Towards Real-World Adverse Weather Image Restoration: Enhancing Clearness and Semantics with Vision-Language Models

Jiaqi Xu[1], Mengyang Wu[1], Xiaowei Hu[2,⋆],
Chi-Wing Fu[1], Qi Dou[1], and Pheng-Ann Heng[1]

[1] The Chinese University of Hong Kong
[2] Shanghai Artificial Intelligence Laboratory

ECCV 2024

Presenter: Hao Wang

Advisor: Prof. Chia-Wen Lin

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

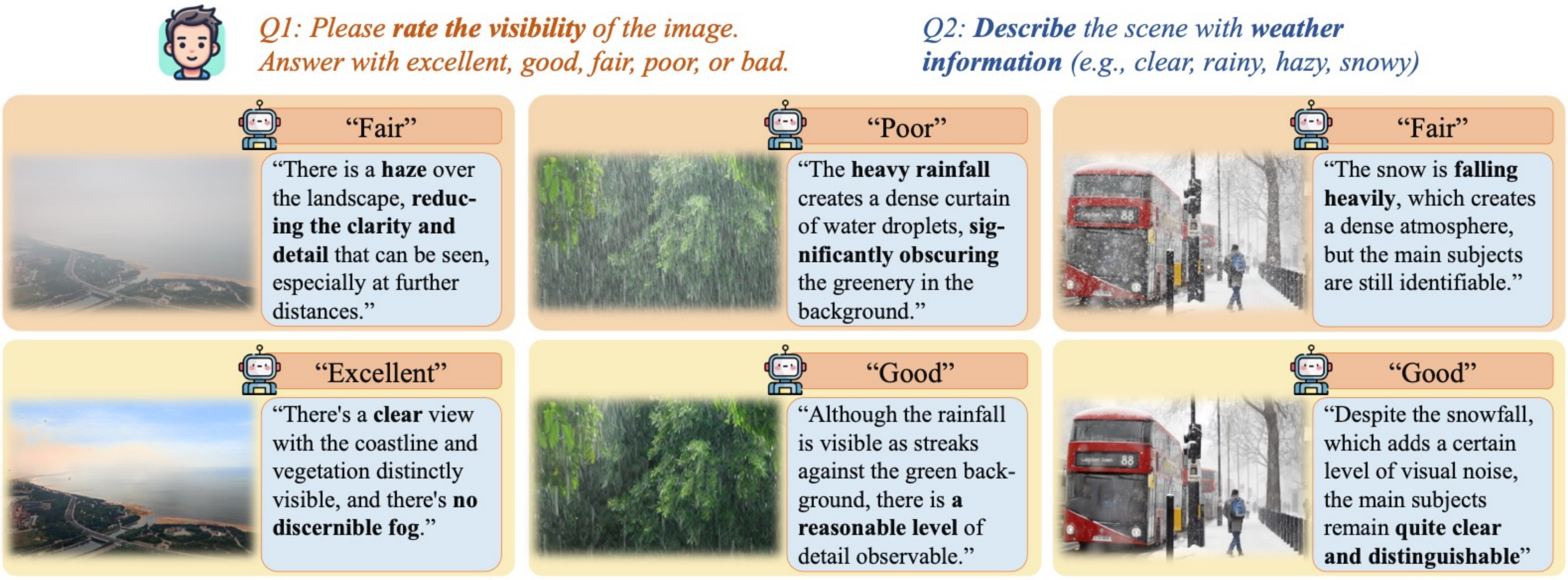
Introduction

- Formulate a **semi-supervised learning framework** employing **vision-language models** to enhance restoration performance across diverse adverse weather conditions in real-world settings

- This approach involves assessing **image clearness** and providing **semantics** using vision-language models on real data, serving as supervision signals for training restoration models.

- Achieves superior results in real-world adverse weather image restoration, demonstrated through qualitative and quantitative comparisons with state-of-the-art works.

Introduction

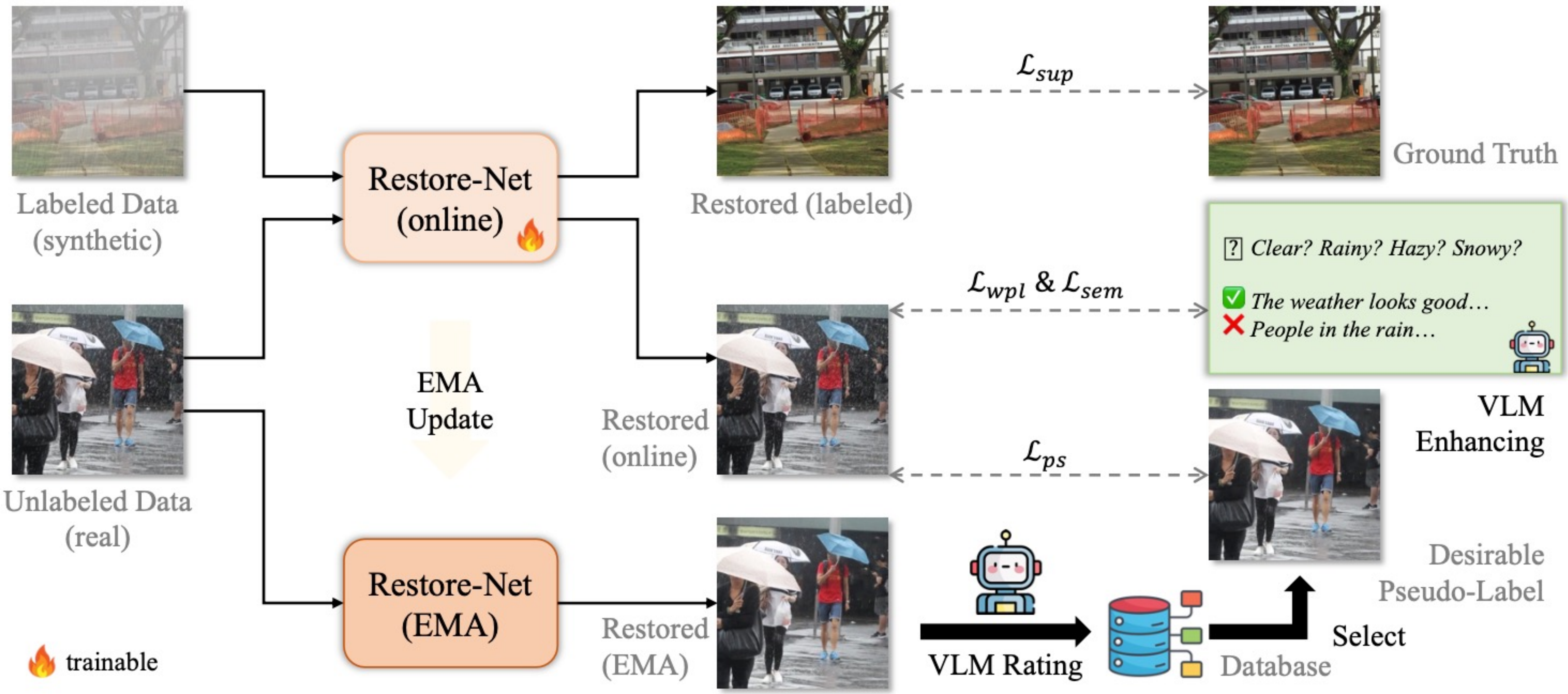


- The clearness level and the semantics information of real-world adverse weather images are provided by large vision-language models.

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Framework


Labeled Data
(synthetic)
Unlabeled Data
(real)
Restore-Net
(online)
EMA
Update
Restore-Net
(EMA)
trainable
Restored (labeled)
Restored
(online)
Restored
(EMA)
$\mathcal{L}_{sup}$
$\mathcal{L}_{wpl}$ & $\mathcal{L}_{sem}$
$\mathcal{L}_{ps}$
Ground Truth
Clear? Rainy? Hazy? Snowy?
The weather looks good…
People in the rain…
VLM
Enhancing
Desirable
Pseudo-Label
VLM Rating
Database
Select

Outline

Image Assessment

(a)

User <img> Rate the visibility of the image against rain, fog, and snow. Please answer with excellent, good, fair, poor, or bad.

Visual Encoder

Tokenizer

VLM
(e.g. LLaVA)

The | visibility | of | the | image | is | ...... | <rating>

token	excellent	good	fair	poor	bad
logit	13.1	16.2	18.8	12.6	11.5

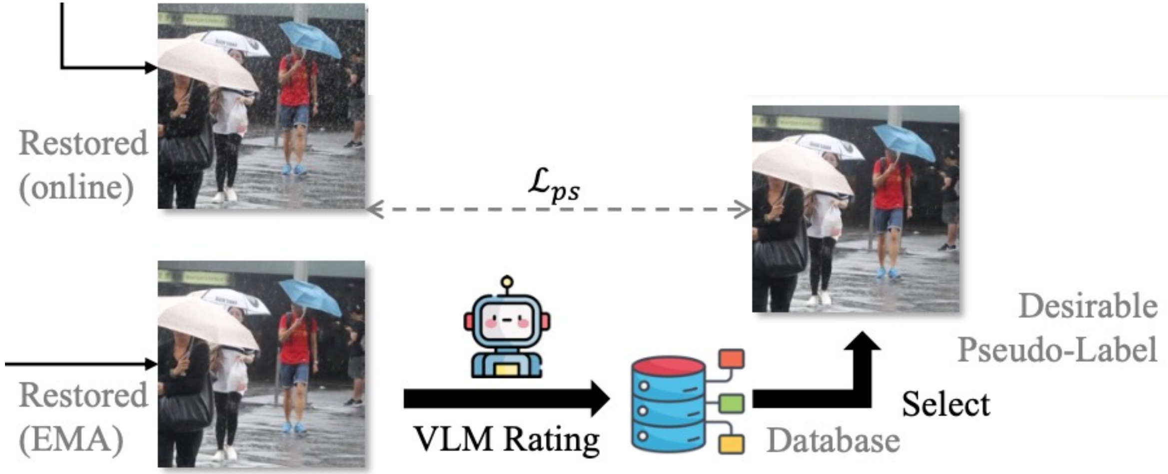
$$r^{vlm} = \sum_{i=1}^{5} i \times p_i \, , p_i = \sigma(l)_i = \frac{e^{l_i}}{\sum_{j=1}^{5} e^{l_j}}$$

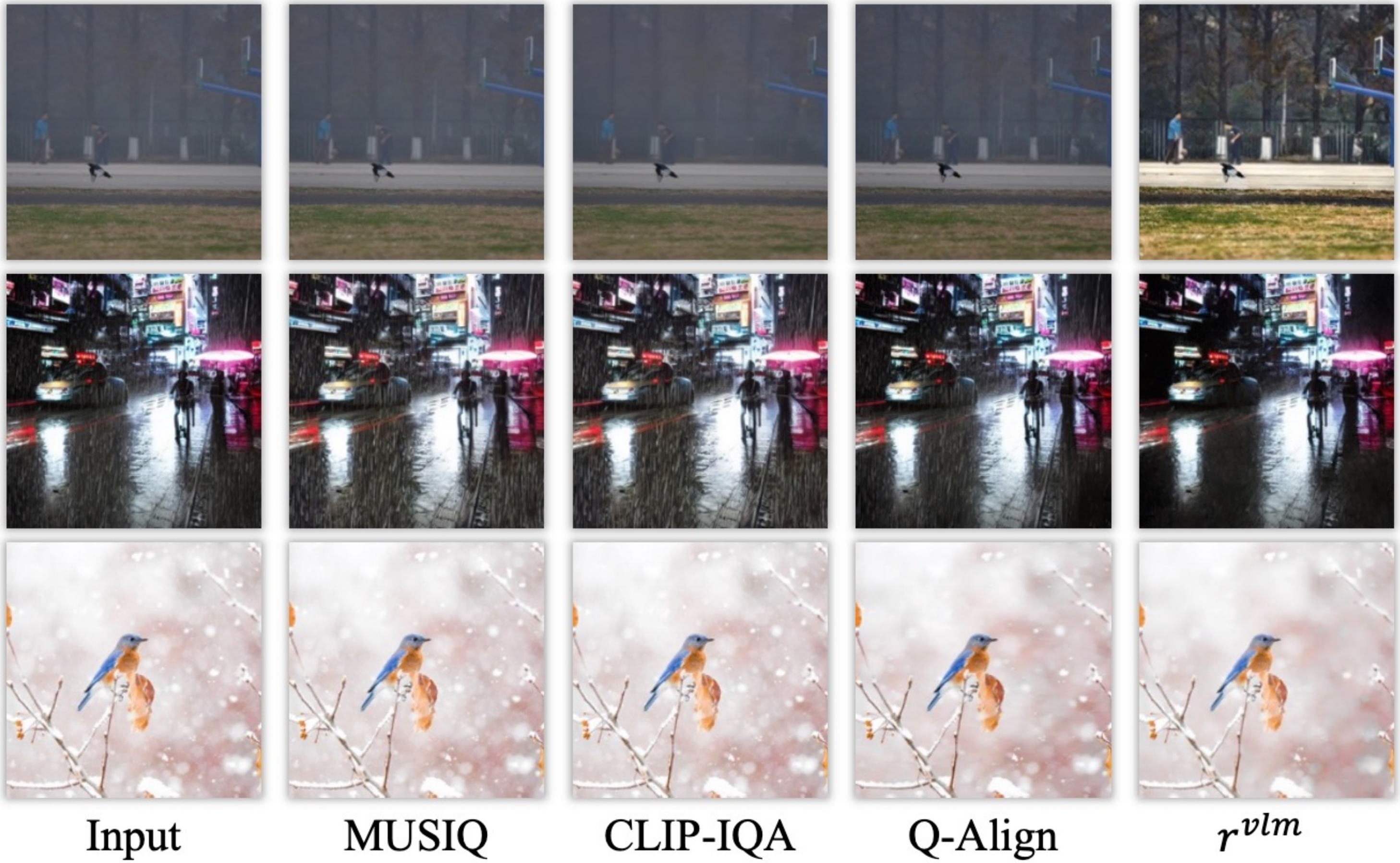
softmax

token	excellent	good	fair	poor	bad
prob	3.1e-3	0.069	0.93	1.9e-3	6.3e-4

fair

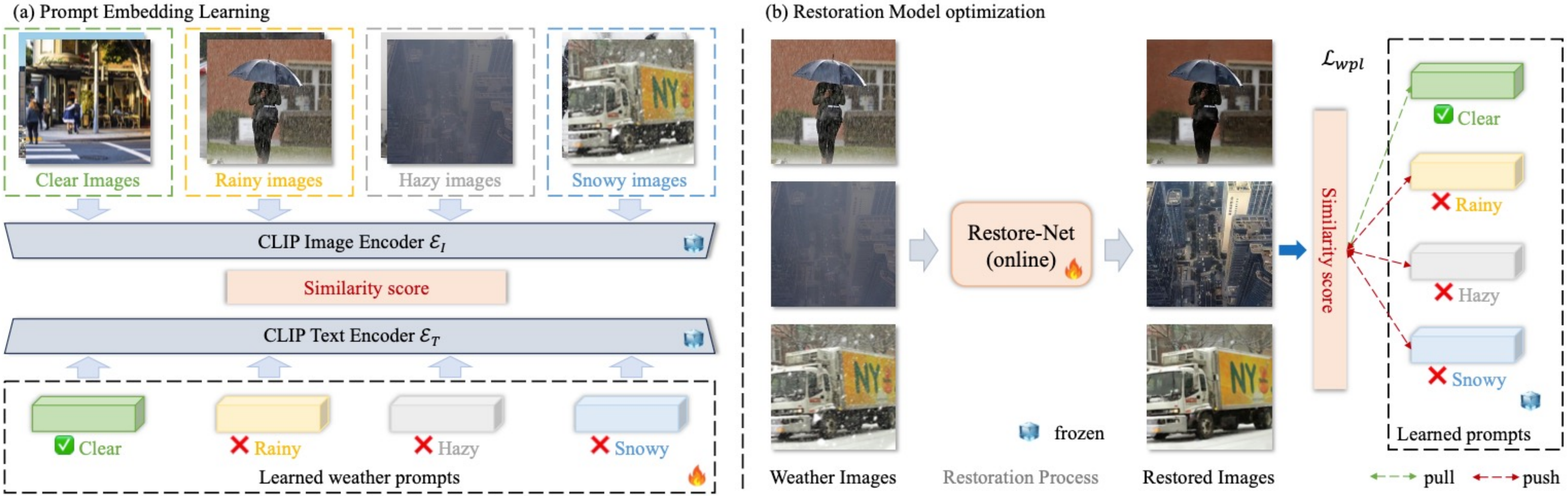
Pseudo-Labeling





$$\mathcal{L}_{ps} = \mathcal{L}_{app}(\hat{y}_i, y_i^{ps})$$

Weather Prompt Learning



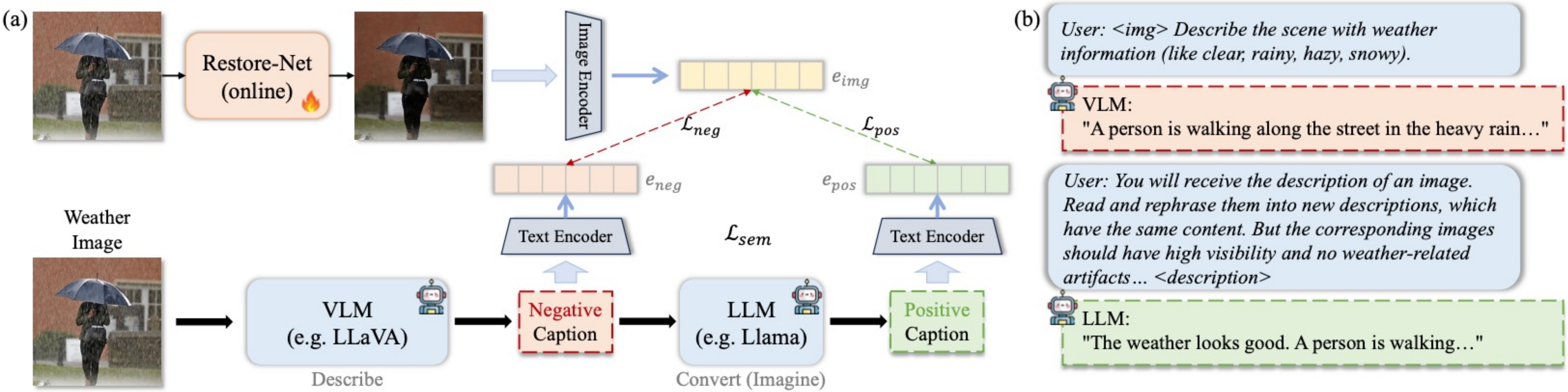
$$\mathcal{L}_{wpl} = \frac{e^{cos(\mathcal{E}_I(\hat{y}),\mathcal{E}_T(t_c))}}{\sum_{t\in\{t_c,t_r,t_h,t_s\}} e^{cos(\mathcal{E}_I(\hat{y}),\mathcal{E}_T(t))}} .$$

Feature similarity loss

$$\mathcal{L}_{feat} = \frac{1}{HW}\sum_{i=1}^{HW}\left(1 - cos(\hat{g}_i, g_i^*)\right)$$

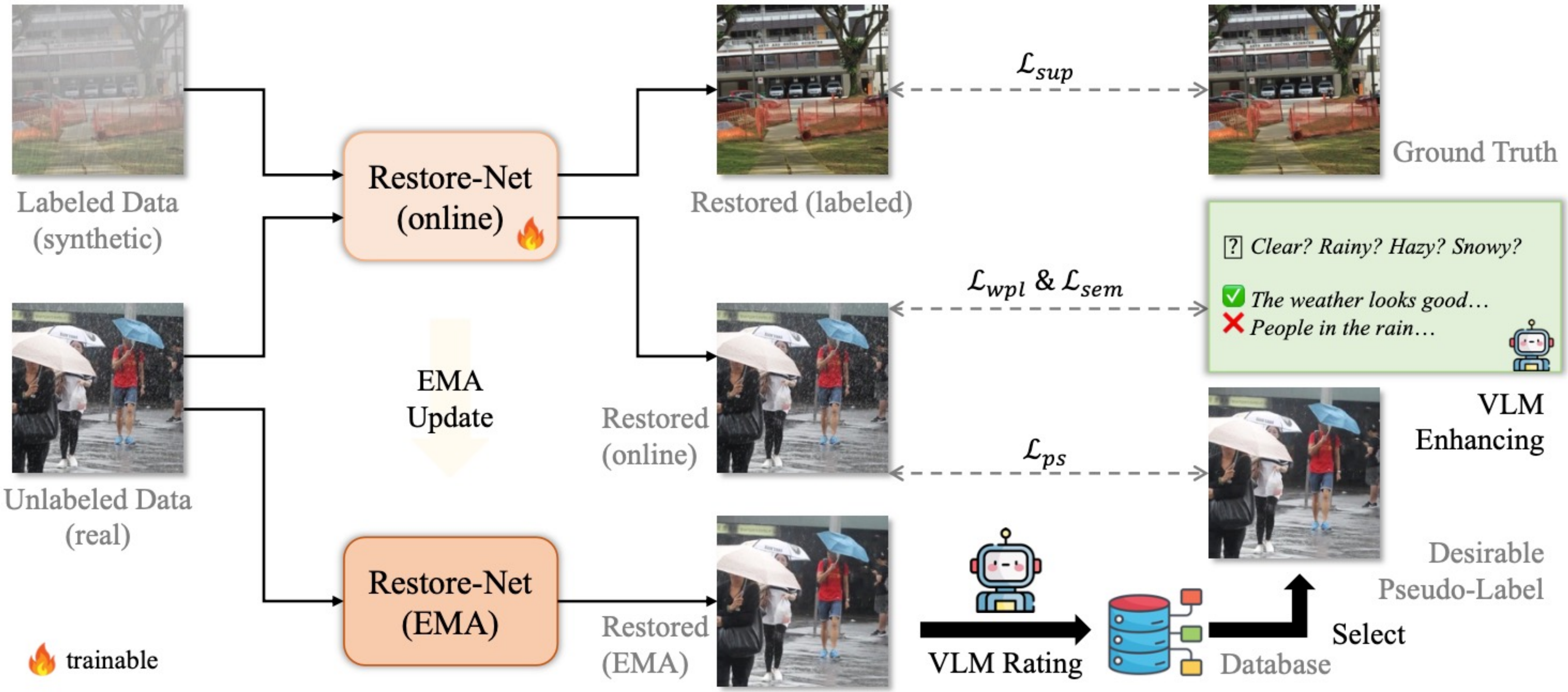
- the resulting image exhibits noticeable noise
- align the model’s prediction with both the pseudo-label and the input
- adopt the visual encoder of Depth Anything for feature extraction

Description-assisted semantic enhancement



$$\mathcal{L}_{sem} = \frac{e^{\cos(\mathcal{E}_I(\hat{y}), \mathcal{E}_T(d_{pos}))}}{\sum_{d \in \{d_{pos}, d_{neg}\}} e^{\cos(\mathcal{E}_I(\hat{y}), \mathcal{E}_T(d))}}$$

Total loss

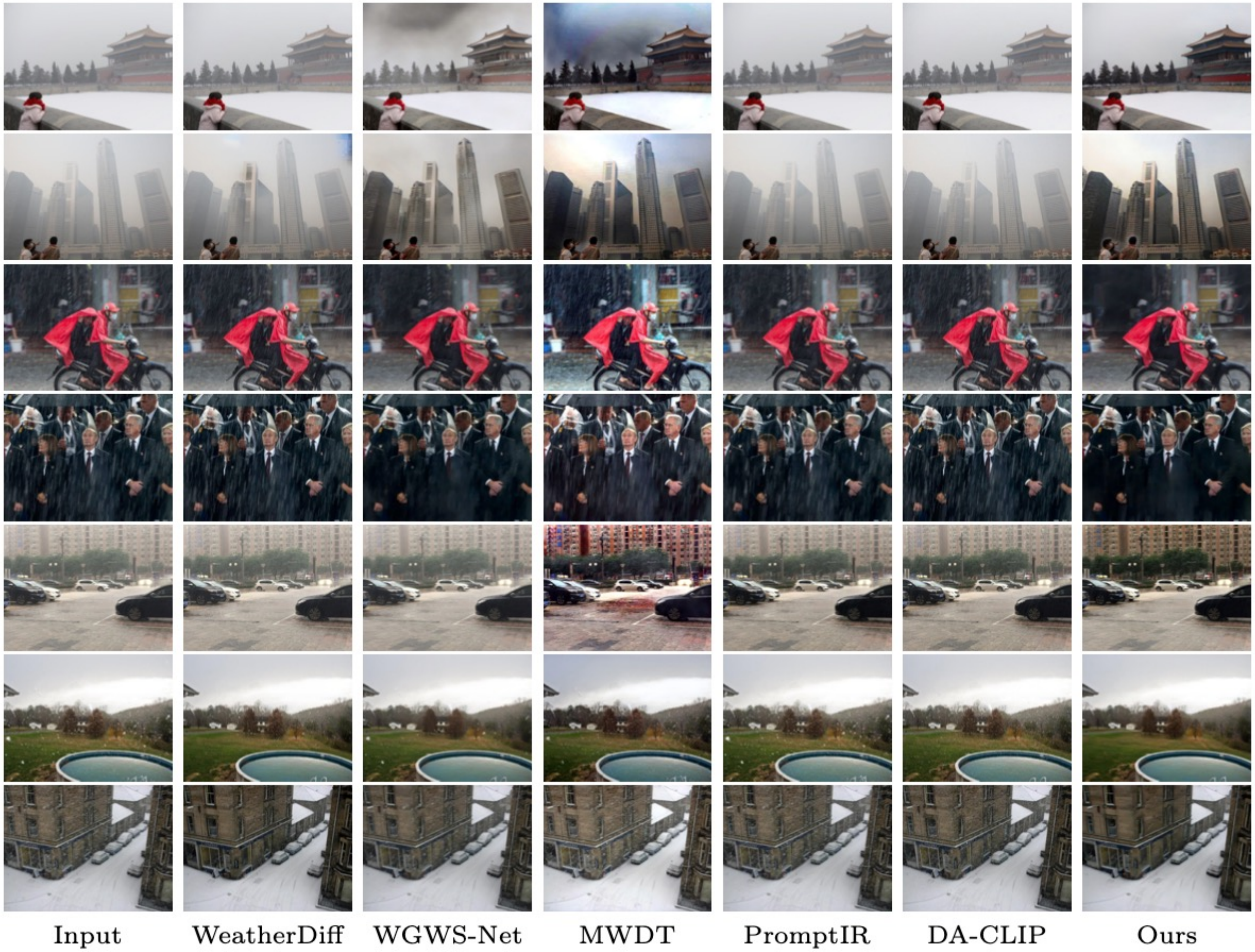


$$\mathcal{L} = \mathcal{L}_{sup} + w_1 \times \mathcal{L}_{ps} + w_2 \times \mathcal{L}_{wpl} + w_3 \times \mathcal{L}_{sem} + w_4 \times \mathcal{L}_{feat}$$

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Results

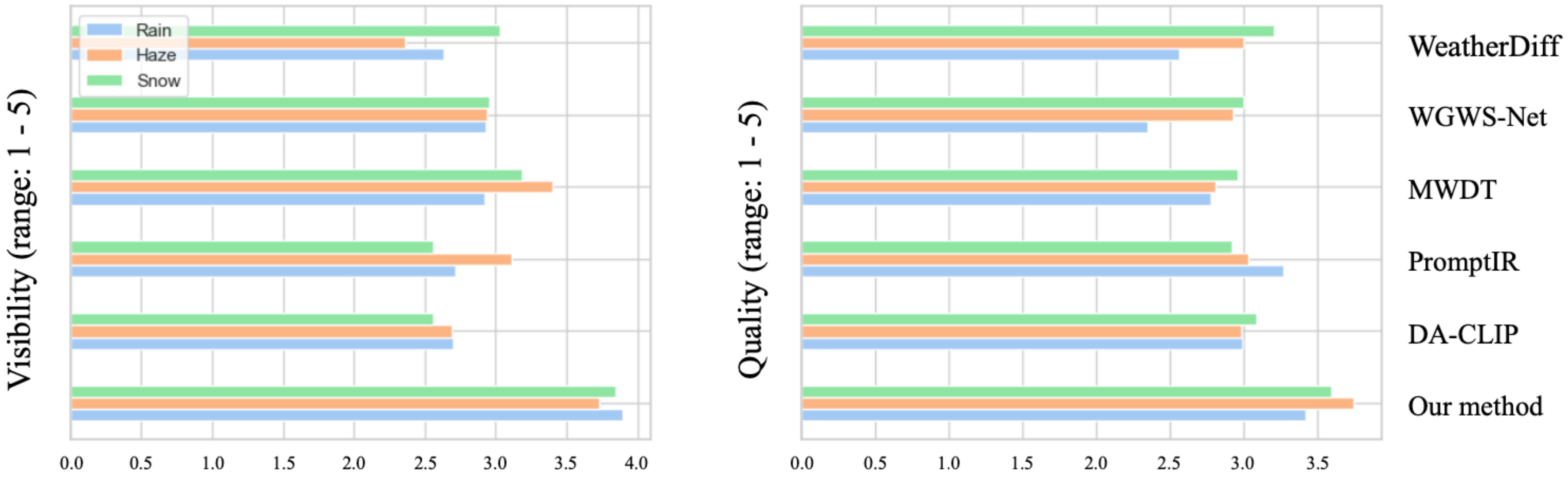


Results

Method	NIMA [36] ↑ / MUSIQ [14] ↑ / CLIP-IQA [41] ↑			
	Rain	Haze	Snow	Overall
Restormer [54]	5.151 / 54.69 / 0.437	4.804 / 53.27 / 0.366	5.020 / 61.18 / 0.510	4.992 / 56.38 / 0.438
TransWeather [40]	5.068 / 51.06 / 0.358	4.716 / 46.27 / 0.292	4.928 / 59.38 / 0.416	4.904 / 52.24 / 0.355
TKL [4]	5.099 / 50.96 / 0.392	4.697 / 48.21 / 0.318	4.905 / 59.24 / 0.428	4.900 / 52.80 / 0.379
WeatherDiff [27]	5.054 / 51.82 / 0.395	4.616 / 47.70 / 0.326	4.917 / 60.52 / 0.466	4.862 / 53.35 / 0.396
WGWS-Net [59]	5.035 / 51.46 / 0.389	4.815 / 45.76 / 0.310	4.779 / 57.95 / 0.395	4.876 / 51.72 / 0.365
MWDT [28]	5.104 / 52.47 / 0.377	4.741 / 51.23 / 0.315	5.034 / 60.16 / 0.407	4.960 / 54.62 / 0.366
PromptIR [29]	5.174 / 53.48 / 0.439	4.823 / 53.88 / **0.372**	5.032 / 60.86 / 0.517	5.009 / 56.07 / 0.443
DA-CLIP [25]	5.168 / 52.98 / 0.412	4.851 / 53.23 / 0.325	5.012 / 60.57 / 0.499	5.010 / 55.59 / 0.412
Our method	**5.291** / **59.80** / **0.477**	**4.906** / **56.09** / 0.371	**5.057** / **62.12** / **0.519**	**5.084** / **59.34** / **0.456**

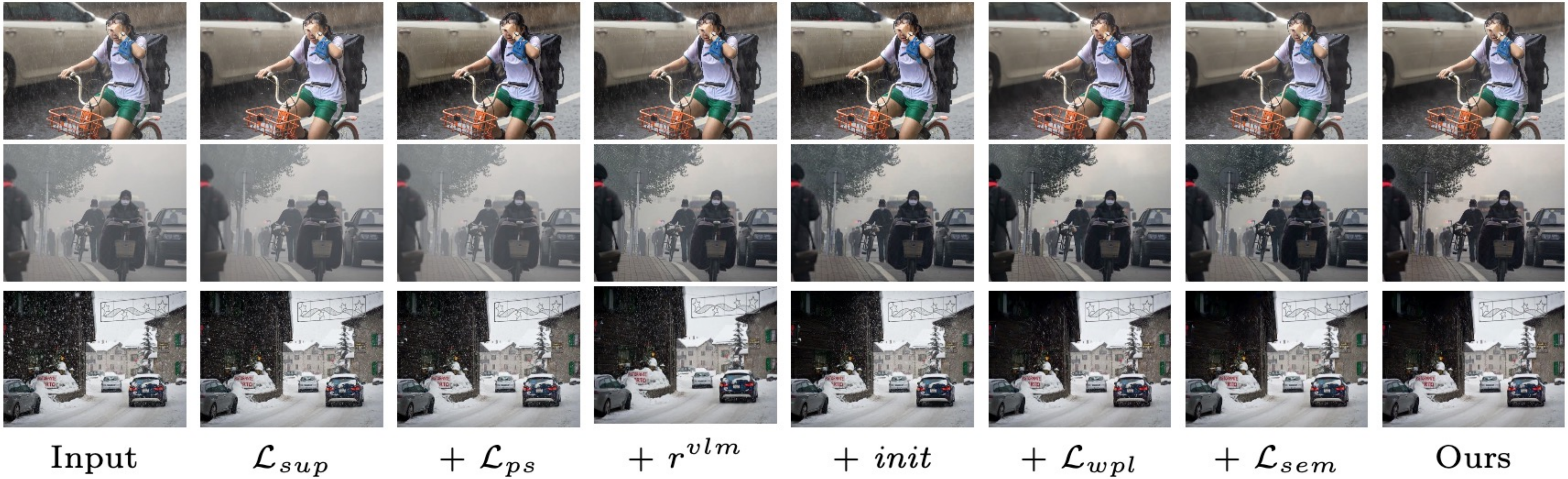
Method	LIQE [56] / Q-Align [47] ↑ / VLM-Vis ↑			
	Rain	Haze	Snow	Overall
Restormer [54]	2.277 / 3.795 / 0.417	1.918 / 3.068 / 0.218	3.172 / 3.646 / 0.395	2.456 / 3.503 / 0.343
TransWeather [40]	1.924 / 3.545 / 0.402	1.502 / 2.809 / 0.223	2.770 / 3.537 / 0.384	2.065 / 3.297 / 0.336
TKL [4]	2.028 / 3.588 / 0.406	1.590 / 2.908 / 0.238	2.830 / 3.557 / 0.393	2.149 / 3.351 / 0.346
WeatherDiff [27]	2.050 / 3.640 / 0.411	1.520 / 2.843 / 0.217	2.950 / 3.573 / 0.397	2.173 / 3.352 / 0.342
WGWS-Net [59]	1.965 / 3.592 / 0.411	1.506 / 2.915 / 0.238	2.619 / 3.490 / 0.383	2.030 / 3.332 / 0.344
MWDT [28]	2.068 / 3.548 / 0.426	1.720 / 2.861 / 0.273	2.903 / 3.569 / 0.412	2.230 / 3.326 / 0.370
PromptIR [29]	2.250 / 3.770 / 0.419	1.941 / 3.093 / 0.226	3.121 / 3.609 / 0.384	2.437 / 3.491 / 0.343
DA-CLIP [25]	2.250 / 3.732 / 0.412	2.014 / 3.071 / 0.230	3.050 / 3.637 / 0.395	2.438 / 3.480 / 0.346
Our method	**2.563** / **3.843** / **0.440**	**2.064** / **3.176** / **0.289**	**3.293** / **3.702** / **0.431**	**2.640** / **3.574** / **0.387**

User study



Rain
Haze
Snow
Visibility (range: 1 - 5)
Quality (range: 1 - 5)
WeatherDiff
WGWS-Net
MWDT
PromptIR
DA-CLIP
Our method
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5

Ablation Study



Ablation Study

$\mathcal{L}_{sup}$	$\mathcal{L}_{ps}$	r^{vlm}	$init$	$\mathcal{L}_{wpl}$	$\mathcal{L}_{sem}$	$iter$	MUSIQ ↑	CLIP-IQA ↑	VLM-Vis ↑
✓							53.41	0.388	0.343
✓	✓						54.08	0.396	0.354
✓	✓	✓					56.68	0.429	0.366
✓	✓	✓	✓				57.34	0.425	0.370
✓	✓	✓	✓	✓			58.13	0.437	0.376
✓	✓	✓	✓	✓	✓		58.91	0.445	0.381
✓	✓	✓	✓	✓	✓	✓	**59.34**	**0.456**	**0.387**

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Conclusion

- By evaluating clearness and semantics in natural images, our semi-supervised approach trains models on real, unlabeled images using vision-language models.

- Dual-step strategy, combining **image assessment** and **weather prompt learning**, enhances clearness with real data. Further, **semantics enhancement** adjusts weather conditions in vision-language model descriptions, addressing context semantics in adverse weather.

- Experimental results show that this method outperforms state of the arts. Yet, the computational burden of using large VLMs remains a limitation.