on?

Runway

OpenAI GPT 4o

EXPLAIN HOW THESE TYPES OF DEGRADATION WILL IMPACT YOUR ANSWERS WHEN YOU ARE ANSWER THIS QUESTION IN THE FORM OF PARAGRAPH

<INFLUENCE>
The gaussian noise and block exchange create a layer of random artifacts and irregular patterns, making it challenging to visually extract precise shapes and boundaries.
Lens flare disrupts Mean shift slightly alters the color consistency, potentially misleading the understanding of uniformity in surfaces like runways or grasslands.
<INFLUENCE_END>

OpenAI GPT 4o

ANALYZE THE QUESTION AND PROVIDE REASONING IN THE FORM OF PARAGRAPH TO ANSWER

<REASONING>
The question involves identifying the surface on which the plane is situated. Despite the quality issues, certain details are discernible. The plane appears to be
The pixel variations, despite blurred edges due to motion blur and atmospheric disturbances, still allow some visible segments that align with a runway's characteristics.
<REASONING_END>

OpenAI GPT 4o

SUMMARIZE THE CONTENT OF REASONING AND THEN GIVE THE CONCLUSION

<CONCLUSION>
The visible traits amidst gaussian noise, lens flare, and other factors suggest a runway is the most plausible interpretation.
<CONCLUSION_END>

<ANSWER>
Runway
<ANSWER_END>

Original QAs $\Psi_{INFLUENCE}$ Influence $\Psi_{REASONING}$ Reasoning $\Psi_{CONCLUSION}$ Conclusion

2. Degradation-aware Reasoning Chain Generation

Data Construction - Pipeline

Original Images



Acquisition	
Degradation Intensity	
Lens Blur	×
Lens Flare	0.31
Motion Blur	0.47
Dirty Lens	×
Saturation	×

Transmission	
Degradation Intensity	
Compression	×
Block Change	0.33
Shifting	0.14
Scan Lines	×

Environment	
Degradation Intensity	
Darkness	×
Atmospheric Turbulence	0.24
Noise	0.5
Color Diffusion	×

Postprocessing	
Degradation Intensity	
Sharpness Change	0.04
Graffiti	×
Watermark Damage	×

Degraded Images



<TYPE>	
Lens Flare	0.31
Motion Blur	0.47
Block Change Shifting	0.33
Atmospheric Turbulence	0.24
Noise	0.5
Sharpness Change	0.04
Overall	2.03
<TYPE_END>	

What is the plane on?

OpenAI GPT 4o

<INFLUENCE>
The gaussian noise and block exchange create a layer of random artifacts and irregular patterns, making it challenging to visually extract precise shapes and boundaries.
Lens flare disrupts Mean shift slightly alters the color consistency, potentially misleading the understanding of uniformity in surfaces like runways or grasslands.
<INFLUENCE_END>

OpenAI GPT 4o

<REASONING>
The question involves identifying the surface on which the plane is situated. Despite the quality issues, certain details are discernible. The plane appears to be
The pixel variations, despite blurred edges due to motion blur and atmospheric disturbances, still allow some visible segments that align with a runway's characteristics.
<REASONING_END>

OpenAI GPT 4o

<CONCLUSION>
The visible traits amidst gaussian noise, lens flare, and other factors suggest a runway is the most plausible interpretation.
<CONCLUSION_END>
<ANSWER>
Runway
<ANSWER_END>

OpenAI GPT 4o

ORIGINAL CHAIN
SCALING CHAIN BY INTENSITY

Dynamic Compression

EXPLAIN HOW THESE TYPES OF DEGRADATION WILL IMPACT YOUR ANSWERS WHEN YOU ARE ANSWER THIS QUESTION IN THE FORM OF PARAGRAPH

ANALYZE THE QUESTION AND PROVIDE REASONING IN THE FORM OF PARAGRAPH TO ANSWER

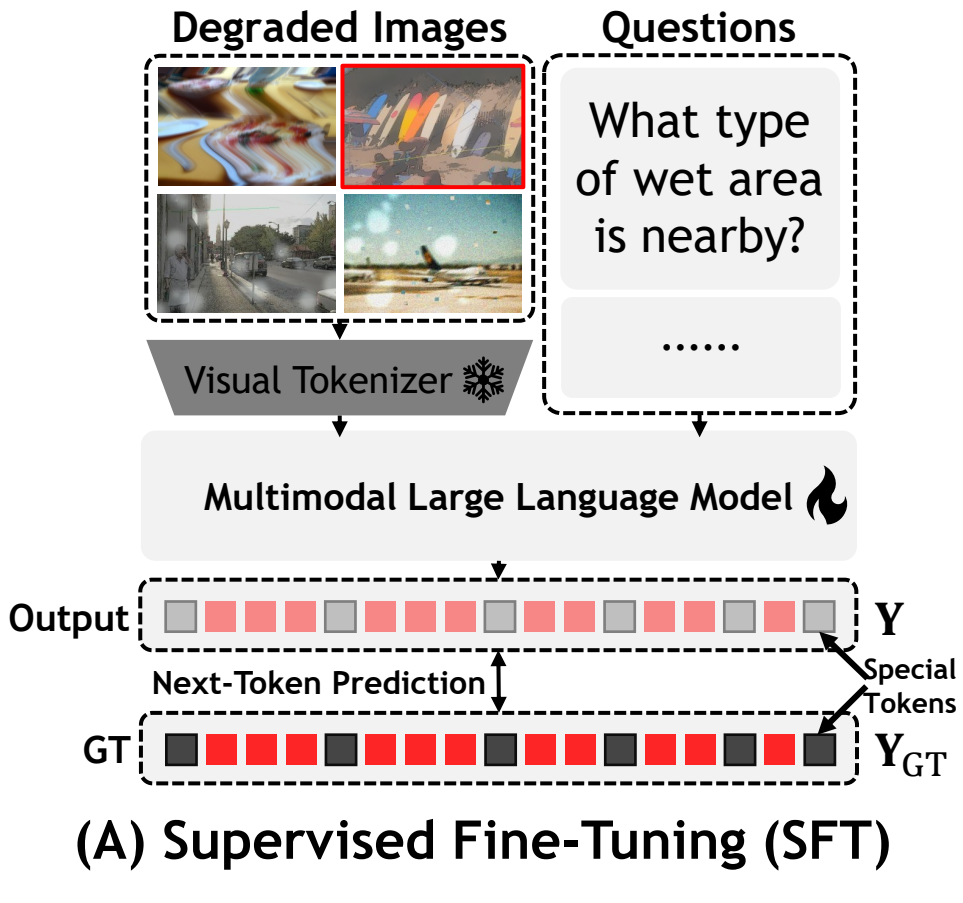
SUMMARIZE THE CONTENT OF REASONING AND THEN GIVE THE CONCLUSION

COMPRESS THE FOLLOWING REASONING CHAIN BASED ON DEGRADATION INTENSITY. PRESERVE CRITICAL INFORMATION REMOVING REDUNDANCY

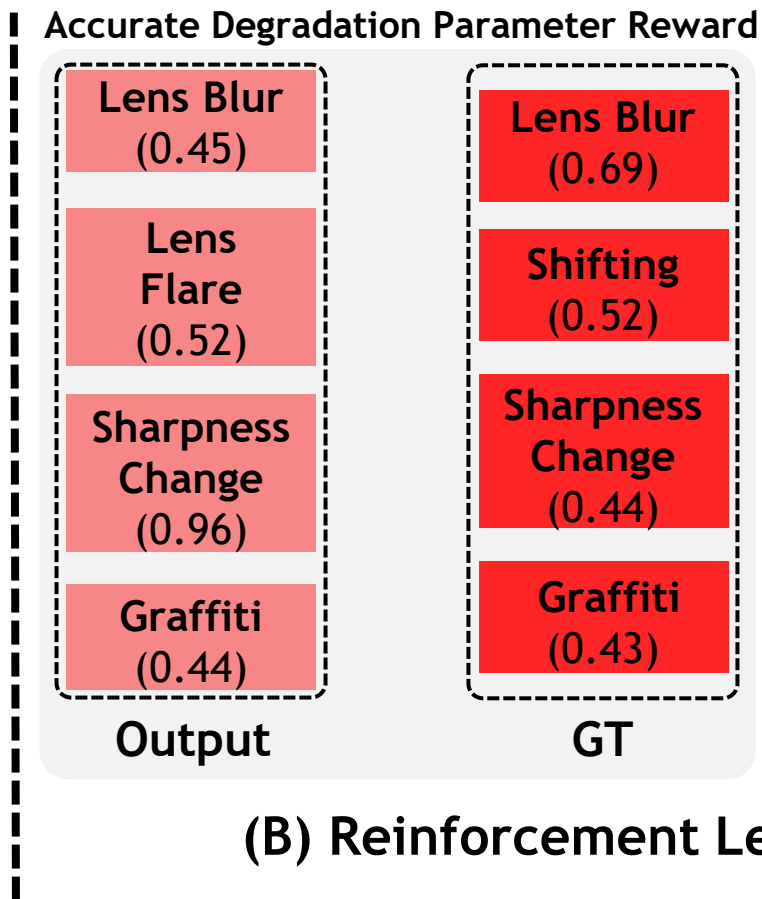
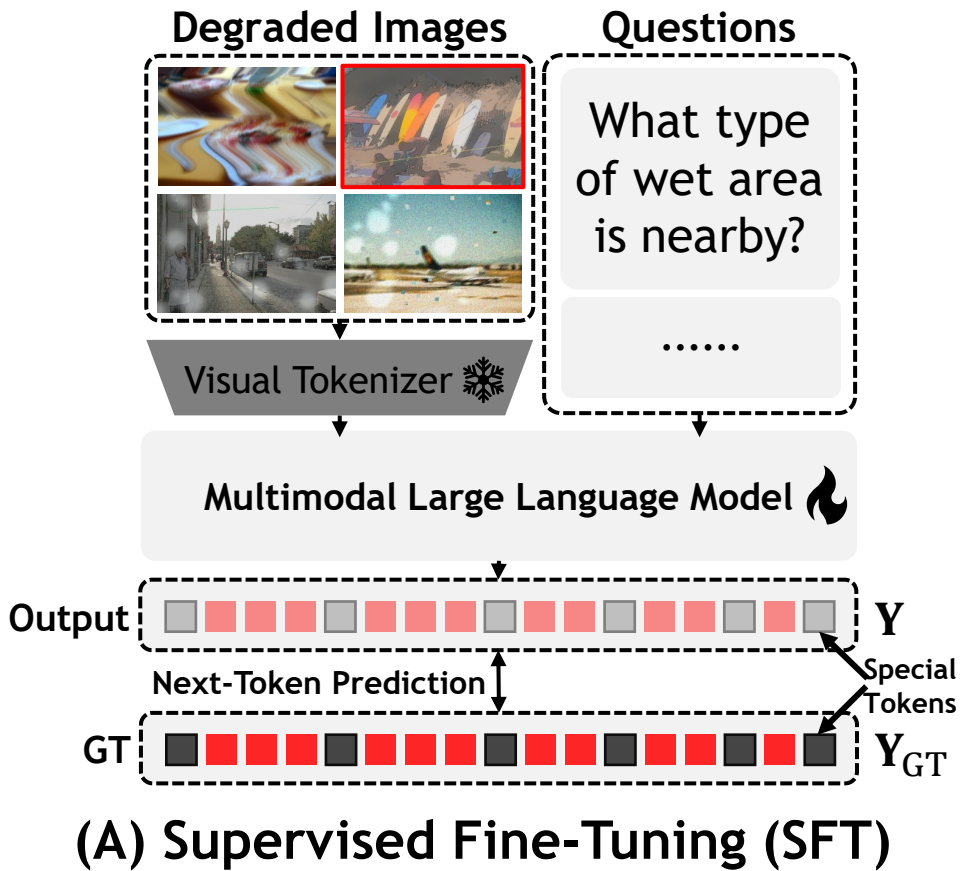
Original QAs $\Psi_{INFLUENCE}$ Influence $\Psi_{REASONING}$ Reasoning $\Psi_{CONCLUSION}$ Conclusion Ψ_{Len} Scaling

3. Scaling to Suitable Length

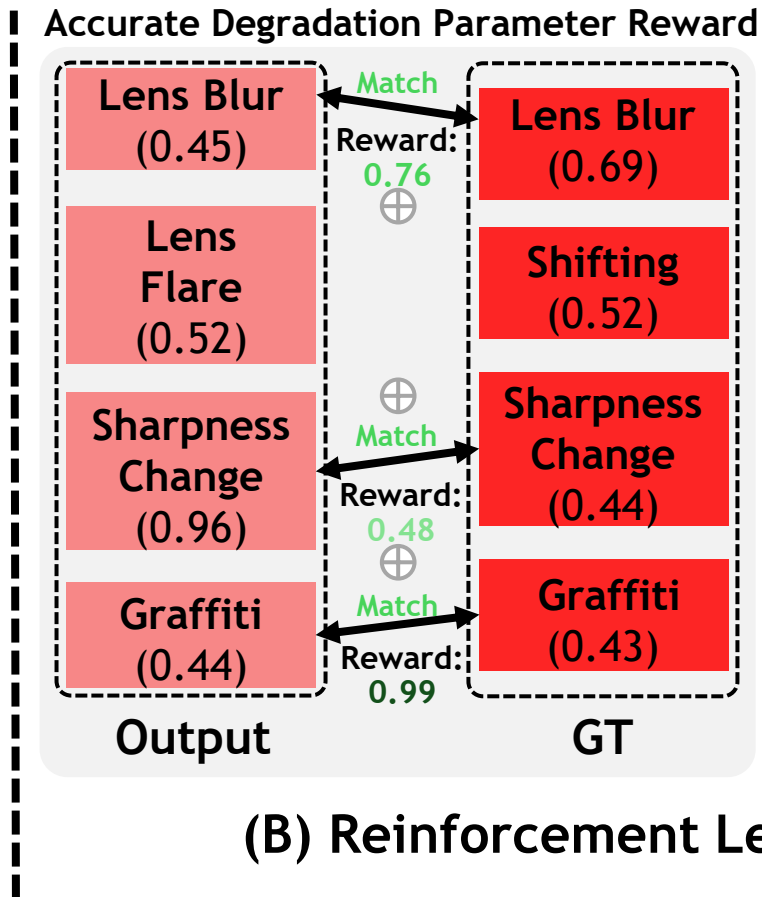
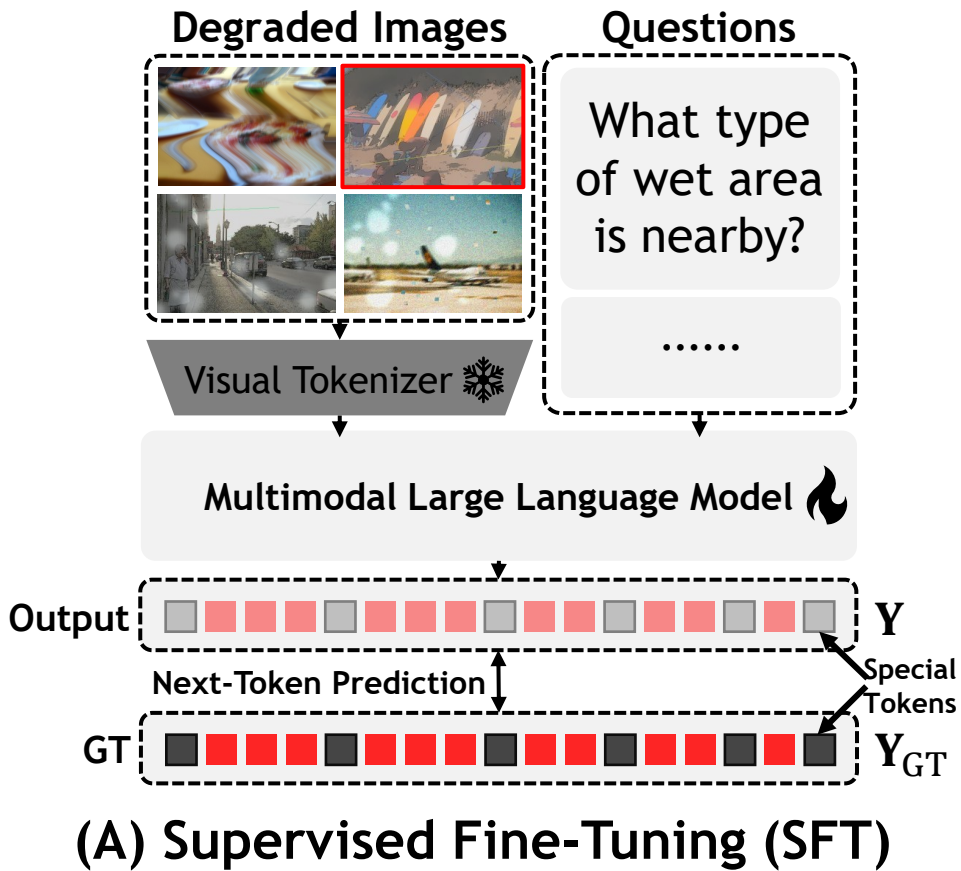
Methodology - SFT



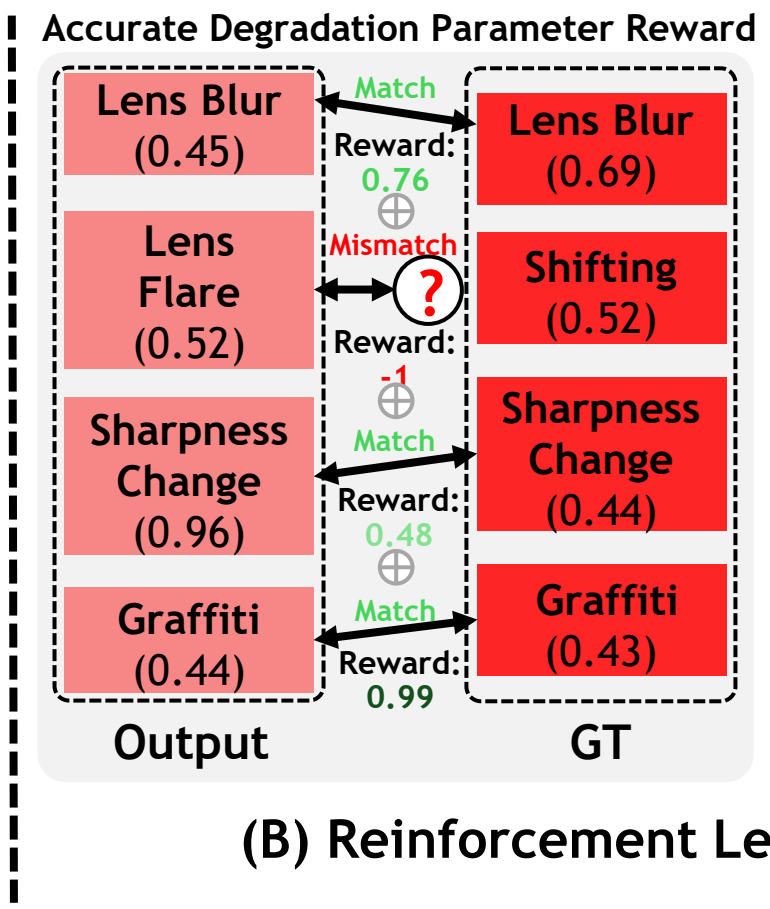
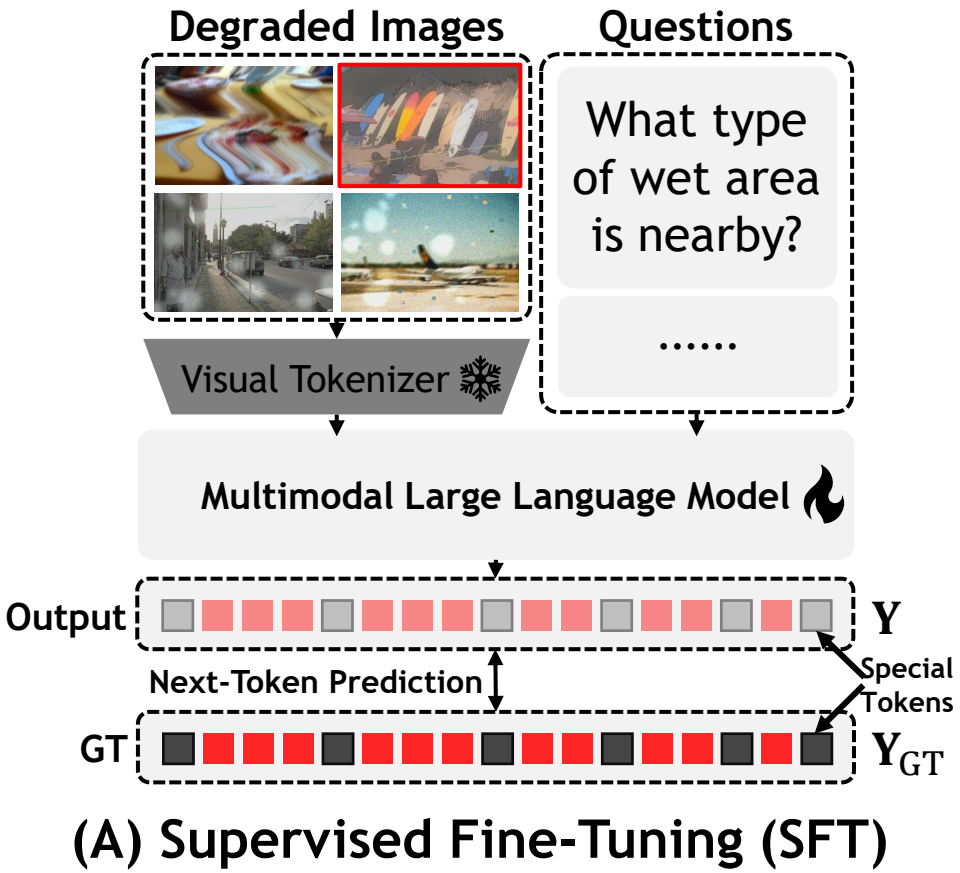
Methodology - RL



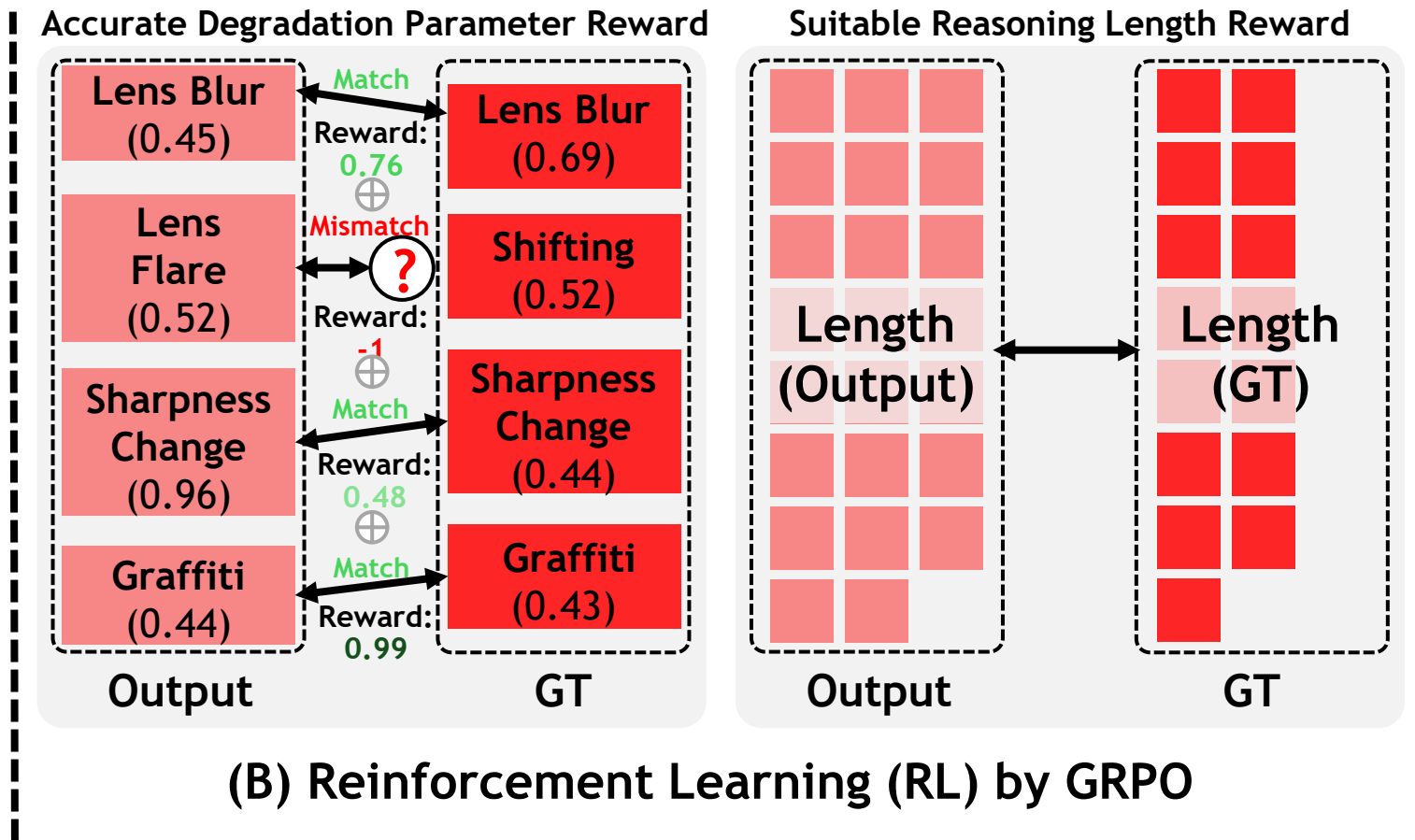
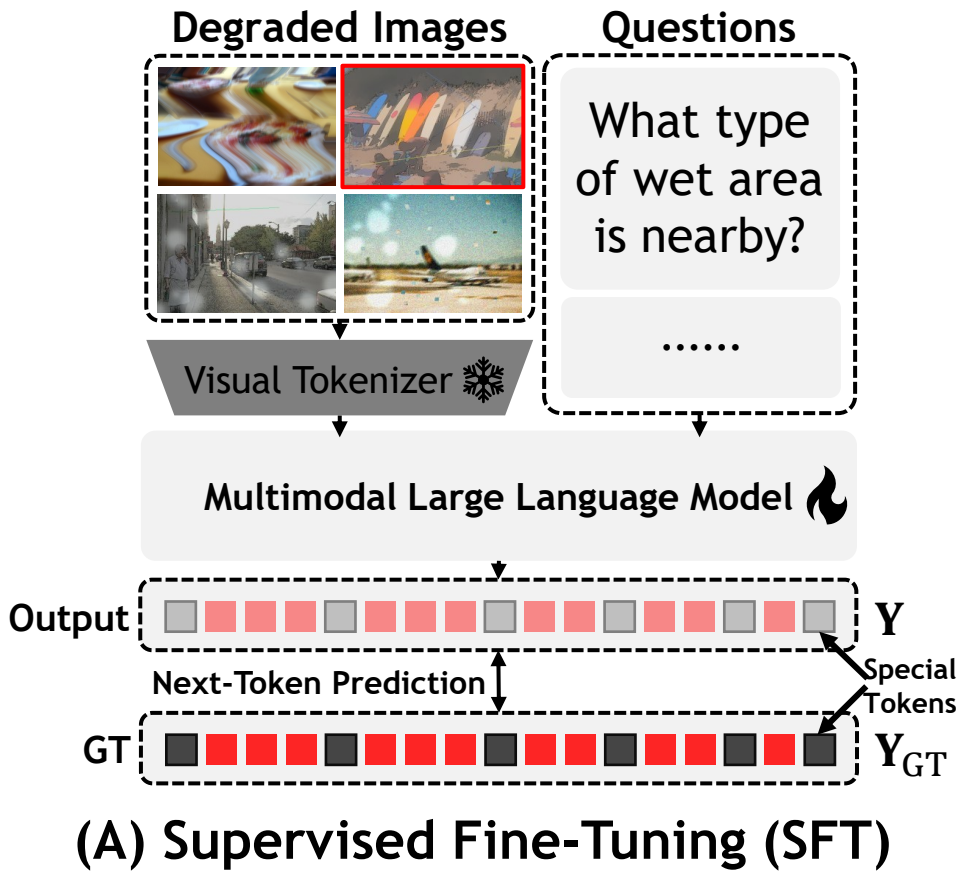
Methodology - RL



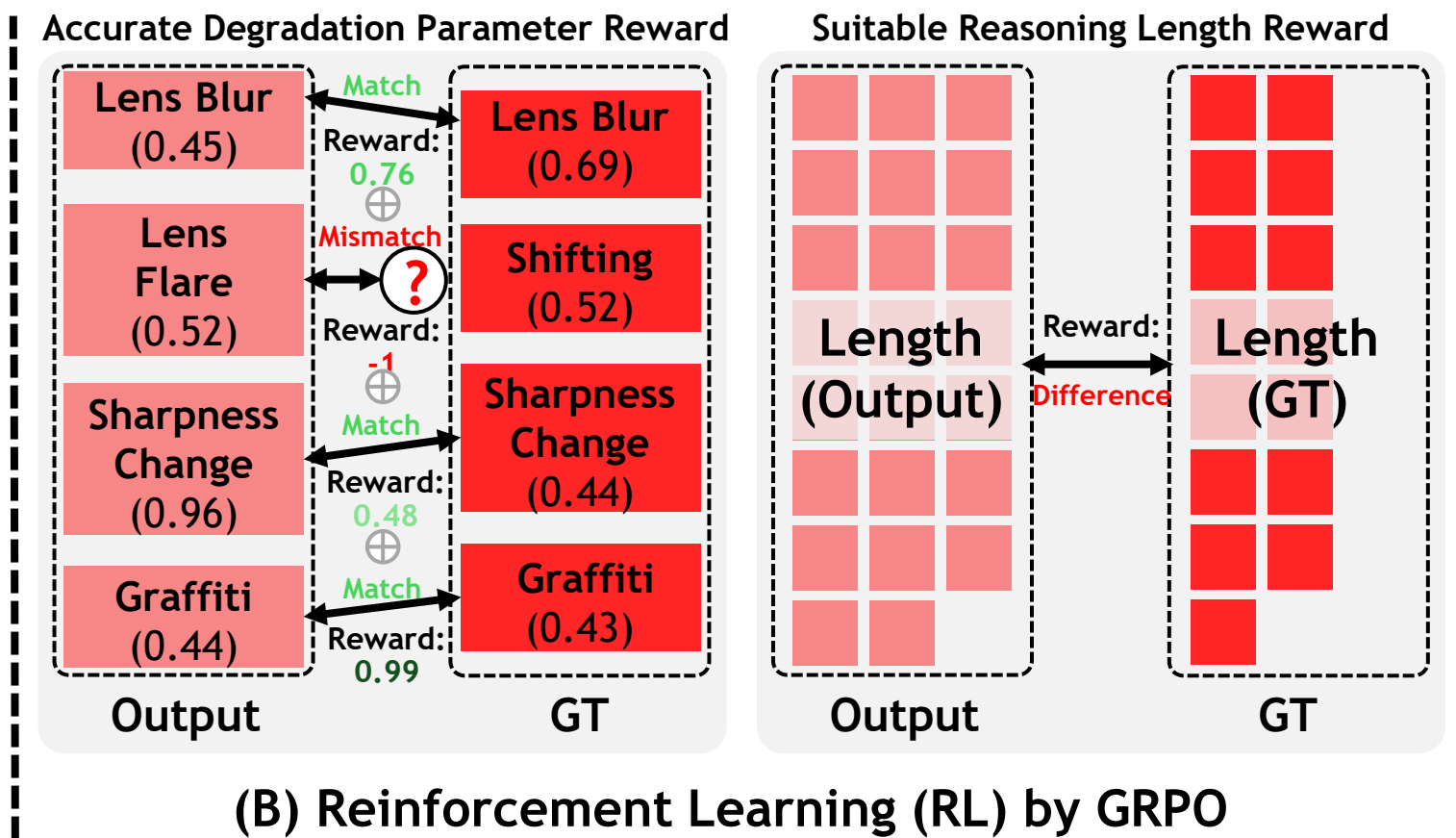
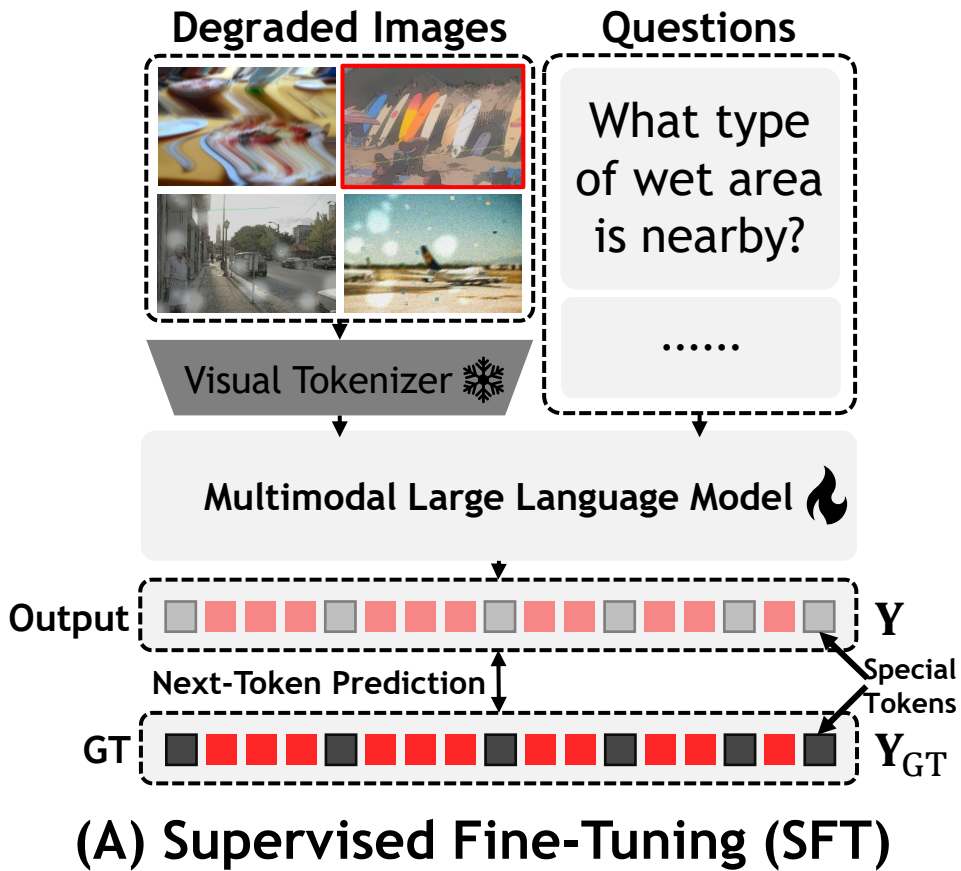
Methodology - RL



Methodology - RL



Methodology - RL



Experiments

✓ Baselines

1. General MLLMs, including **Qwen2.5-VL-3B** (Bai et al. 2025), **Gemma3-4B** (Team et al. 2025), and **InternVL-4B** (Chen et al. 2024);
2. Robust MLLMs, including **TeCoA** (Wang et al. 2024), **Robust CLIP** (Schlarmann et al. 2024), and **Robust LLaVA** (Malik et al. 2025).

✓ Benchmarks

1. Real-World Robustness: Directly assessing robust visual understanding ability on **R-Bench** (Li et al. 2024);
2. Adversarial Robustness: Evaluation under synthetic degradation attacks by applying multi-type, multi-level real-world degradations to visual content in **MMMB** (Sun et al. 2025), **MMStar** (Chen et al. 2024), and **RealWorldQA** (xAI 2024).

Experiments

✓ Baselines

1. General MLLMs, including Qwen2.5-VL-3B (Bai et al. 2025), Gemma3-4B (Team et al. 2025), and InternVL-4B (Chen et al. 2024);
2. Robust MLLMs, including TeCoA (Wang et al. 2024), Robust CLIP (Schlarmann et al. 2024), and Robust LLaVA (Malik et al. 2025).

✓ Benchmarks

1. Real-World Robustness: Directly assessing robust visual understanding ability on **R-Bench** (Li et al. 2024);
2. Adversarial Robustness: Evaluation under synthetic degradation attacks by applying multi-type, multi-level real-world degradations to visual content in **MMMB** (Sun et al. 2025), **MMStar** (Chen et al. 2024), and **RealWorldQA** (xAI 2024).

Experiments - Real-World Robustness (R-Bench)



Category	Method	MCQ			VQA			CAP			Overall
		low	mid	high	low	mid	high	low	mid	high	
General MLLM	Qwen2.5-VL-3B (Bai et al. 2025)	0.6411	0.6022	0.5732	0.4872	0.4854	0.4904	0.3778	0.3704	0.3330	0.4845
	Gemma3-4B (Team et al. 2025)	0.5823	0.5776	0.5060	0.4865	0.4630	0.4419	0.4048	0.3746	0.3480	0.4649
	InternVL-4B (Chen et al. 2024b)	0.6235	0.6024	0.5914	0.4982	0.4539	0.5108	0.3667	0.3041	0.2851	0.4706
Robust MLLM	TeCoA (Wang et al. 2024b)	0.4647	0.4223	0.4024	0.4687	0.3994	0.4461	0.2111	0.2195	0.1937	0.3586
	Robust CLIP (Schlarmann et al. 2024)	0.4705	0.4658	0.4024	0.4503	0.4339	0.4743	0.2290	0.2219	0.1983	0.3718
	Robust LLaVA (Malik et al. 2025)	0.3352	0.2608	0.3048	0.2607	0.2212	0.2443	0.0068	0.0065	0.0067	0.1830
Ours	SFT	0.6176	0.6087	0.5610	0.4804	0.4836	0.5012	0.4080	0.3858	0.3518	0.4886
	SFT + RL	0.6529	0.6391	0.6097	0.4914	0.4909	0.4980	0.4068	0.3781	0.3484	0.5017

- MCQ/VQA/CAP tasks with three degradation strength levels (from low to high).
- The best/second best results are shown in Red/Blue respectively.

Experiments - Anti-Degradation

Category	Method	MMMB (Sun et al. 2025)				MMStar (Chen et al. 2024a)				RealWorldQA (xAI 2024)			
		clean	25%	Intensity 50%	100%	clean	25%	Intensity 50%	100%	clean	25%	Intensity 50%	100%
General MLLM	Qwen2.5-VL-3B (Bai et al. 2025)	80.60	79.19	78.68	74.50	54.73	52.90	51.86	48.66	65.22	64.96	63.39	60.65
	Gemma3-4B (Team et al. 2025)	71.01	70.30	70.20	69.14	43.93	43.20	42.60	41.33	55.42	54.77	53.72	52.81
	InternVL-4B (Chen et al. 2024b)	77.97	77.47	76.66	74.59	51.53	50.26	49.60	46.93	57.38	58.16	57.64	54.90
Robust MLLM	TeCoA (Wang et al. 2024b)	57.17	65.71	56.11	51.76	30.46	30.60	30.73	28.06	40.00	39.73	39.47	38.69
	Robust CLIP (Schlarmann et al. 2024)	58.83	58.28	57.97	53.33	33.00	32.26	31.80	29.46	43.26	42.48	42.61	41.43
Ours	SFT	80.85	79.45	78.68	74.94	55.20	53.00	51.86	49.53	68.23	67.58	67.32	63.92
	SFT + RL	81.41	79.49	79.04	75.35	56.86	54.40	53.60	49.53	67.71	66.40	67.05	63.26

- Three visual understanding benchmarks with three degradation intensity levels (from 25% to 100%).
- The best/second best results are shown in Red/Blue respectively.

Experiments - Qualitative Evaluation

➤ Better in Performance and Efficiency

What is fake in this photo?

- 0. bears
- 1. plants
- 2. grass
- 3. trees



lens flare (0.05)
lens blur (0.29)

Qwen2.5-VL

The fake elements in the photo include:

- 0. bears
- 1. plants
- 2. trees

All of these elements are artificial or sculpted, not real.

Wrong

Ours (SFT)

<TYPE>
lens blur(0.47), jpeg compression(0.58)
<TYPE_END> (WRONG TYPE)
<INFLUENCE>The lens blur at 0.47 can cause a general softening effect across the entire image, making it difficult to discern fine details such as textures or small objects. JPEG compression at 0.58 may introduce artifacts that further obscure clarity, particularly affecting areas with high contrast or intricate patterns.<INFLUENCE_END>
<REASONING>The image depicts a garden scene with bear statues and surrounding elements like rocks, trees, and grass. The bear statues appear to be made of stone or metal, which is typically used for decorative purposes rather than natural materials like wood or living plants. The presence of artificial structures like the bear statues suggests they are not real bears but rather sculptures or models. The other options, such as plants, grass, and trees, are natural elements that would likely show more detail and texture if they were real.<REASONING_END>
<CONCLUSION>The bear statues in the image are fake, as they are made from materials that are not natural and are designed for decorative purposes. Therefore, the correct answer is 0.<CONCLUSION_END>
<ANSWER>0<ANSWER_END>

Correct

Ours (SFT+RL)

<TYPE>
lens blur(0.31), lens flare(0.05)
<TYPE_END>
<INFLUENCE>The lens blur affects the clarity of the image, making it slightly less detailed. This can make it challenging to discern finer details or textures, such as the texture of the bears' fur or the surface of the rocks.<INFLUENCE_END>
<REASONING>The image depicts three bear statues placed in a garden setting. The bears appear to be made of stone or a similar material, which is consistent with the appearance of fake objects.<REASONING_END>
<CONCLUSION>The bears in the image are fake, as they are made of stone or a similar material, and the surrounding environment appears artificial. Therefore, the correct answer is 0.<CONCLUSION_END>
<ANSWER>0<ANSWER_END>

Correct

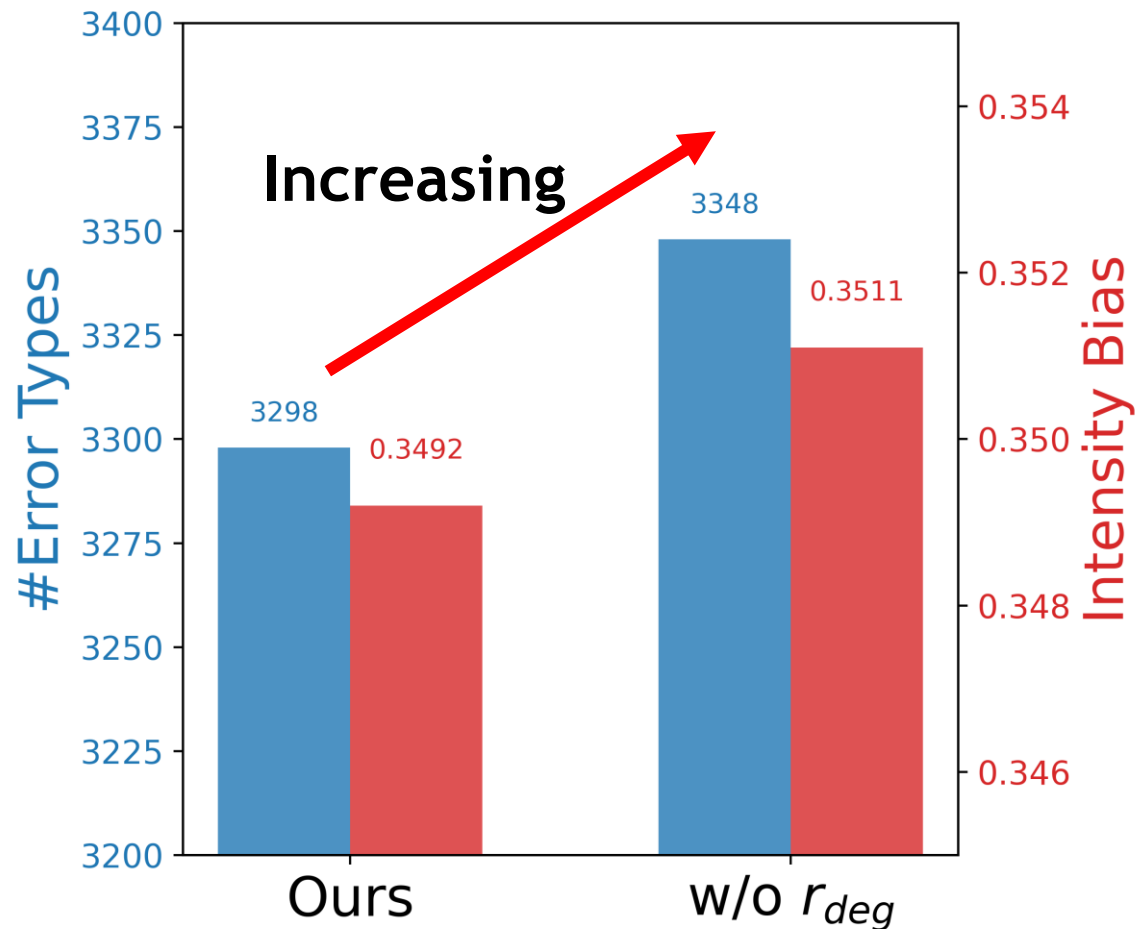
Experiments - Ablation Study

Method	MCQ			VQA			CAP			Overall
	low	mid	high	low	mid	high	low	mid	high	
Qwen2.5-VL-3B (Bai et al. 2025)	0.6411	0.6022	0.5732	0.4872	0.4854	0.4904	0.3778	0.3704	0.3330	0.4845
Ours (w/o Reasoning)	0.6588	0.5901	0.4756	0.4905	0.4900	0.4862	0.2901	0.2673	0.2758	0.4471
Ours (w/o r_{deg})	0.6647	0.6398	0.5505	0.4912	0.4894	0.5056	0.3684	0.3578	0.3248	0.4880
Ours (w/o r_{len})	0.6647	0.6354	0.5975	0.4904	0.4887	0.4877	0.3656	0.3678	0.3189	0.4907
Ours	0.6529	0.6391	0.6097	0.4914	0.4909	0.4980	0.4068	0.3781	0.3484	0.5017

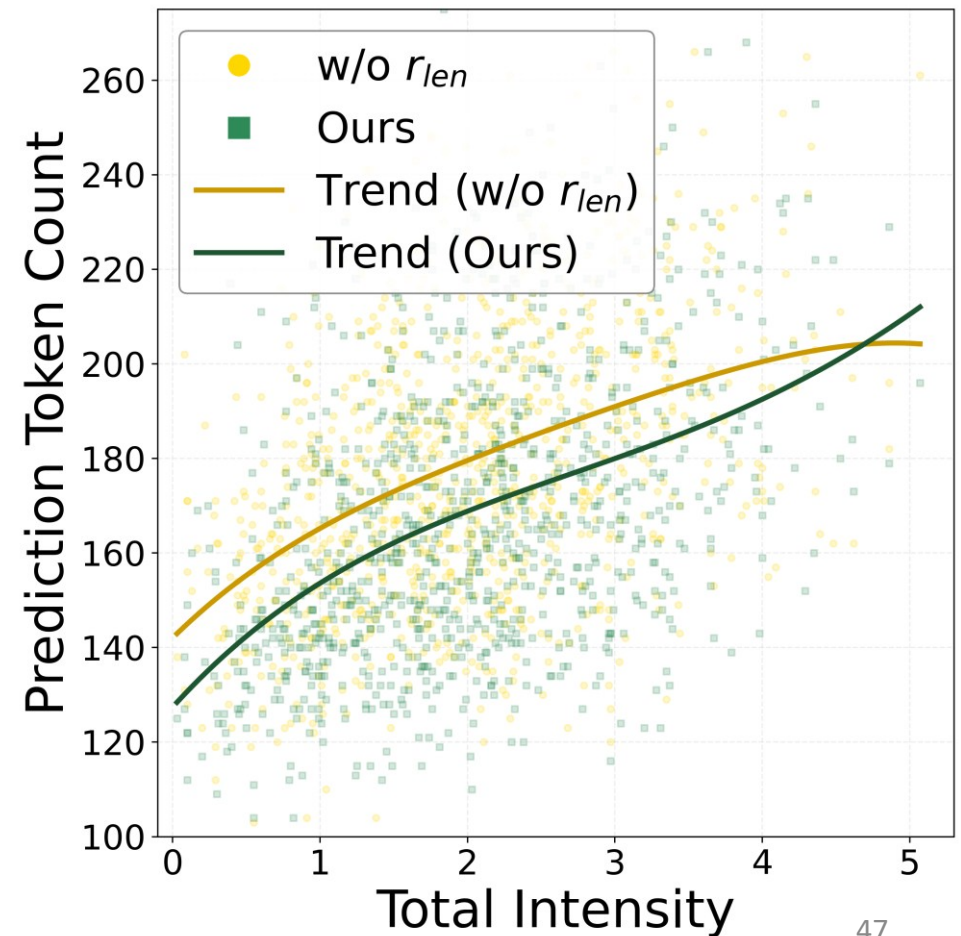
➤ The best/second best results are shown in Red/Blue respectively.

Experiments - Ablation Study

(A) Error Analysis of Accurate Degradation Parameter Reward



(B) Token Efficiency of Suitable Reasoning Length Reward



Conclusion

- ✓ **Robust-R1** is a new paradigm for robust adversarial learning of multimodal large models by **reasoning**.
- ✓ A new direction for building more robust, interpretable, and efficient multimodal systems in visually challenging environments.

Conclusion

- ✓ Robust-R1 is a new paradigm for robust adversarial learning of multimodal large models by **reasoning**.
- ✓ A **new direction** for building more robust, interpretable, and efficient multimodal systems in visually challenging environments.

Thanks

01/2026



Personal Page
jqt.me



WeChat



Fully Open-Sourced, Please Enjoy!

- Code: github.com/jqtangust/Robust-R1
- Data: huggingface.co/datasets/Jiaqi-hkust/Robust-R1
- Model: huggingface.co/Jiaqi-hkust/Robust-R1
- Space: huggingface.co/spaces/Jiaqi-hkust/Robust-R1



香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY