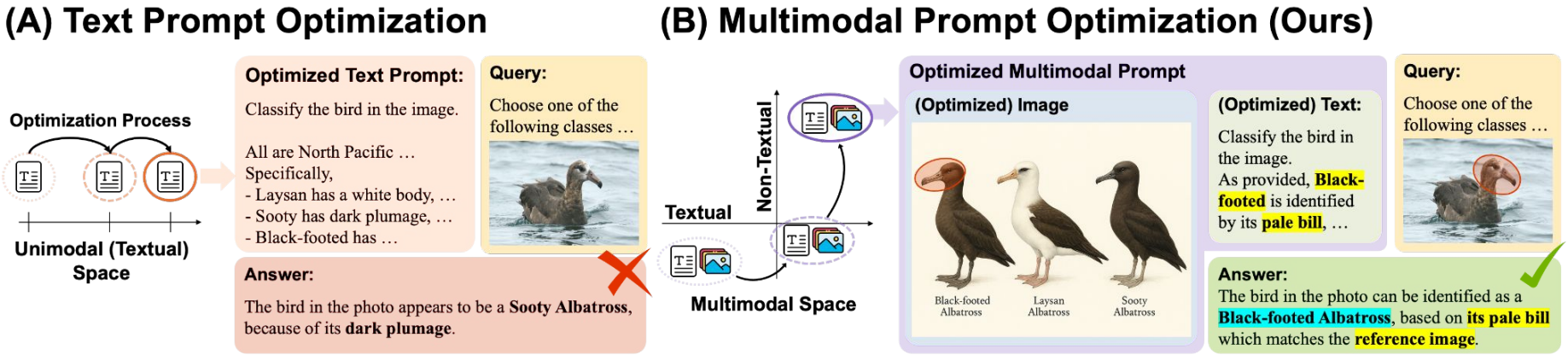
Multimodal Prompt Optimization:
Why Not Leverage Multiple Modalities for MLLMs

Yumin Choi[*1] **Dongki Kim**[*1] **Jinheon Baek**[†1] **Sung Ju Hwang**[†1,2]

[1]KAIST [2]DeepAuto.ai
{yuminchoi, cleverki, jinheon.baek, sungju.hwang}@kaist.ac.kr

Motivation: Beyond Text-Only Prompt Optimization

- Prior automatic prompt optimization has largely been developed for **text-only prompts**.
- However, modern **MLLMs can reason over images, videos, molecules**, and other modalities.
- In many tasks, critical context is **easier to convey through non-text inputs** than through language alone.
- As a result, optimizing only text fails to fully leverage the multimodal capabilities of MLLMs.



From text-only to multimodal prompt optimization

Problems and Challenges

Problem Definition

- We define a multimodal prompt as a pair of a textual prompt (t) and a non-textual prompt (m).
- The goal is to optimize both prompts jointly to maximize task performance.

$$(\boldsymbol{t}^*, \boldsymbol{m}^*) = \underset{(\boldsymbol{t},\boldsymbol{m})\in\mathcal{T}\times\mathcal{M}}{\operatorname{argmax}} \mathbb{E}_{(\boldsymbol{q},\boldsymbol{a})\sim\mathcal{D}}\Big[f\big(\texttt{MLLM}(\boldsymbol{t},\boldsymbol{m},\boldsymbol{q}),\boldsymbol{a}\big)\Big]$$

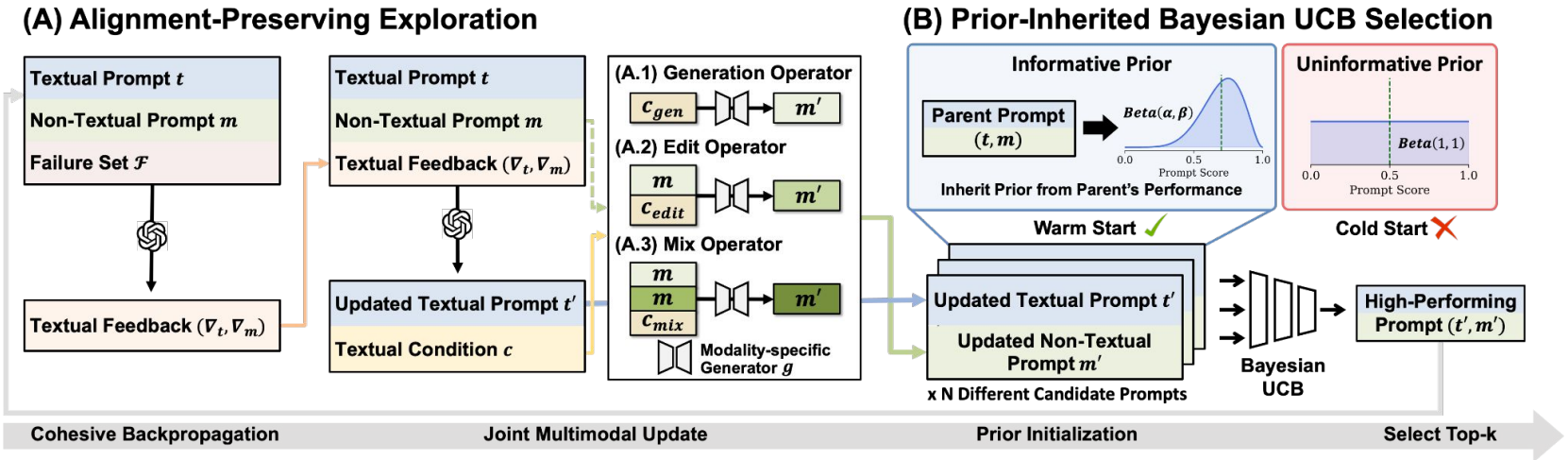
Challenge 1: Cross-modal alignment

- The textual and non-textual prompts must provide consistent and complementary information.
- If they are optimized independently, they can easily become semantically misaligned.

Challenge 2: Efficient search in a much larger space

- Joint optimization over multiple modalities greatly enlarges the prompt search space.
- High-performing prompts become sparser, making candidate evaluation more costly and less reliable.

Methodology: Multimodal Prompt Optimizer



Alignment-Preserving Exploration

- MPO first analyzes failure cases and produces a unified feedback signal for modality-specific generator.
- This shared feedback is used to update both the textual prompt and the non-textual prompt simultaneously.
- By coupling the updates, MPO preserves cross-modal consistency during optimization.c

Prior-Inherited Bayesian UCB Selection

- Evaluating many multimodal prompt candidates is expensive.
- MPO uses the parent prompt's performance as an informative prior for its children.
- This warm-started selection process allocates evaluation budget to more promising candidates.

Main Results: Multimodal Prompt Optimization Outperforms Text-Only Methods

Overall Performance

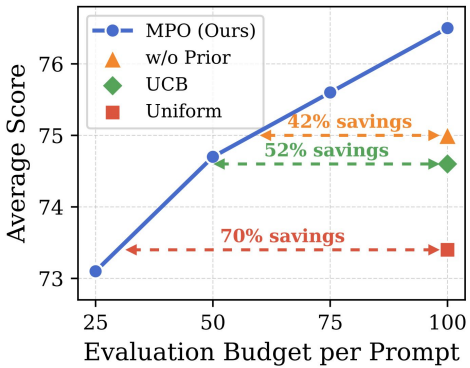
- MPO achieves the best average performance across **image**, **video**, and **molecular** benchmarks
- Representative improvements are observed on CUB (**76.4** vs. 69.0) and Drive&Act (**76.7** vs. 71.4).

Efficiency of Prior-Inherited Bayesian UCB Selection

- Prior-inherited selection substantially reduces search cost: 42% less budget than w/o Prior, 52% less than standard UCB, and the same performance as Uniform with only 30% of the budget.

	Image					Video		Molecule						
	PlantVillage*	CUB*	SLAKE*	DrivingVQA	RSVQA	Drive&Act	VANE.	Absorption*		BBBP		CYP Inhibit.*		
Methods	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	F1	Acc.	F1	Acc.	F1	**Avg.**
Human	42.2	47.9	35.2	49.7	51.0	47.3	47.0	38.5	36.3	39.4	38.6	43.1	37.1	44.1
CoT	43.1	49.0	30.8	52.9	49.6	37.2	31.6	39.6	36.7	33.6	32.5	40.1	32.3	40.8
1-Shot	39.7	54.7	31.4	54.5	48.5	50.4	62.4	37.8	35.7	36.1	34.8	56.2	48.3	47.2
3-Shot	48.2	58.8	30.6	53.9	52.2	54.2	56.0	46.1	44.2	42.7	42.6	51.9	47.3	49.5
5-Shot	46.5	58.1	28.0	45.9	49.2	54.3	61.4	48.1	45.5	49.3	49.3	52.0	47.0	49.3
APE	55.8	67.3	34.3	52.8	54.4	50.3	64.3	45.7	40.4	36.0	34.7	52.3	50.9	51.3
OPRO	54.1	59.7	33.9	52.7	51.0	46.4	51.0	37.6	35.4	39.2	38.3	43.0	37.1	46.9
EvoPrompt	56.1	59.6	34.8	52.9	50.5	46.7	56.5	48.2	46.5	38.7	37.7	51.1	49.7	49.5
PE2	67.9	71.6	35.8	53.7	55.2	50.8	63.0	64.5	56.8	61.3	58.2	58.5	55.1	58.2
ProTeGi	64.4	70.0	35.4	54.4	54.2	53.0	65.5	71.1	58.2	72.1	65.7	59.8	57.0	60.0
SEE	69.0	71.6	35.0	52.2	53.4	51.7	57.9	71.4	60.0	67.0	62.3	61.4	56.7	59.1
MPO (Ours)	**76.4**	**78.6**	**38.2**	**56.0**	**55.9**	**58.3**	**71.2**	**76.7**	**64.5**	**75.3**	**67.6**	**64.3**	**60.2**	**65.1**

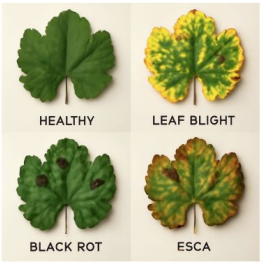
MPO outperforms text-only prompt optimization methods (e.g., ProTeGi, SEE).



Efficiency of Prior-Inherited Bayesian UCB

Why Does MPO Work?

- Text-only optimization improves instructions, but **remains limited to linguistic guidance**.
- In many tasks, critical evidence is easier to **convey through non-text modalities** than through text alone.
- Joint multimodal optimization provides both **clearer task instructions** and **richer task-specific evidence.**
- As a result, the textual and non-textual prompts reinforce each other and guide the model more effectively.



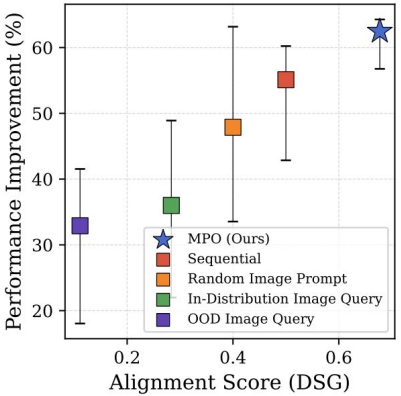
PlantVillage: Plant Leaf Images

Analyze the provided grape leaf image and classify it into one of the following categories: [‘Healthy’, ‘Leaf Blight’, ‘Black Rot’, ‘Esca’]. Use the hybrid reference image for guidance, focusing on the following critical visual features:

1. **Healthy**: Look for a vibrant, uniform green color and a smooth texture without blemishes.
2. **Leaf Blight**: Identify distinct yellowing edges along with well-defined small dark spots that are clearly visible.
3. **Black Rot**: Check for sharply defined, dark, sunken lesions that are prominent on the leaf surface, often accompanied by slight shriveling.
4. **Esca**: Look for distinct irregular brown patches, significant necrosis, and curling of the leaf edges.

In cases where symptoms overlap, prioritize the most severe characteristics. For example, if both dark spots and sunken lesions are present, classify based on the prominence of the lesions. Ensure that you assess each feature carefully, referencing the hybrid image to visualize these distinctions accurately.

Qualitative examples of the optimized multimodal (image and text) prompts.



Better cross-modal alignment leads to larger performance gains.

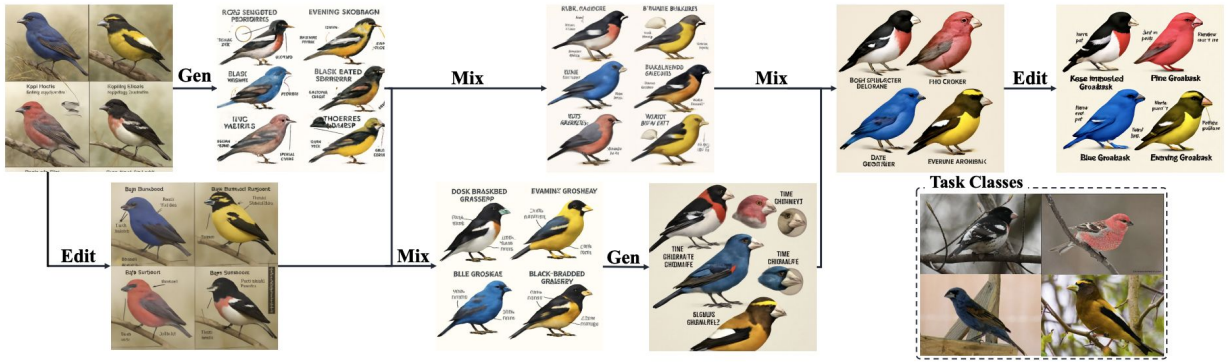
Conclusion

- We introduce the **new problem of multimodal prompt optimization**.

- MPO addresses it with alignment-preserving joint updates and prior-inherited Bayesian UCB selection.

- MPO consistently achieves better performance and greater search efficiency than text-only prompt optimization methods.

- Our main takeaway is that we should **extend the prompt optimization space beyond text** to fully leverage the capability of MLLMs.



Example of image prompt optimization process on the CUB dataset.

Multimodal Prompt Optimization:
Why Not Leverage Multiple Modalities for MLLMs

Thank You!

Yumin Choi[*1] **Dongki Kim**[*1] **Jinheon Baek**[†1] **Sung Ju Hwang**[†1,2]

[1]KAIST [2]DeepAuto.ai
{yuminchoi, cleverki, jinheon.baek, sungju.hwang}@kaist.ac.kr