ICLR
International Conference On Learning Representations

RAPPER: Reinforced Rationale-Prompted Paradigm for Natural Language Explanation in Visual Question Answering

Kai-Po Chang[1], Chi-Pin Haung[1], Wei-Yuan Cheng[1], Fu-En Yang[1,2], Chien-Yi Wang[2], Yung-Hsuan Lai[1], Yu-Chiang Frank Wang[1,2]

[1]National Taiwan University [2]NVIDIA

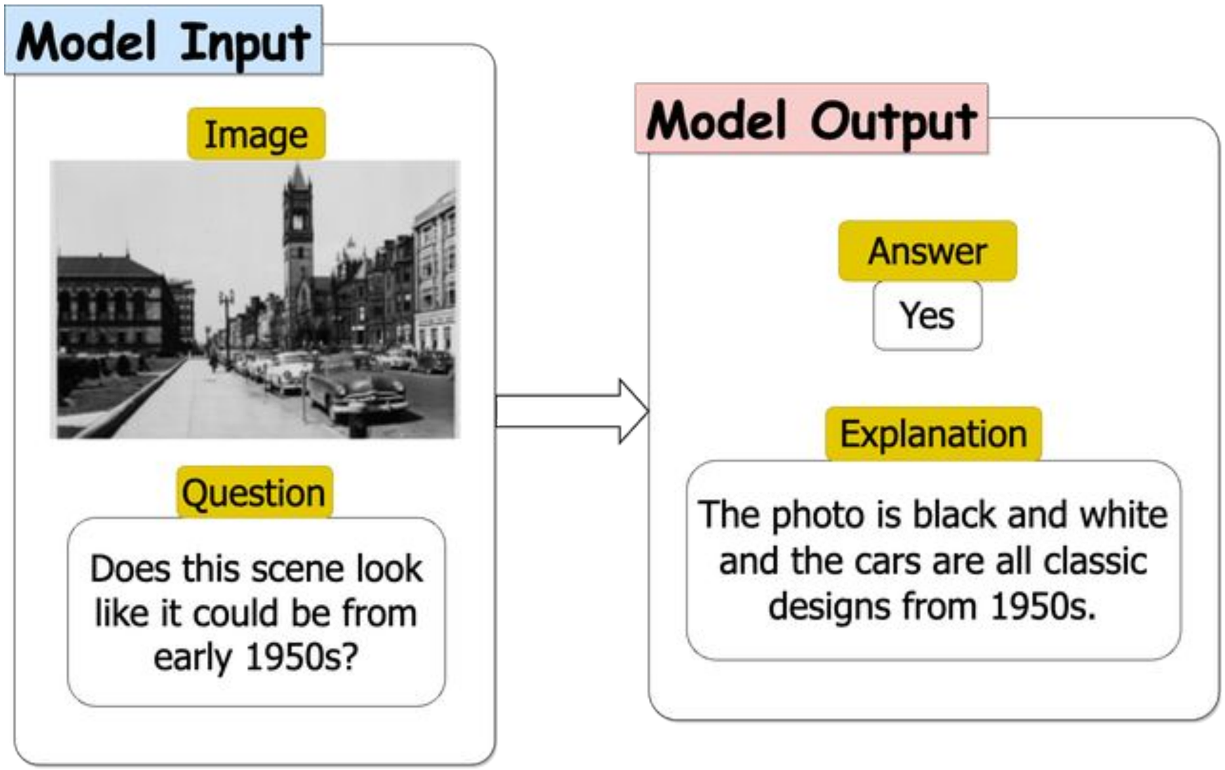
ICLR 2024

國立臺灣大學 National Taiwan University

NVIDIA

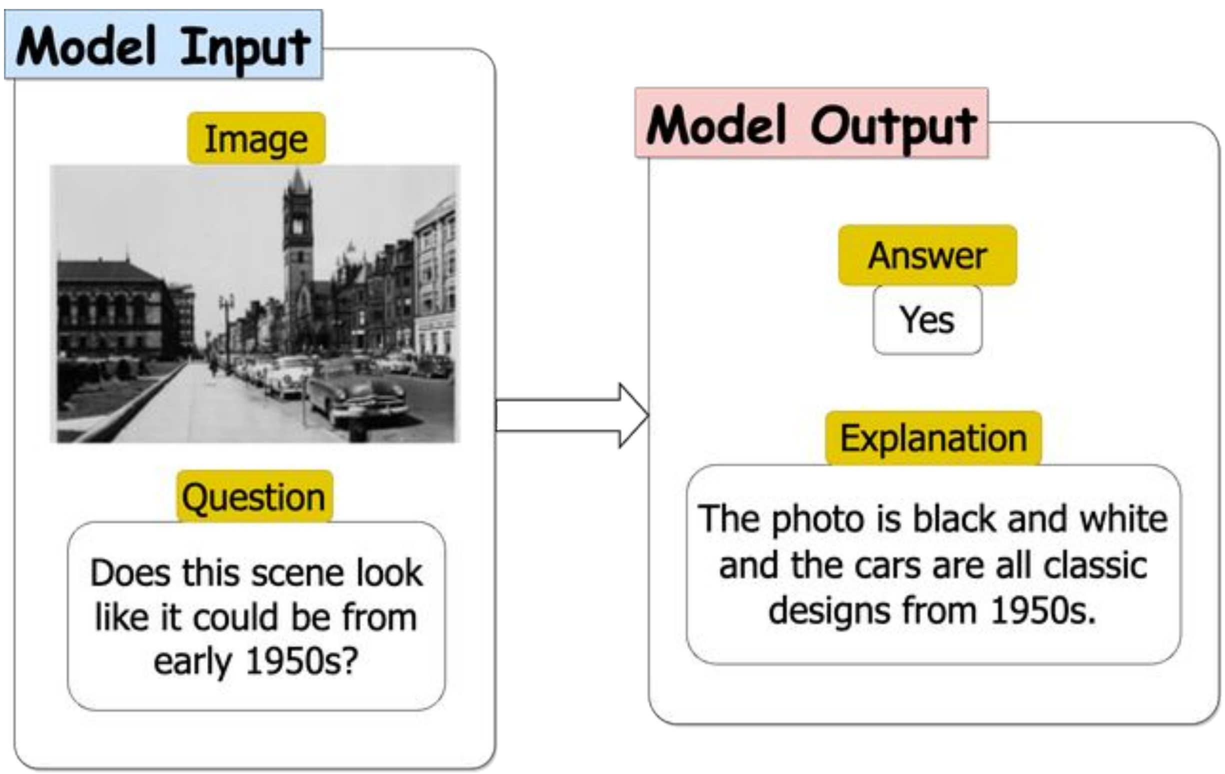
Visual Question Answering with Natural Language Explanation (VQA-NLE) (1/2)

- **Goal:**
  Besides generating answer, vision-language models are required to provide natural language explanations (NLE) that represent their reasoning process.



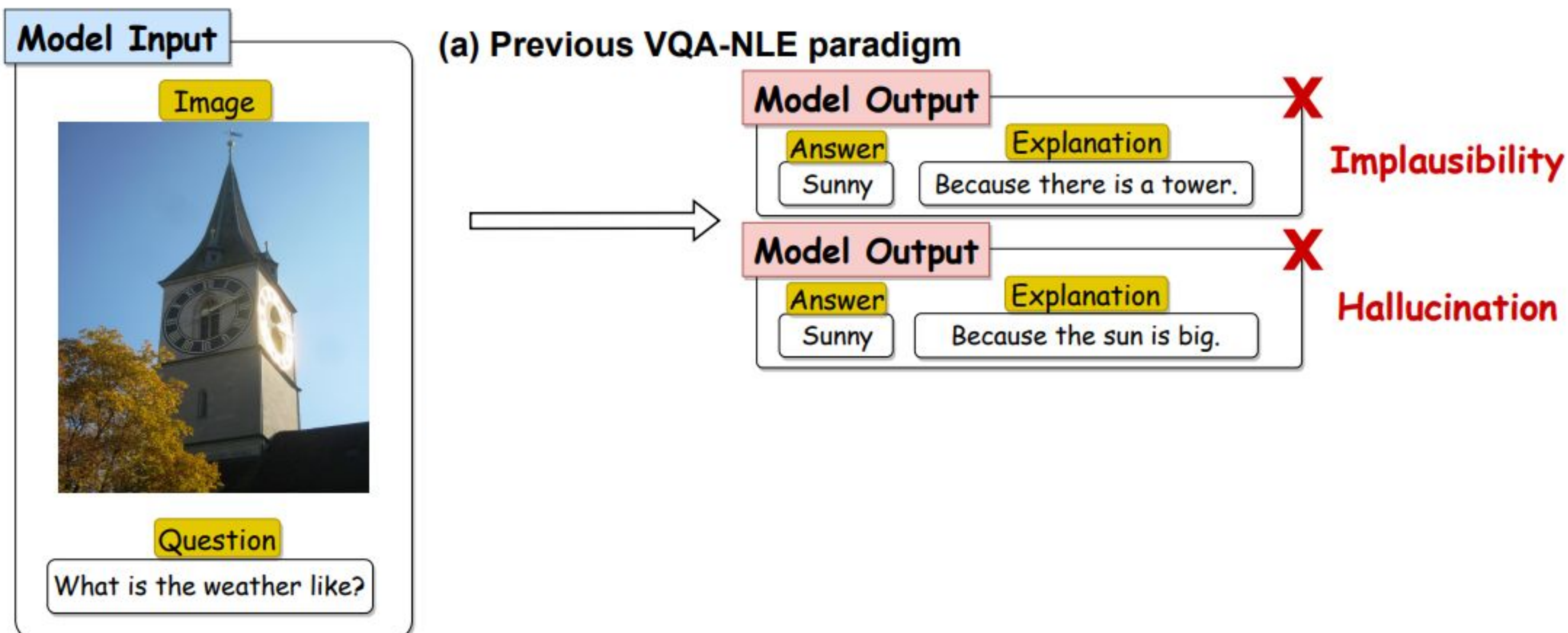
Visual Question Answering with Natural Language Explanation (VQA-NLE) (2/2)

- **Goal:**
  Besides generating answer, vision-language models are required to provide natural language explanations (NLE) that represent their reasoning process.
- **Challenges:**
  The NLE from **VLMs** are often *implausible* and *hallucinated* (next slide).



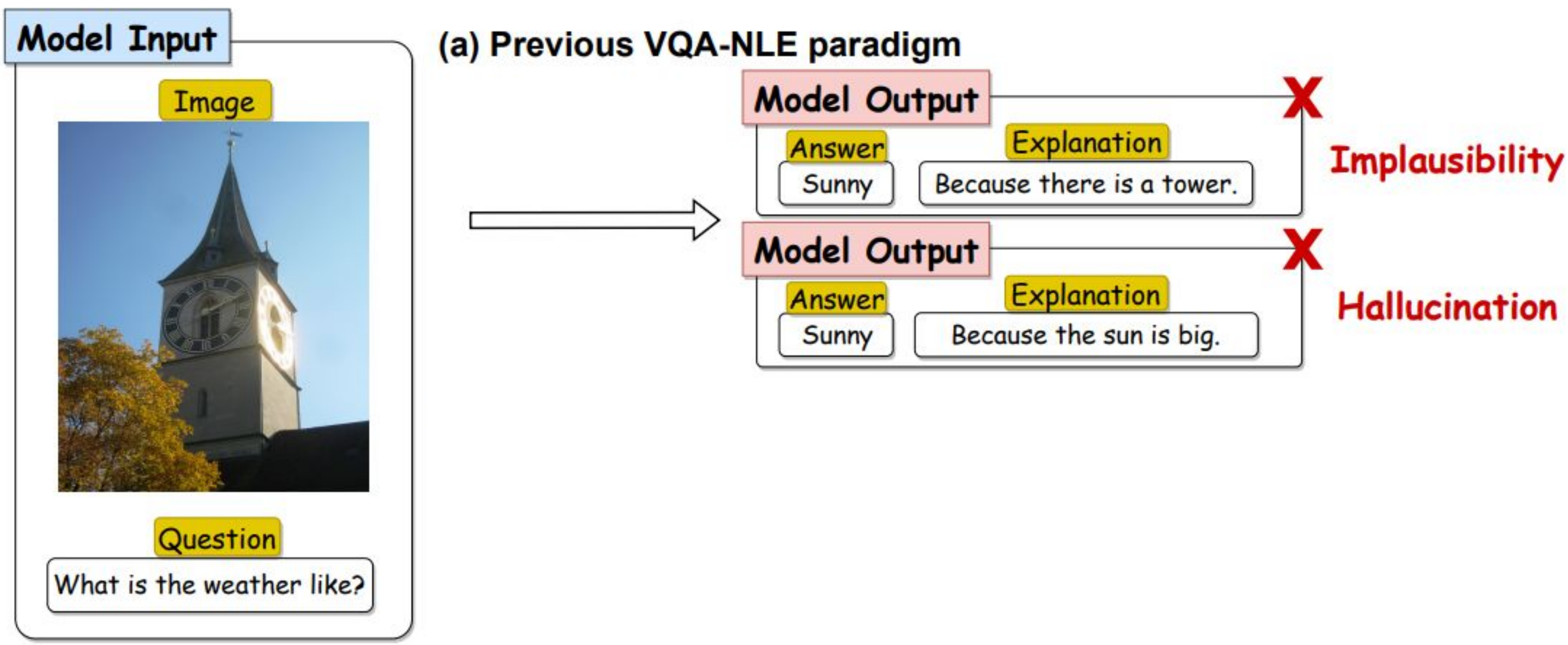
Motivation - challenge (1/3)

- **Implausibility**: NLEs are not relevant to the **question**.
  i.e., the building in image (tower) is not related to the weather.
- **Hallucination**: NLEs are not related to the **image**.
  i.e., the big sun cannot be observed from image.



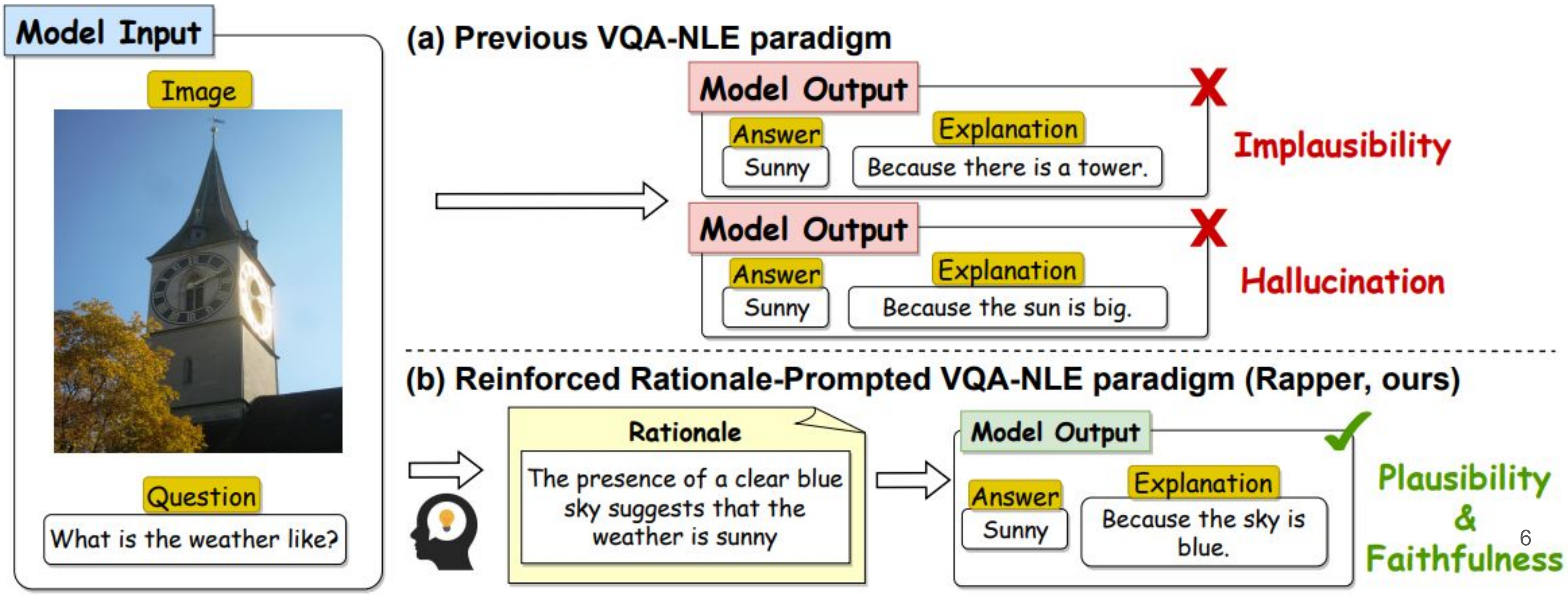
Motivation - challenge (2/3)

- **Implausibility**: NLEs are not relevant to the **question**.
  This happens when VLMs lack the knowledge required to answer this question.
- **Hallucination**: NLEs are not related to the **image**.
  This happens when VLMs explain w/ lang-based fact instead of image understanding.



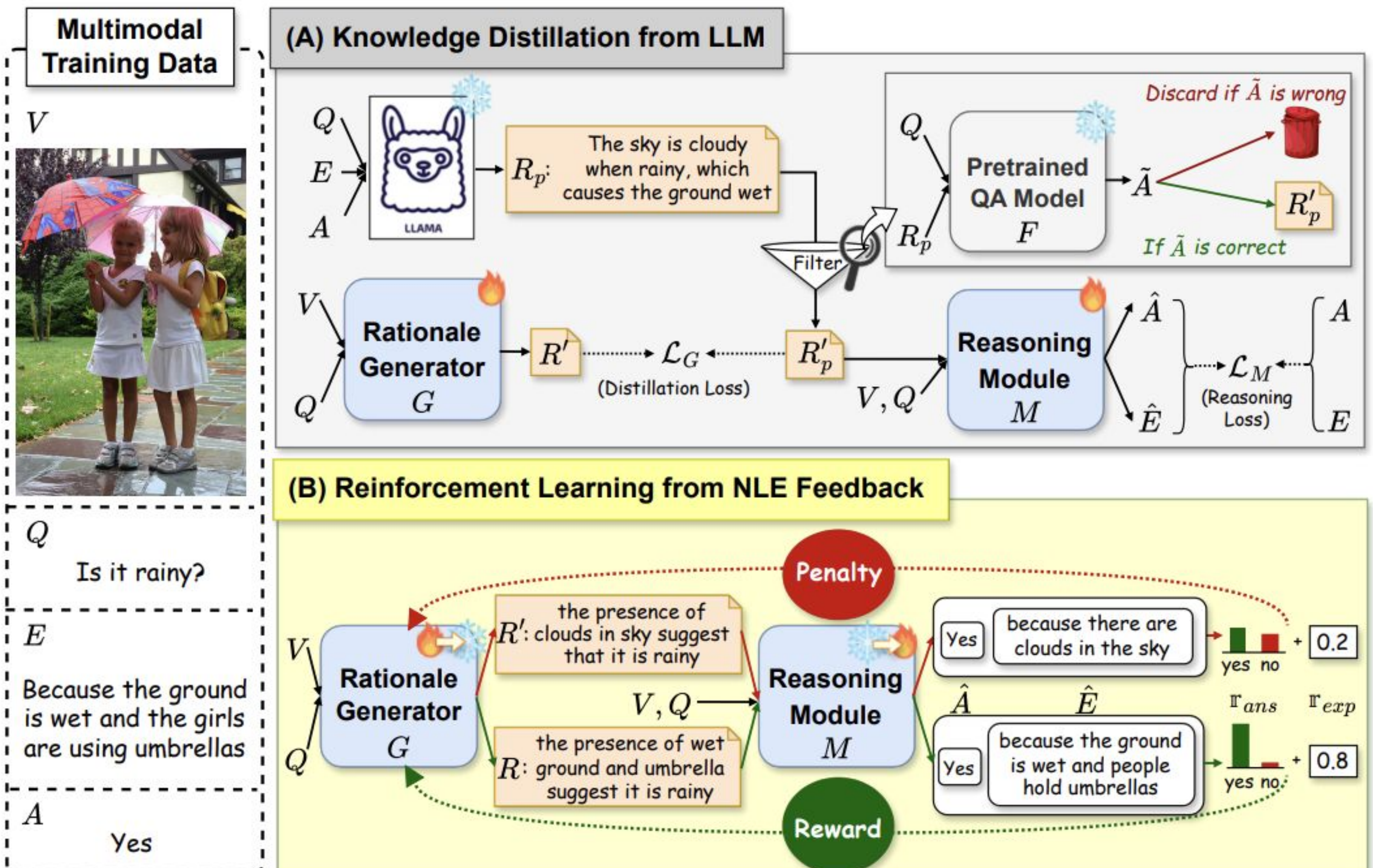
Motivation - solution (3/3)

- **Plausiblity (no implausiblity)**:
  ➔ Exploit the the knowledge inside LLMs.
- **Faithfulness (no hallucinaton)**:
  ➔ Learn to incorporate visual clues from input images.
- Achieved by using <u>rationale</u> as part of input prompt to VLMs.



Method - overview

- To generate plausible and faithful NLE, we learn the rationale through two stages:
  (A) Knowledge Distillation from LLM. ➡ exploiting the knowledge inside LLMs
  (B) Reinforcement Learning from NLE Feedback. ➡ incorporating visual clues from images

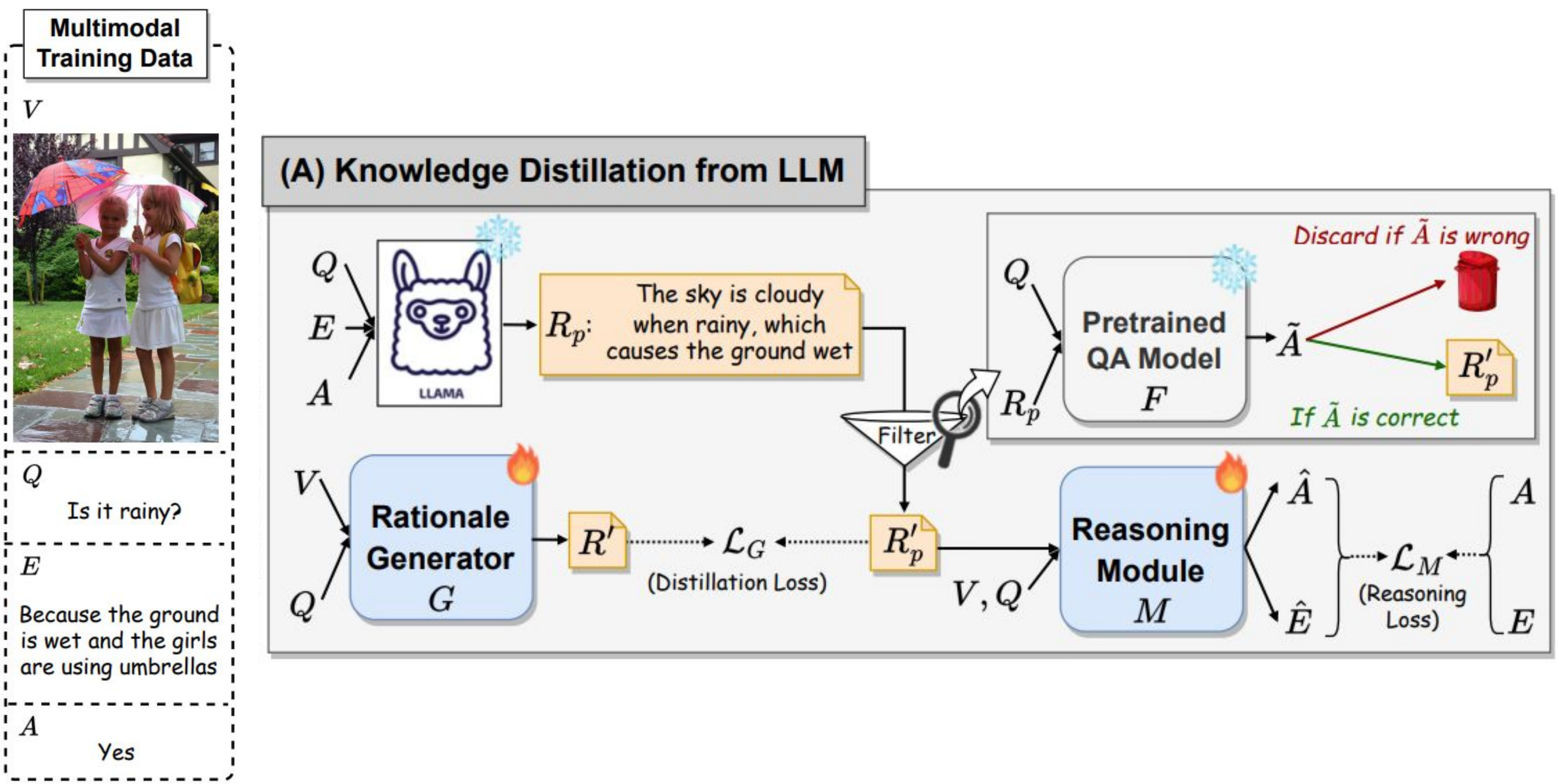


Method - Knowledge Distillation from LLM

- **Goal**: plausible NLE generation

  Step1. KD for fact-based rationale generation (train *G*)

  Step2. Prompting by fact-based rationale for plausible NLE (train *M*)

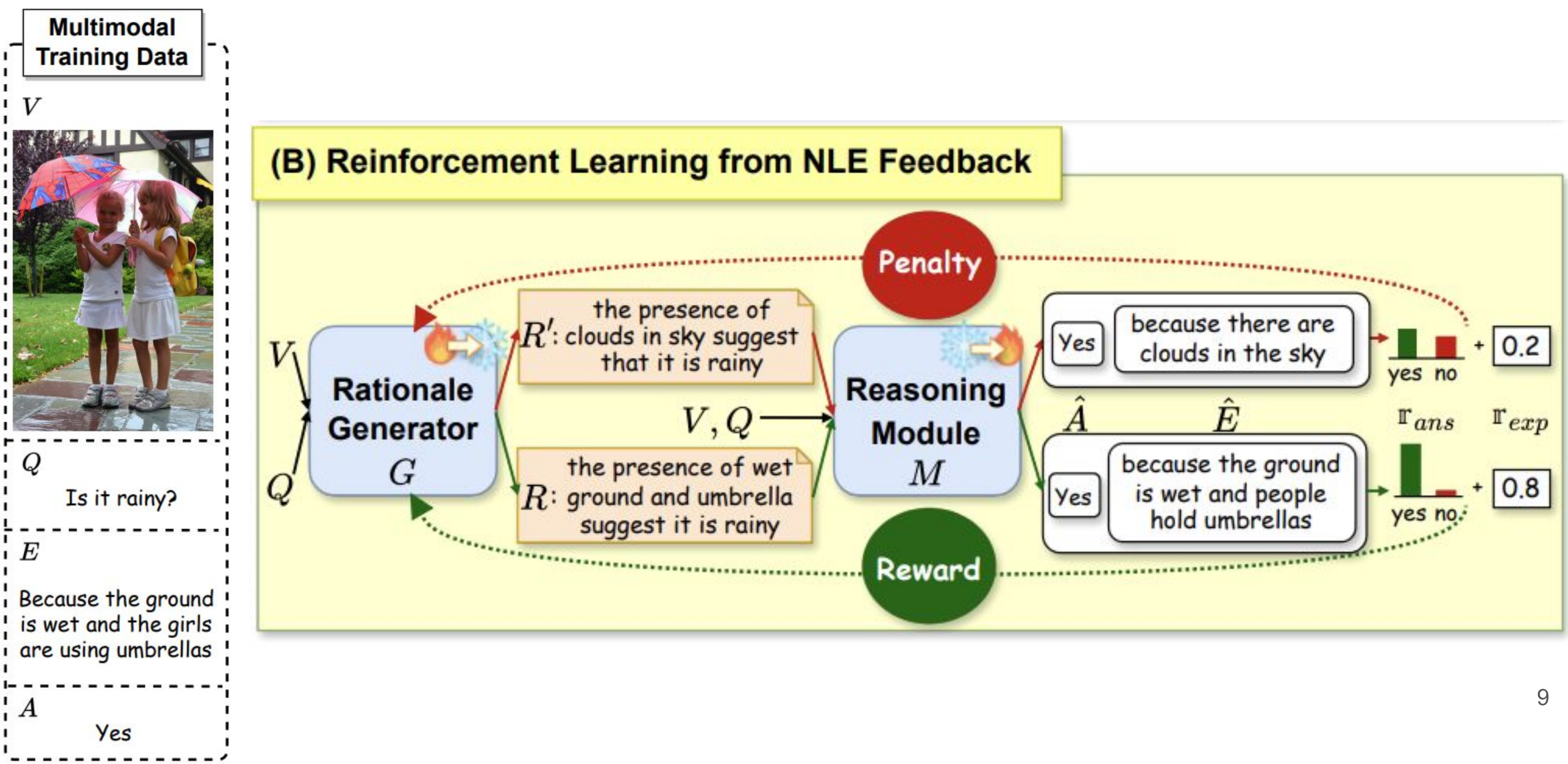


Method - Reinforcement Learning from NLE Feedback (RLNF) (1/3)

- **Goal**: faithful NLE generation

  Step1. RLNF for injecting visual facts into rationales (train *G*, freeze *M*)

  Step2. Prompting by visual-fact-based rationale for faithful NLE (train *M*, freeze *G*)

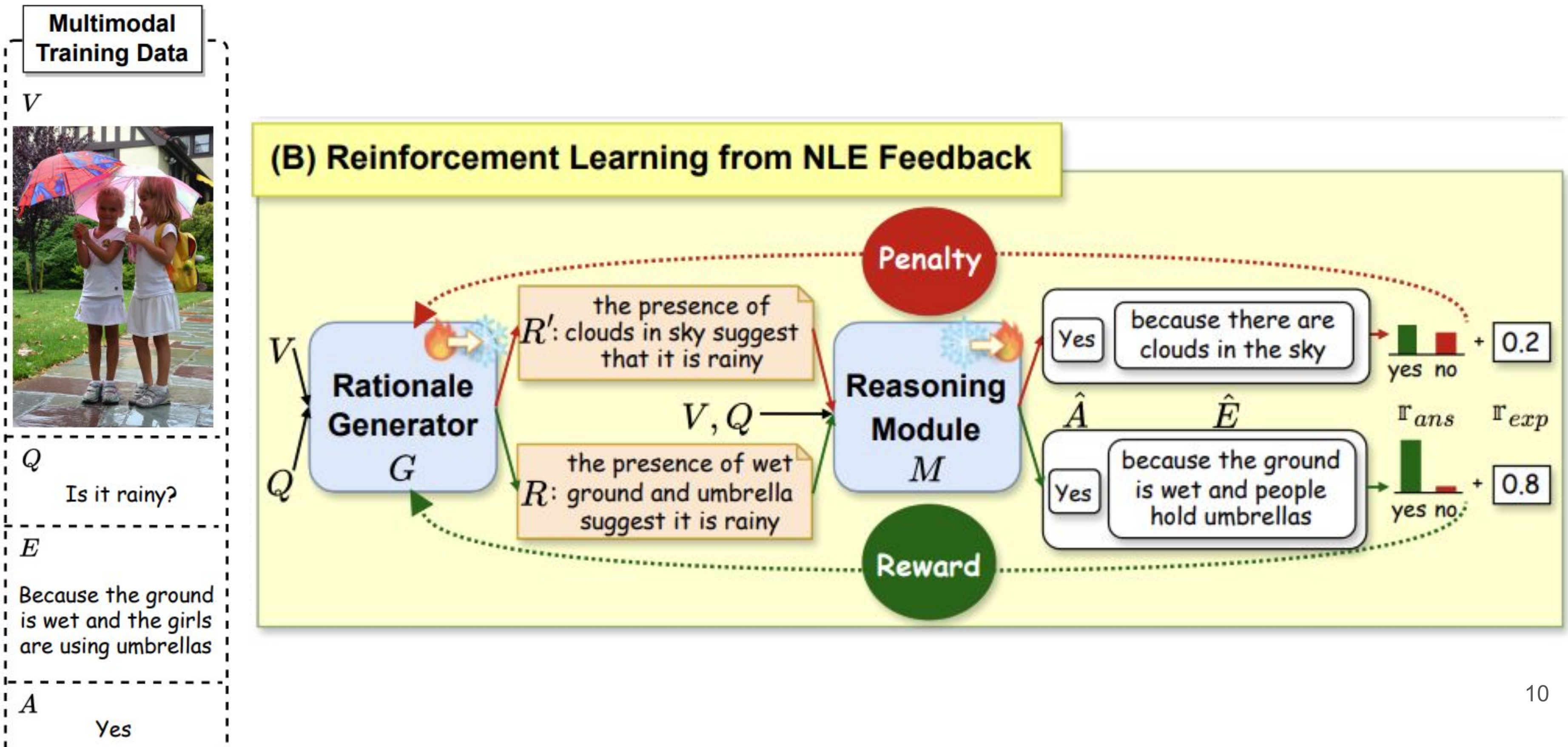


Method - Reinforcement Learning from NLE Feedback (RLNF) (2/3)

- **Goal**: faithful NLE generation

  Step1. RLNF enforces the derivation of visual facts from image to rationale (train *G*, freeze *M*)

  - This is achieved by penalizing the fact-based but hallicinated rationales (*R'*), while rewarding the rationales (*R*) that contain both established facts and visual content.
  - Reward = Prob(gt_ans) + CIDEr(gt_exp, pred_exp)



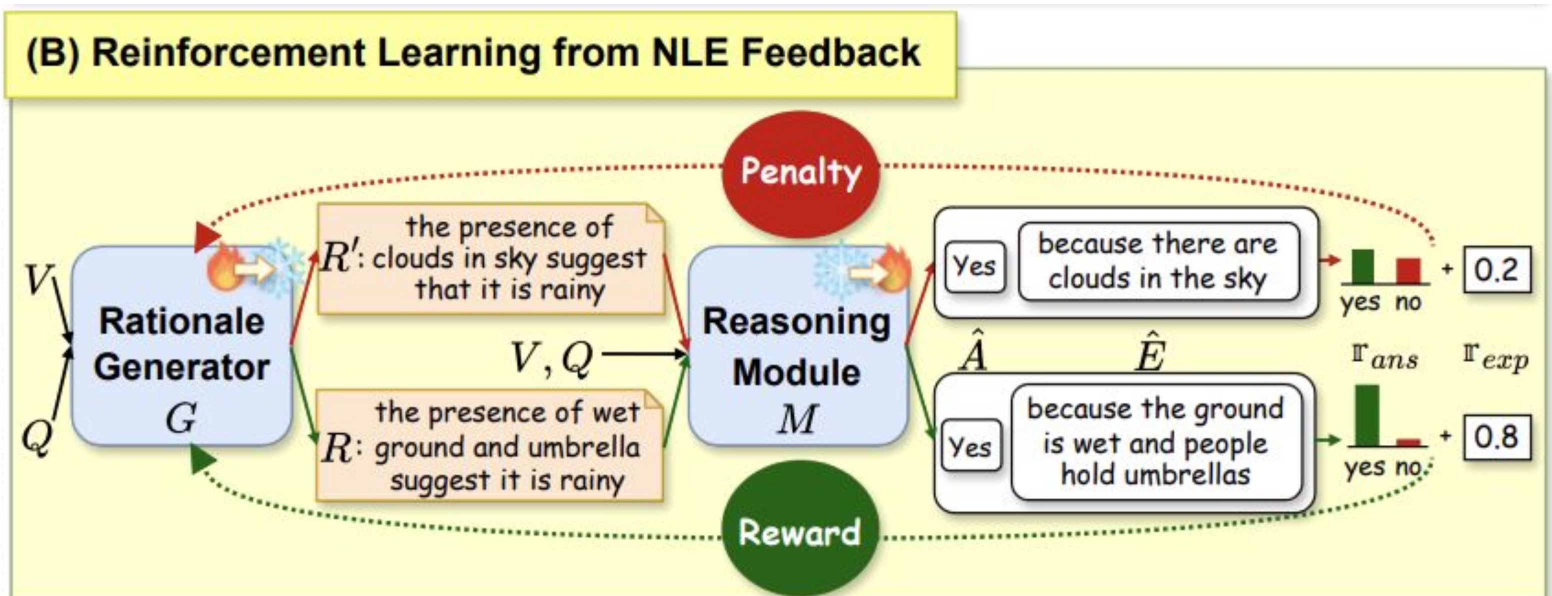
Method - Reinforcement Learning from NLE Feedback (RLNF) (3/3)

- **Goal**: faithful NLE generation

  Step1. RLNF for injecting visual facts into rationales (train *G*, freeze *M*)

  Step2. Prompting by visual-fact-based rationale for faithful NLE (train *M*, freeze *G*)

  - With the visual-fact-based rationales being part of its input prompts, NLEs from reasoning module (*M*) are retained with plausiblity and faithfualness.



Quantitative results (1/2)

- On two VQA-NLE benchmarks, Rapper achieves SOTA performances in terms of *all* natural language generation (NLG) metrics.
- CIDEr and SPICE are considered as metrics reflecting **plausibility** in NLE.

Method	VQA-X								
	B@1	B@2	B@3	B@4	METEOR	ROUGE-L	CIDEr	SPICE	Accuracy
PJ-X (Park et al., 2018)	57.4	42.4	30.9	22.7	19.7	46.0	82.7	17.1	76.4
FME (Wu & Mooney, 2018b)	59.1	43.4	31.7	23.1	20.4	47.1	87.0	18.4	75.5
RVT (Marasović et al., 2020)	51.9	37.0	25.6	17.4	19.2	42.1	52.5	15.8	68.6
QA-only (Kayser et al., 2021)	51.0	36.4	25.3	17.3	18.6	41.9	49.9	14.9	-
e-UG (Kayser et al., 2021)	57.3	42.7	31.4	23.2	22.1	45.7	74.1	20.1	80.5
NLX-GPT (Sammani et al., 2022)	64.2	49.5	37.6	28.5	23.1	51.5	110.6	22.1	83.07
S3C (Suo et al., 2023)	64.7	50.5	38.8	30.7	23.9	52.1	116.7	23.0	85.6
Rapper (ours)	**65.5**	**51.6**	**40.5**	**31.8**	**24.3**	**52.9**	**124.0**	**24.5**	**87.25**

Method	e-SNLI-VE								
	B@1	B@2	B@3	B@4	METEOR	ROUGE-L	CIDEr	SPICE	Accuracy
PJ-X (Park et al., 2018)	29.4	18.0	11.3	7.3	14.7	28.6	72.5	24.3	69.2
FME (Wu & Mooney, 2018b)	30.6	19.2	12.4	8.2	15.6	29.9	83.6	26.9	73.7
RVT (Marasović et al., 2020)	29.9	19.8	13.6	9.6	18.8	27.3	81.7	32.5	72.0
QA-only (Kayser et al., 2021)	29.8	19.7	13.5	9.5	18.7	27.0	80.4	32.1	-
e-UG (Kayser et al., 2021)	30.1	19.9	13.7	9.6	19.6	27.8	85.9	34.5	**79.5**
NLX-GPT (Sammani et al., 2022)	37.0	25.3	17.9	12.9	18.8	34.2	117.4	33.6	73.91
Rapper (ours)	**40.5**	**28.1**	**20.2**	**14.7**	**20.8**	**35.9**	**128.6**	**34.9**	75.73

Quantitative results (2/2)

- The upper part of this table demonstrates *Rapper* enhances the faithfulness in NLE compared to existing SOTA methods.
- The lower part of this table shows that RLNF increases the faithfulness in NLE.
- RefCLIPScore is a metric to reflect **faithfulness** in NLE.

Method	RefCLIPScore(↑)
Much recent VL-NLE works	
NLX-GPT	64.06
S3C	65.09
Our stage-ablated approaches	
Rapper (w/o KD and w/o RLNF)	66.00
Rapper (w/o RLNF)	65.66
Rapper	**67.05**

Qualitative results

- Blue: plausible and faithful NLE
- Orange: implausible NLE
- Red: hallucinated NLE

Multimodal Input / Methods		 **(a)** **Q**: Is the table cluttered? **GT A**: No **GT E**: There is only a single vase with flowers on it	 **(b)** **Q**: Is this in an asian country? **GT A**: Yes **GT E**: there is an asian language used as text font in public	 **(c)** **Q**: What kind of animal is this? **GT A**: Sheep **GT E**: The animal is covered in thick wool
NLX-GPT	$\hat{A}$	No	Yes	Sheep
	$\hat{E}$	There are no objects in the table	There is a train on the tracks	It has a long face and long nose
S^3C	$\hat{A}$	No	Yes	Sheep
	$\hat{E}$	There are only a few items on it	There is a train in the stations	It has a long snout and white fur
Rapper	R	The table is not cluttered because there is only one object on it	The presence of asian writing on the train suggests that it is in an asian country	A sheep is a type of animal that has wool on its body
	$\hat{A}$	No	Yes	Sheep
	$\hat{E}$	There is only one object on it	There is asian writing on the train	Its has wool on its body

Conclusion

- RAPPER enables VLMs generate NLEs with sufficient plausibility and faithfulness on VQA task.

- RAPPER composed of two-training stages:
  1. Knowledge Distillation from LLMs
  2. Reinforcement Learning From NLE Feedback (RLNF)

Thank you for listening!