國立陽明交通大學
NATIONAL YANG MING CHIAO TUNG UNIVERSITY

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Paper Tag: WED-PM-247

Multimodal Prompting with Missing Modalities for Visual Recognition

Yi-Lun Lee

Yi-Hsuan Tsai

Wei-Chen Chiu

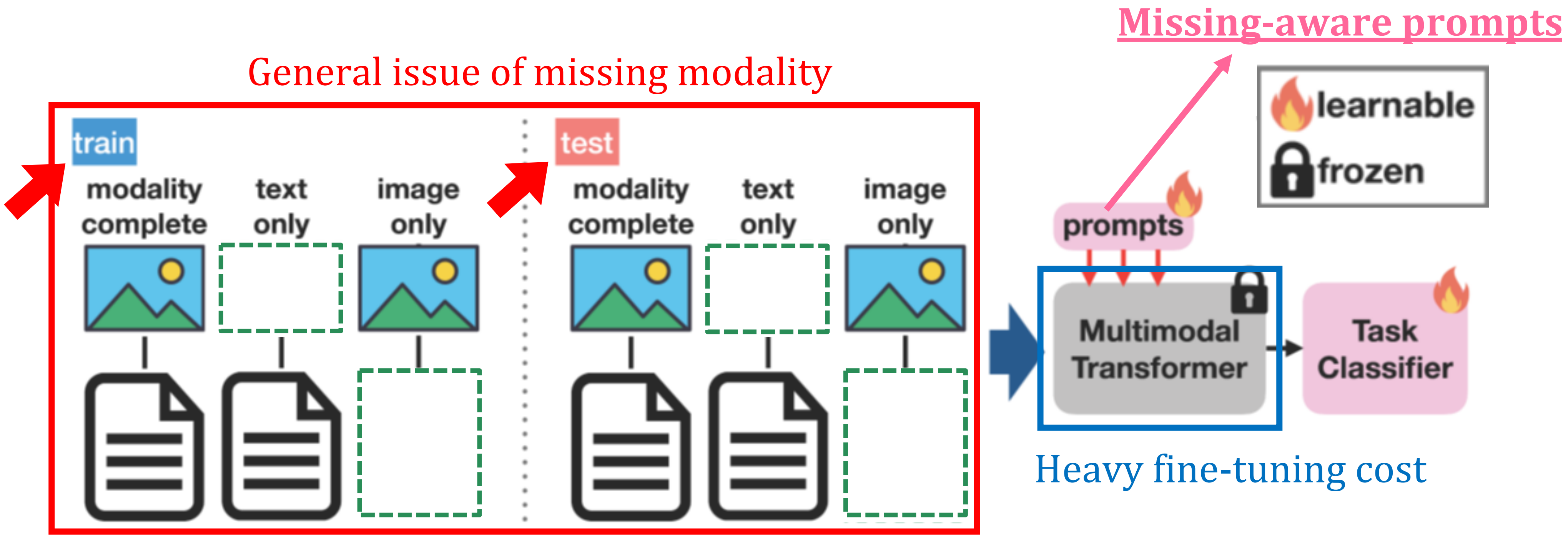
Chen-Yu Lee

Enriched Vision Applications Laboratory

Google

Goal

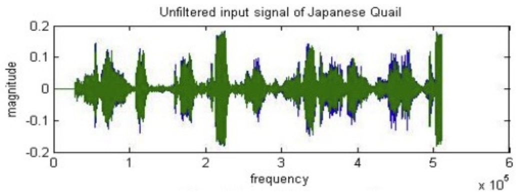
- A simple prompt-learning-based method for multimodal learning:
  - Tackle the general issue of missing modality
  - No need to finetune the heavy pre-trained model (transformer)



Multimodal Learning

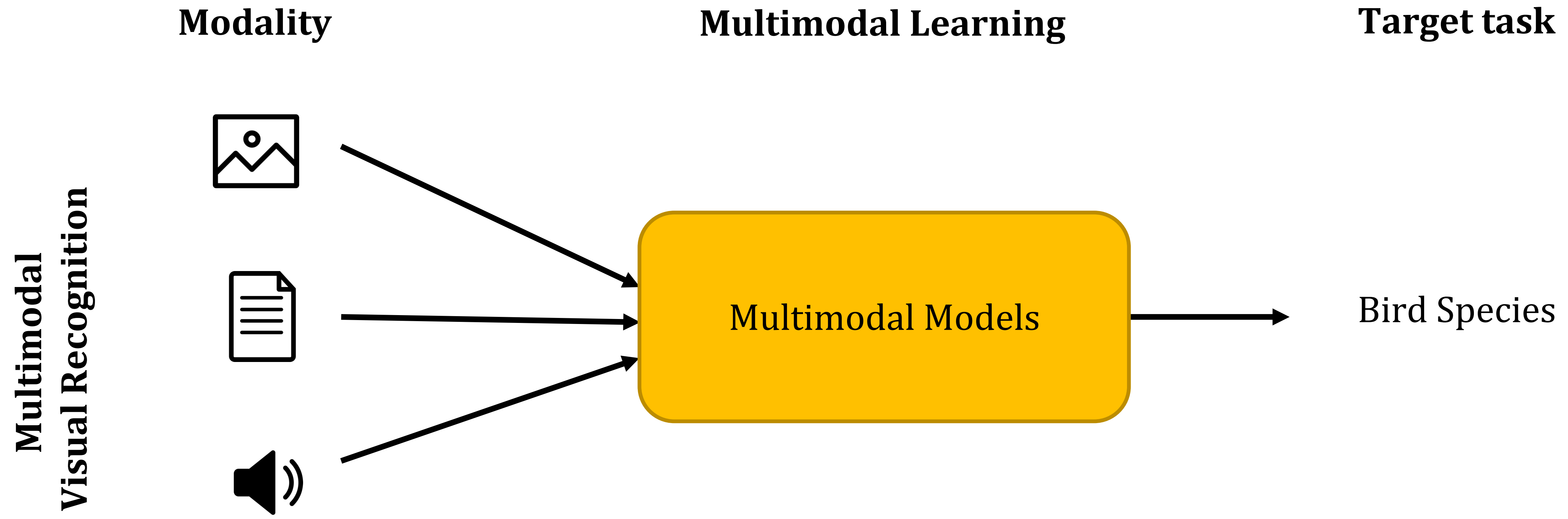
- Our observation perceived in daily life is typically multimodal

Multimodal Visual Recognition

Modality	Data sample	Properties	Target task
		Appearance	
	[From Wikipedia] The morphology of the Japanese quail differs depending on its stage in life. As chicks, both male and female individuals exhibit the same kind of plumage and coloring…	Description	Bird Species
		Sound	

Multimodal Learning

- *GOAL:* leverage the potential complementary properties among modalities to better realize the target tasks

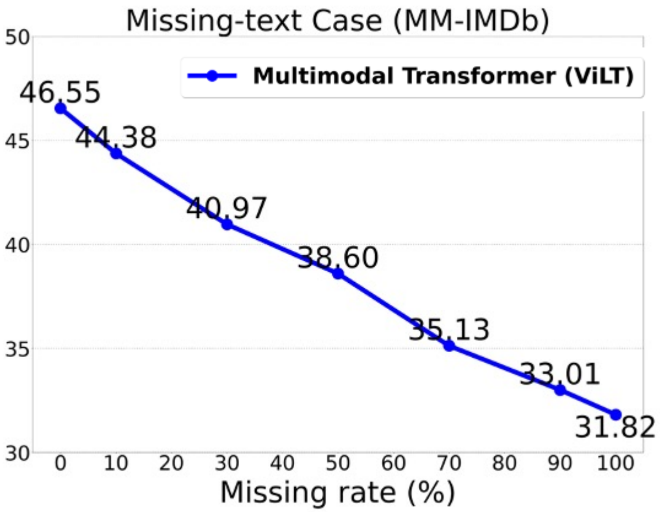


Challenges

- Some practical challenges for multimodal methods
  - Missing modality
  - Heavy cost of finetuning huge pre-trained models
  - Noisy web-crawled data (i.e. incomplete and incorrect)
  - Multimodal data perturbation (i.e. distribution shifts in real world)

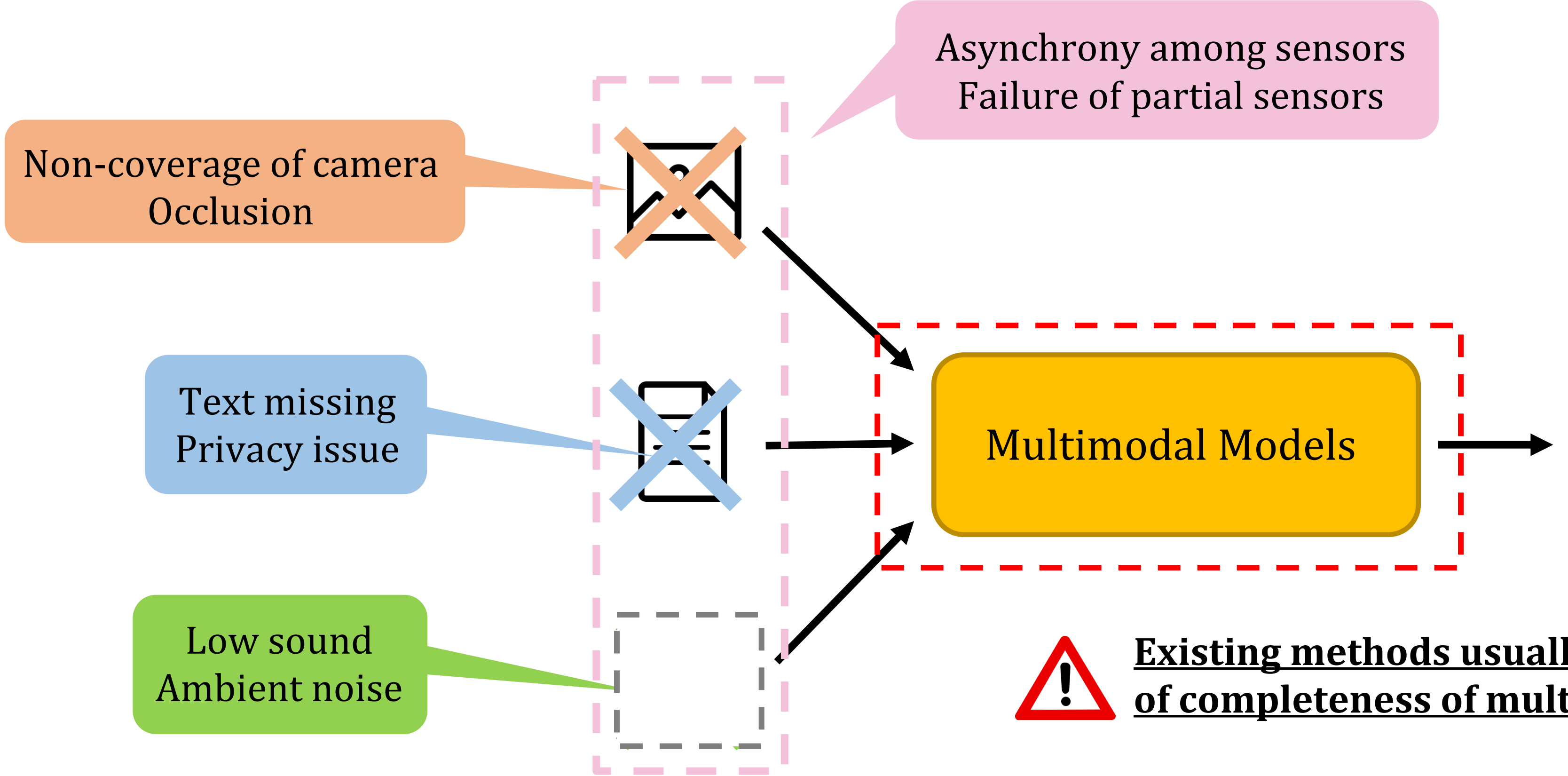
Challenge: Missing Modality

- Missing modality could happen for different reasons



Modality-specific reasons

General reasons



Performance drop

Fail to predict

Existing methods usually have the assumption of completeness of multi-modality.

Challenge: Heavy Training Cost

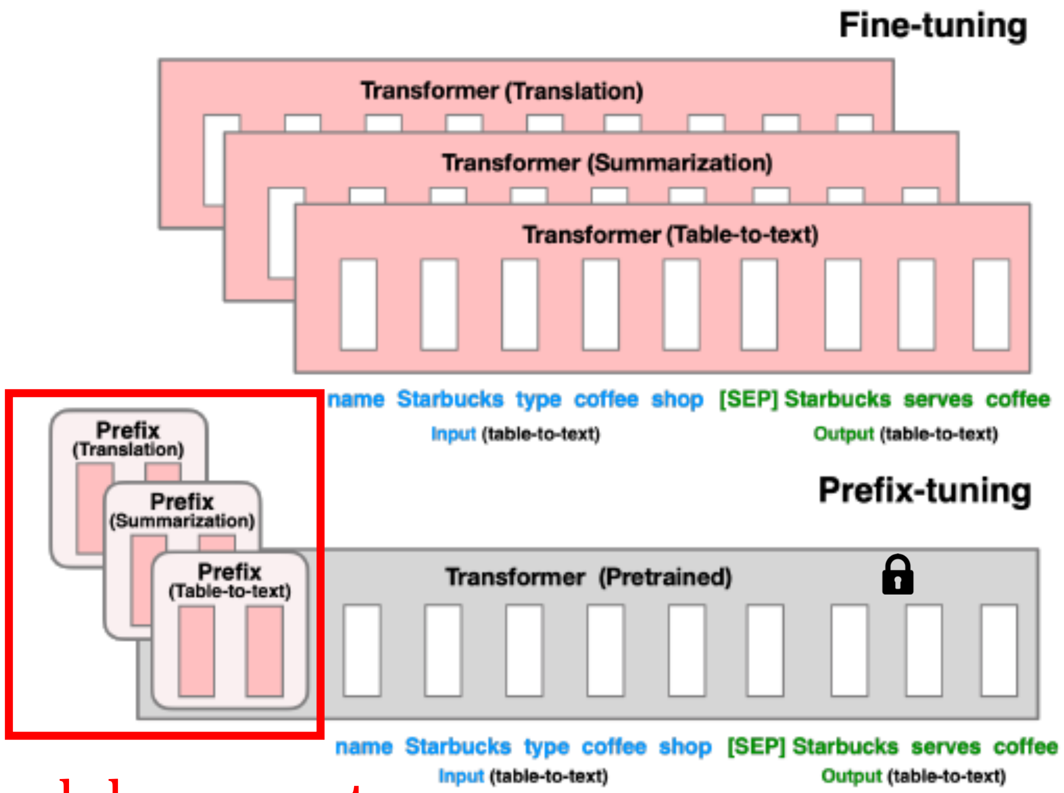
- Heavy cost of finetuning pre-trained models
  - Huge size: billions of parameters
    - E.g. GPT-3 has 175B parameters
  - Long finetuning time
  - Generalization ability (overfitting issue, stability issue)
- If we only have limited computation resource…

Q: How can we efficiently and effectively finetune pre-trained models?

Prompt learning

Prompt Learning

- Learnable "task prompts" instruct models to perform specific downstream tasks

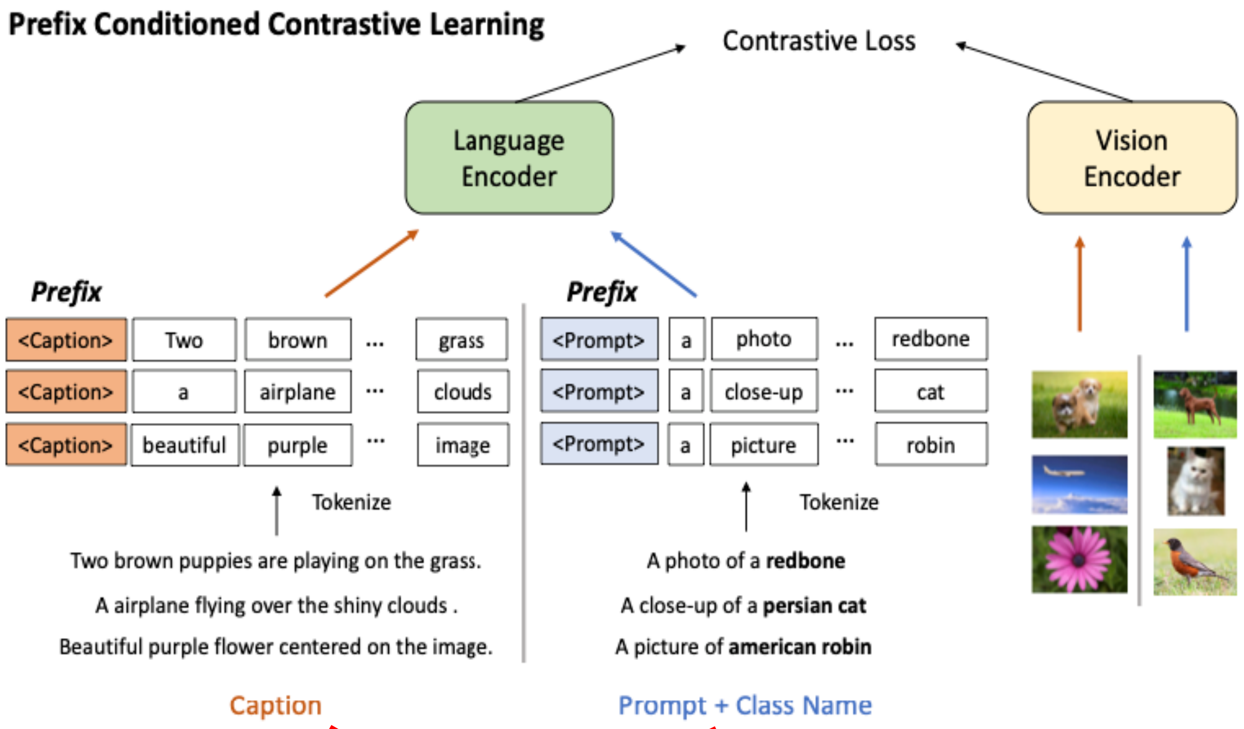


< 1% model parameters

Li et al. "Prefix-Tuning: Optimizing Continuous Prompts for Generation"

Prompt Learning

- Different prompts instruct the model learning with <u>different input distributions</u>

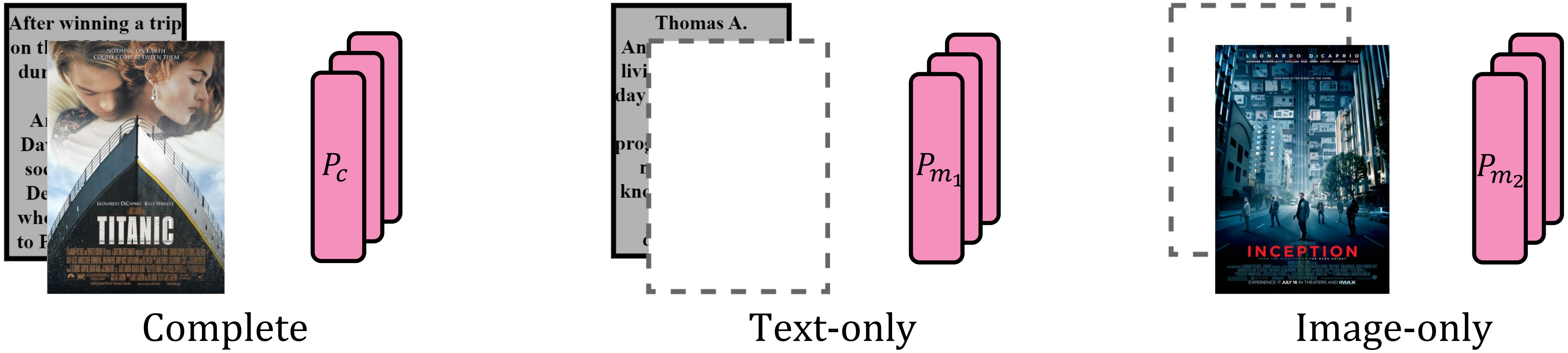


Different input distributions

Saito et al. "Prefix Conditioning Unifies Language and Label Supervision"

Motivation

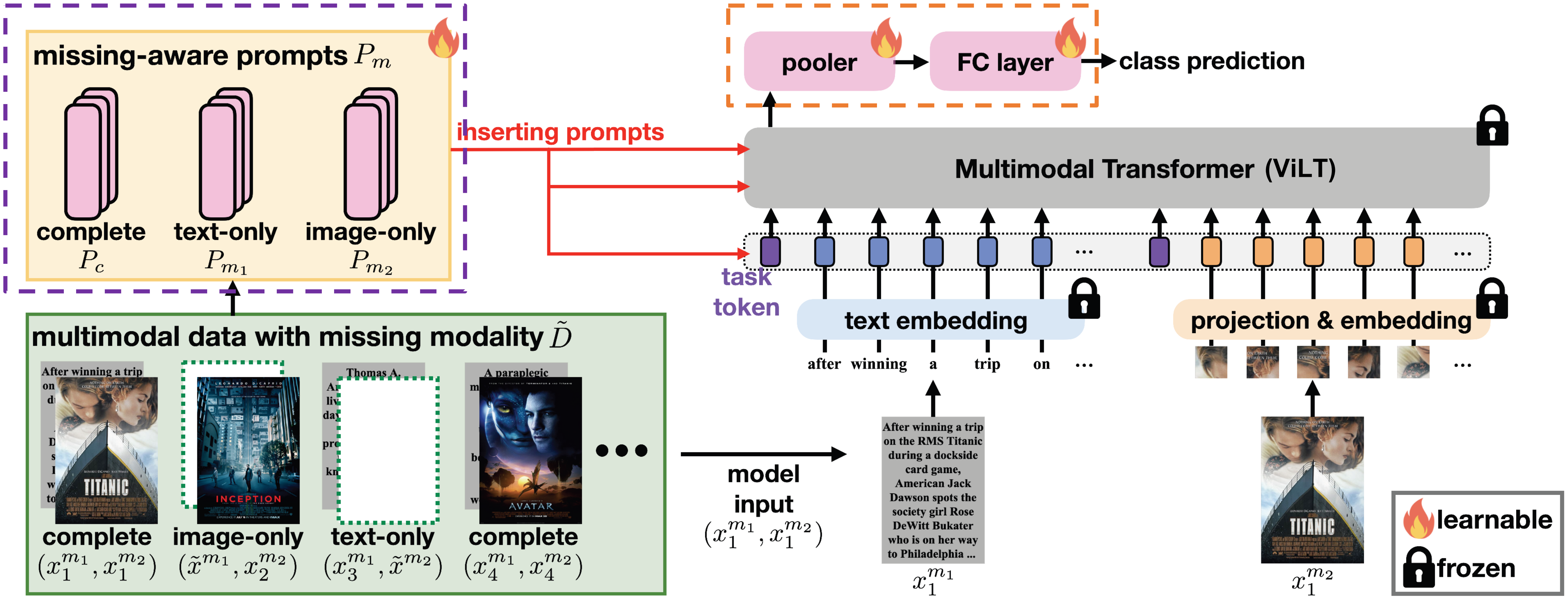
- Missing modalities can be regarded as different input distributions
    - Complete: real text + real image
    - Text-only: real text + dummy image
    - Image-only: dummy text + real image
- Use prompts to learn with modality-missing data accordingly



Proposed Method

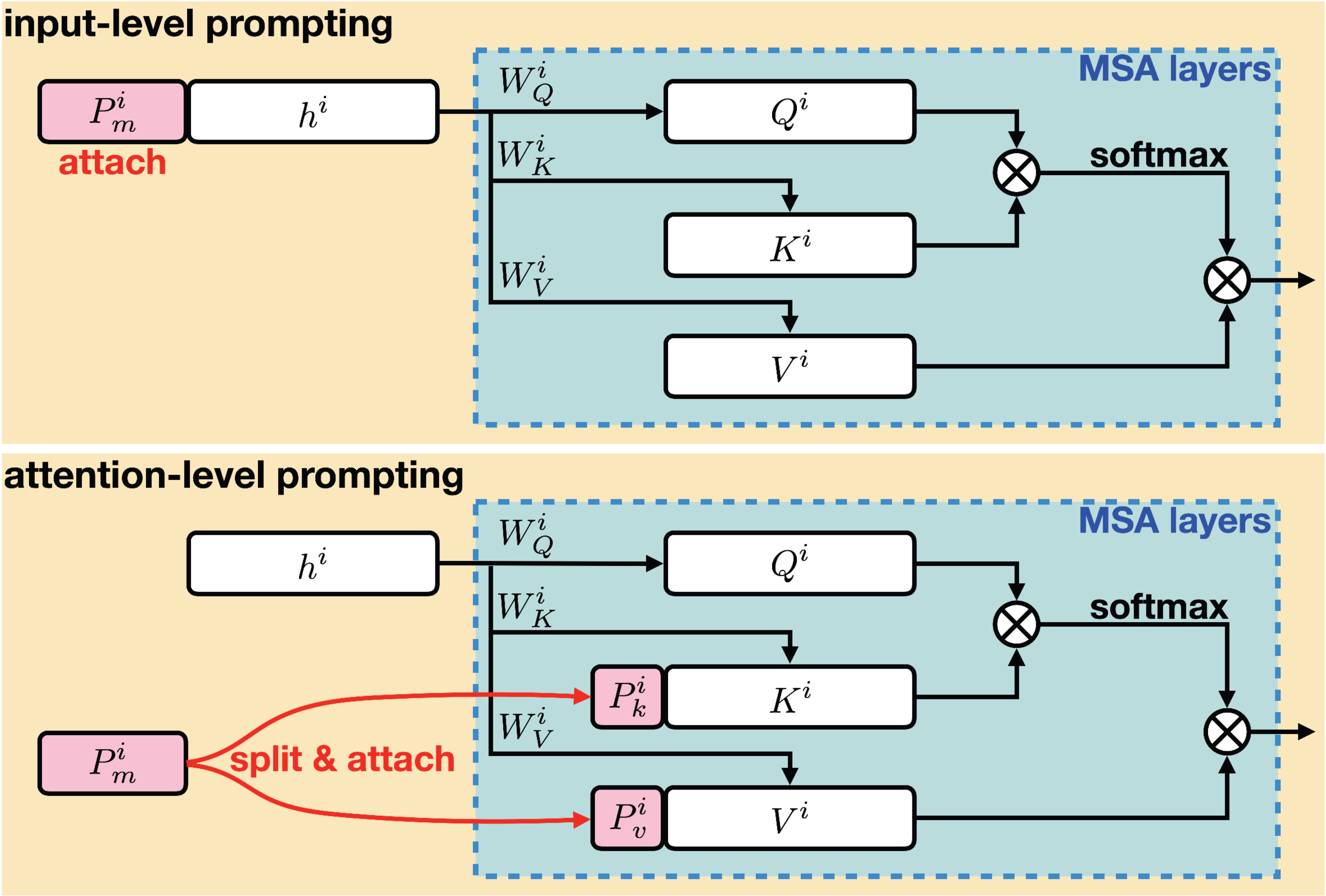
- Our prompting framework

$$L = L_{task}(x_i^{m_1}, x_i^{m_2}, \boxed{\theta_t}, \boxed{\theta_p})$$



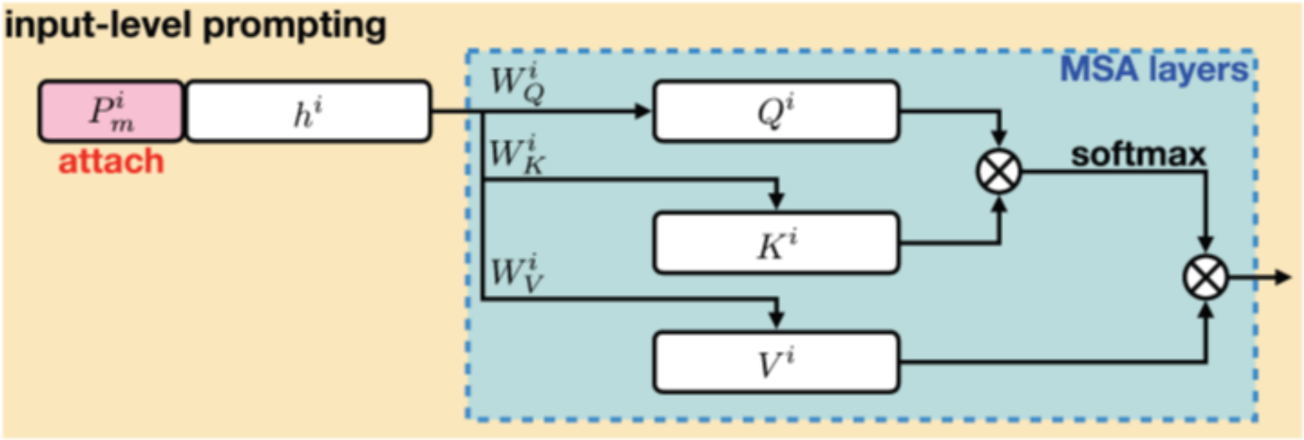
Proposed Method

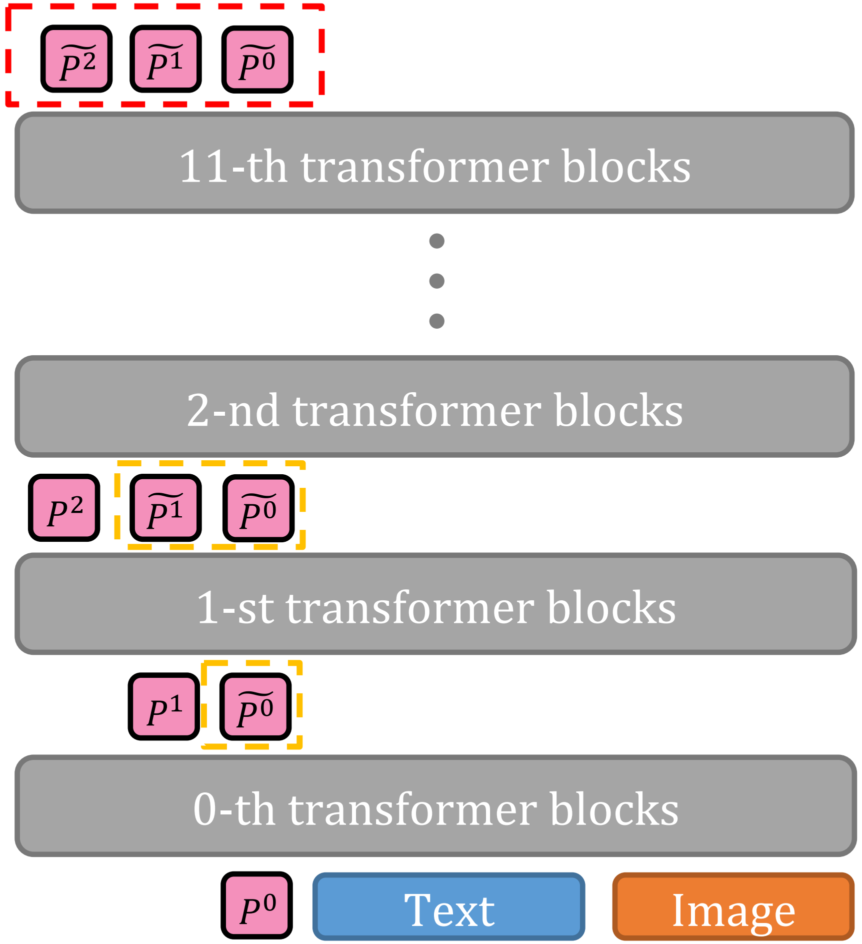
- Prompt design
  - Input-level
  - Attention-level



Proposed Method

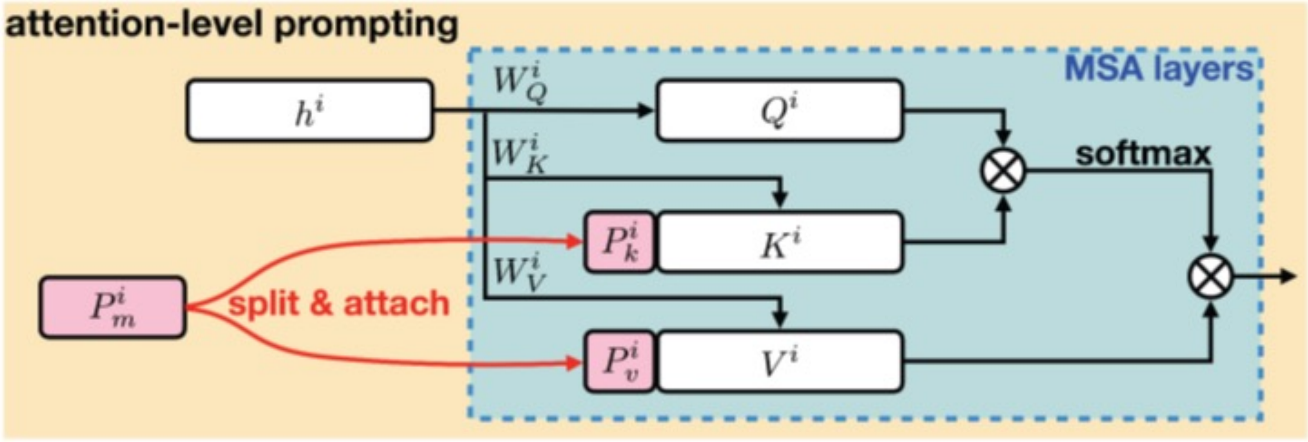
- Input-level prompting
  - Inheriting instruction information from previous layers could be helpful
  - increasing sequence length
    - Sensitive to different datasets





Proposed Method

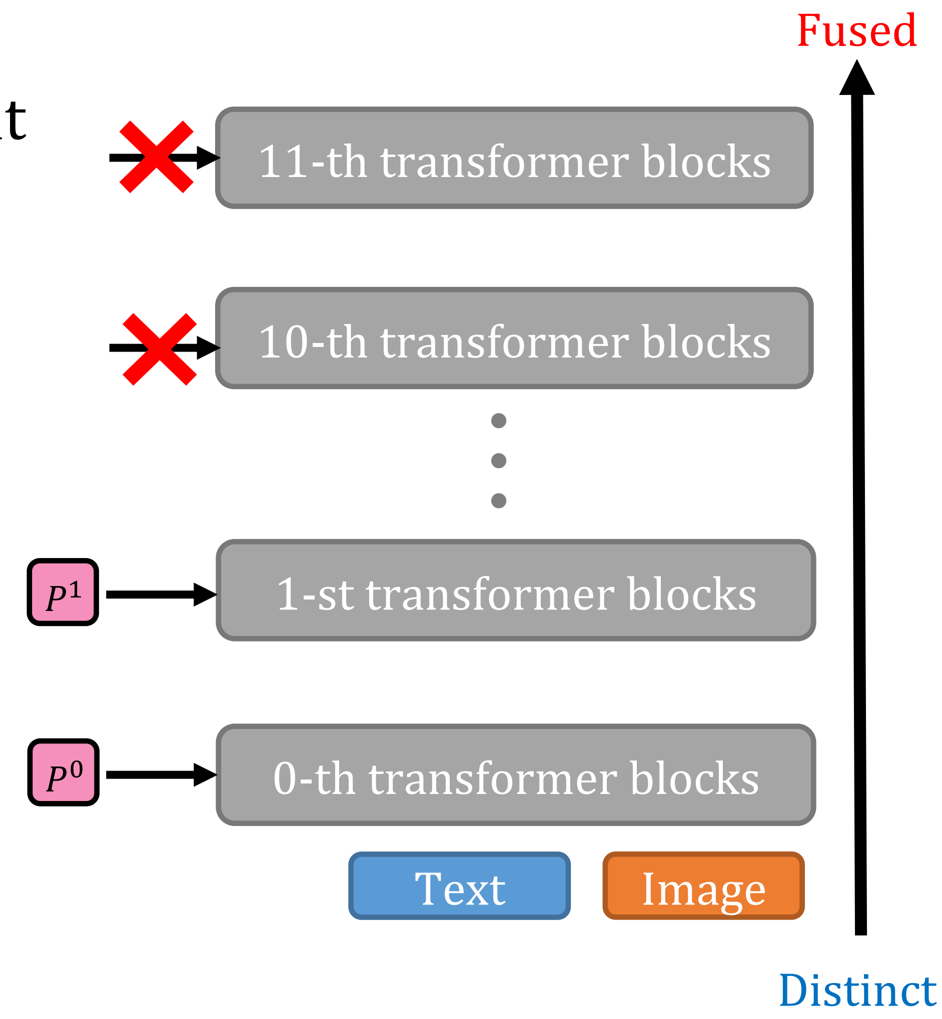
- Attention-level prompting
  - Insert the prompts into key and value of MSA layers
  - Focus on current layer instruction
  - No increasing length -> less-sensitive to datasets



$$\underbrace{\boxed{Attention^{l}}}_{L\times D}=\underbrace{\boxed{softmax\left(\frac{Q^{l}[p_k^{l};K^{l}]^{T}}{\sqrt{d}}\right)}}_{L\times(L+P)}\underbrace{\boxed{[p_v^{l};V^{l}]}}_{(L+P)\times D}$$

Proposed Method

- Location for multi-layer prompting
  - The features of different layers could be different
    - Earlier layer features are more distinct
    - Later layer features are more well-fused
  - Prompting in the earlier layer is the choice



Experiments

- Multimodal vision recognition datasets
  - MM-IMDb – multi-label classification
  - UPMC Food-101 – single-label classification
  - Hateful Memes – binary classification

Datasets	Text length	Image length
MM-IMDb	1024	192-216
UPMC Food-101	512	192-216
Hateful Memes	128	192-216

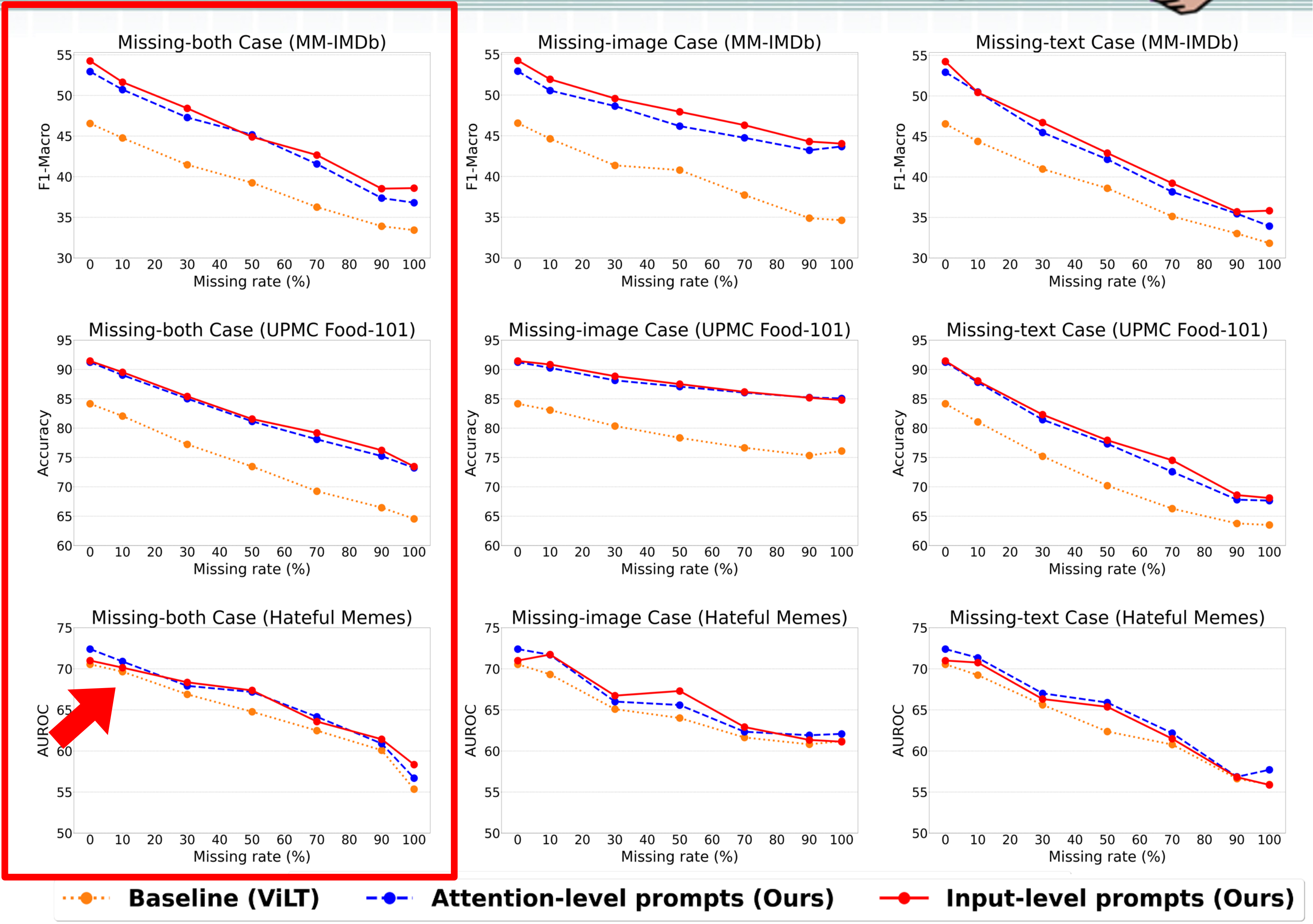
MM-IMDb

UPMC Food-101

Hateful Memes

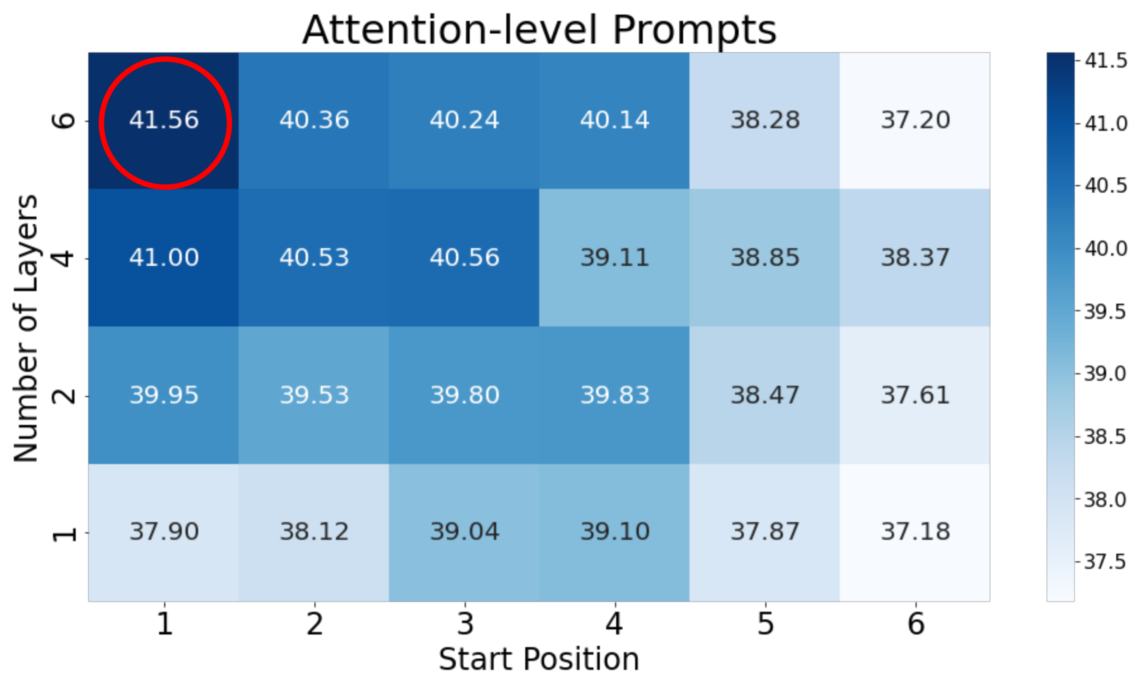
Quantitative Results

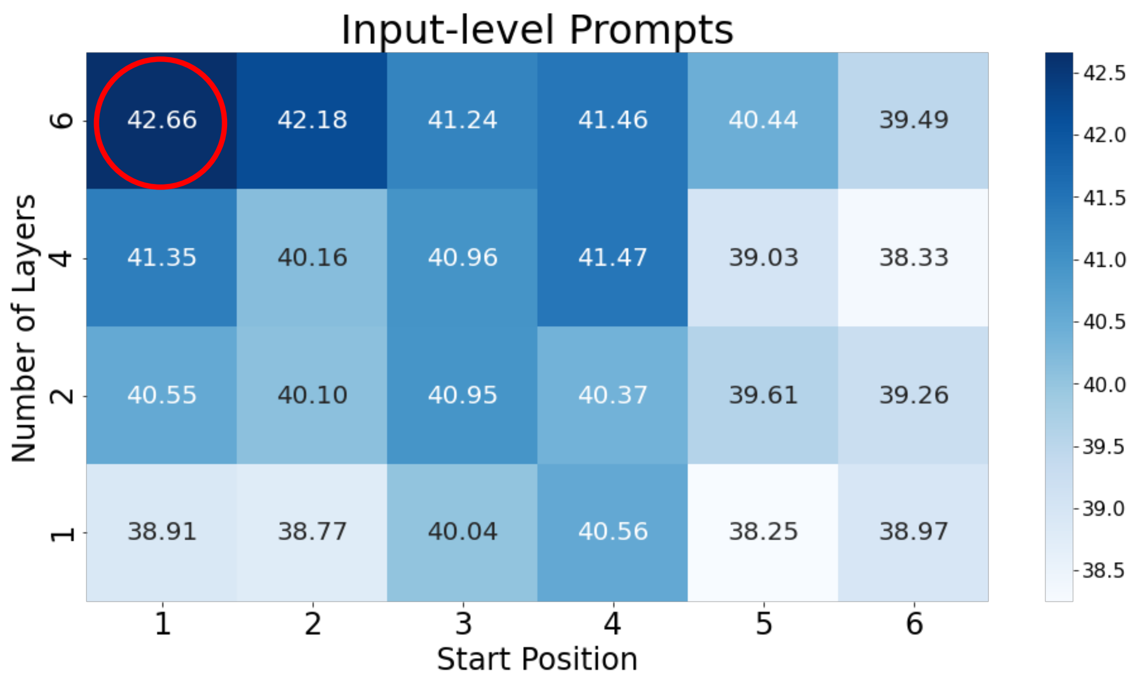
- Baseline
  - Pre-trained ViLT
  - Only train task-related models (i.e. classifier)
- Input-level prompts
  - Better performance
  - Sensitive to datasets
- Attention-level prompts
  - Consistently improvement



Ablation Study

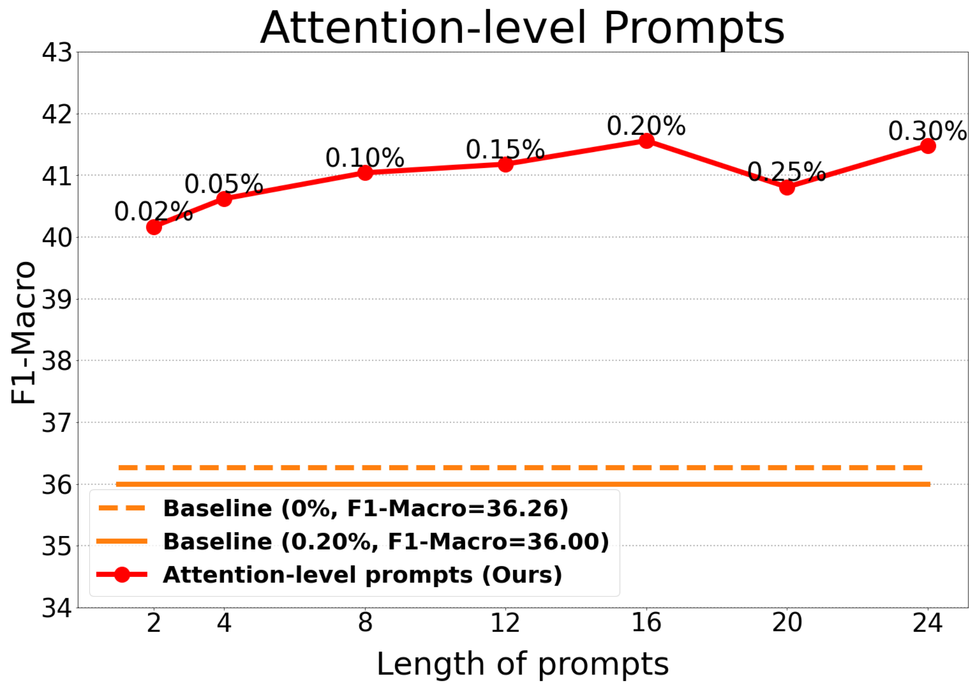
- Prompting position
  - the earlier prompting layers and more prompting layers improve the performance

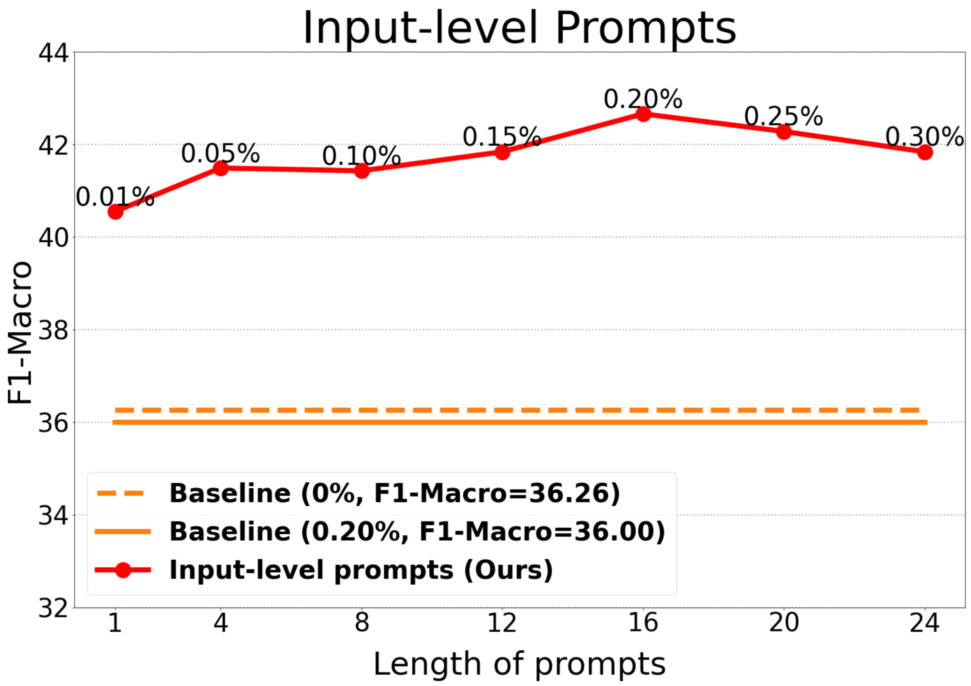




Ablation Study

- Prompt length
  - even with fewer parameters (i.e., reducing the prompt length to 1), the performance is still competitive





國立陽明交通大學
NATIONAL YANG MING CHIAO TUNG UNIVERSITY

Paper Tag: WED-PM-247

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Multimodal Prompting with Missing Modalities for Visual Recognition


Yi-Lun Lee


Yi-Hsuan Tsai


Wei-Chen Chiu


Chen-Yu Lee

Enriched Vision Applications Laboratory

Google

Please visit our github page