Mitigating Intra- and Inter-modal Forgetting in Continual Learning of Unified Multimodal Models

Xiwen Wei, Mustafa Munir, Radu Marculescu

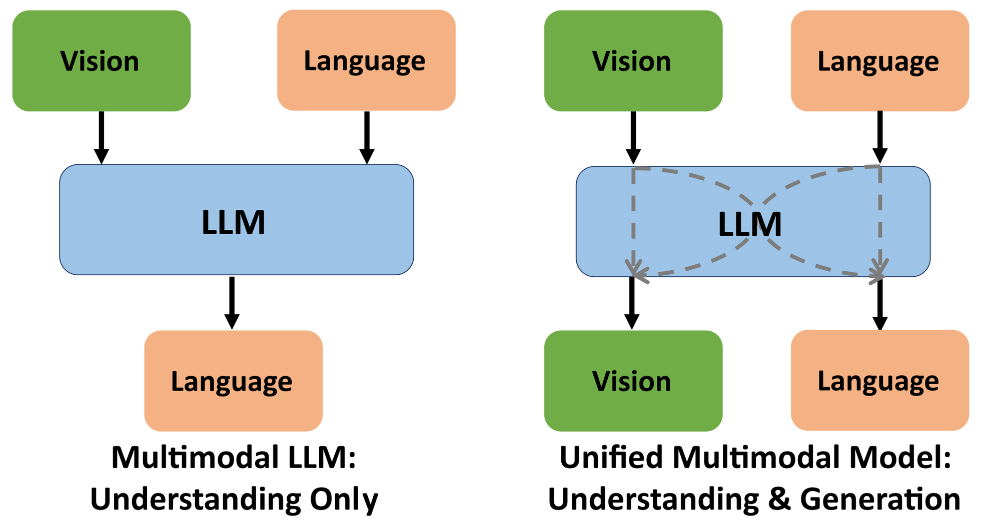
The University of Texas at Austin

NEURAL INFORMATION PROCESSING SYSTEMS

Unified Multimodal Models

Unified Multimodal Generative Models (UMGMs)

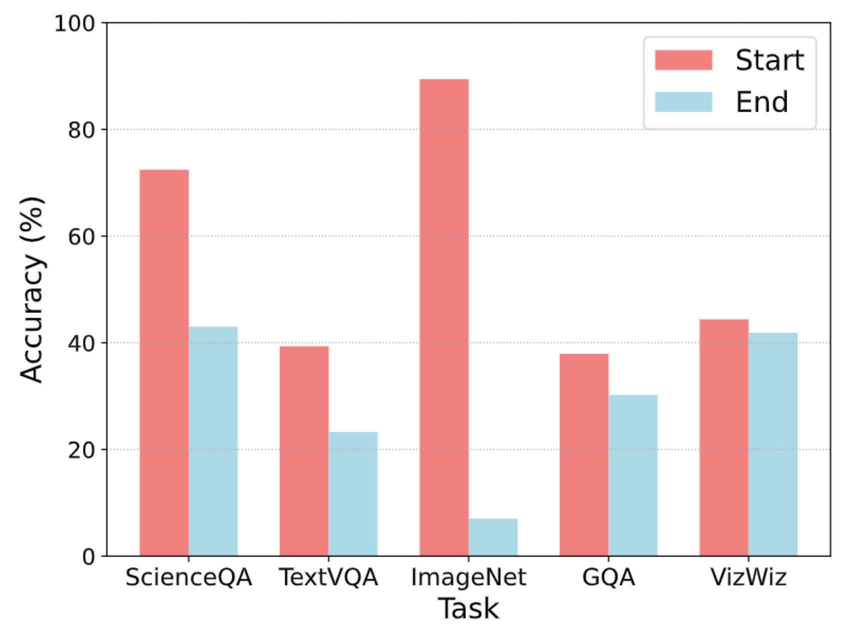
- Integrate both multimodal understanding and multimodal generation.
- Use a single autoregressive backbone.
- For general-purpose multimodal intelligence.



Catastrophic Forgetting in UMGMs

When continually adapted to new tasks, UMGMs suffer from **catastrophic forgetting**, i.e. losing performance on previously learned tasks. We aim to answer:

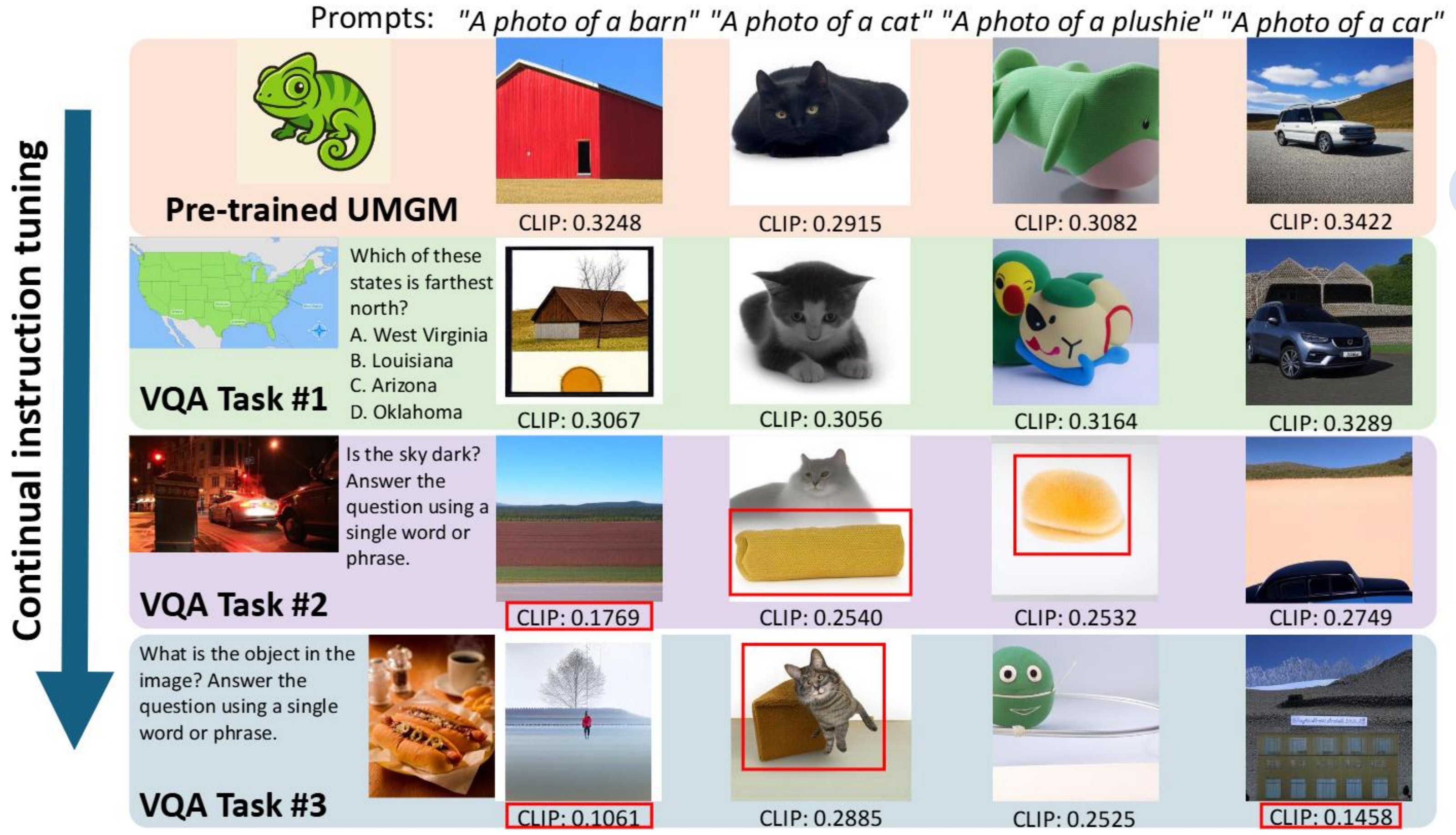
1. Do UMGMs experience both intra- and inter-modal forgetting during continual instruction tuning?
2. How can we mitigate both simultaneously?



Intra-modal forgetting: continuously learning new tasks causes forgetting on previous learned tasks.

- When trained sequentially on multiple VQA tasks, performance on earlier ones, like ScienceQA, drops dramatically.
- Existing continual learning methods mainly target this forgetting.

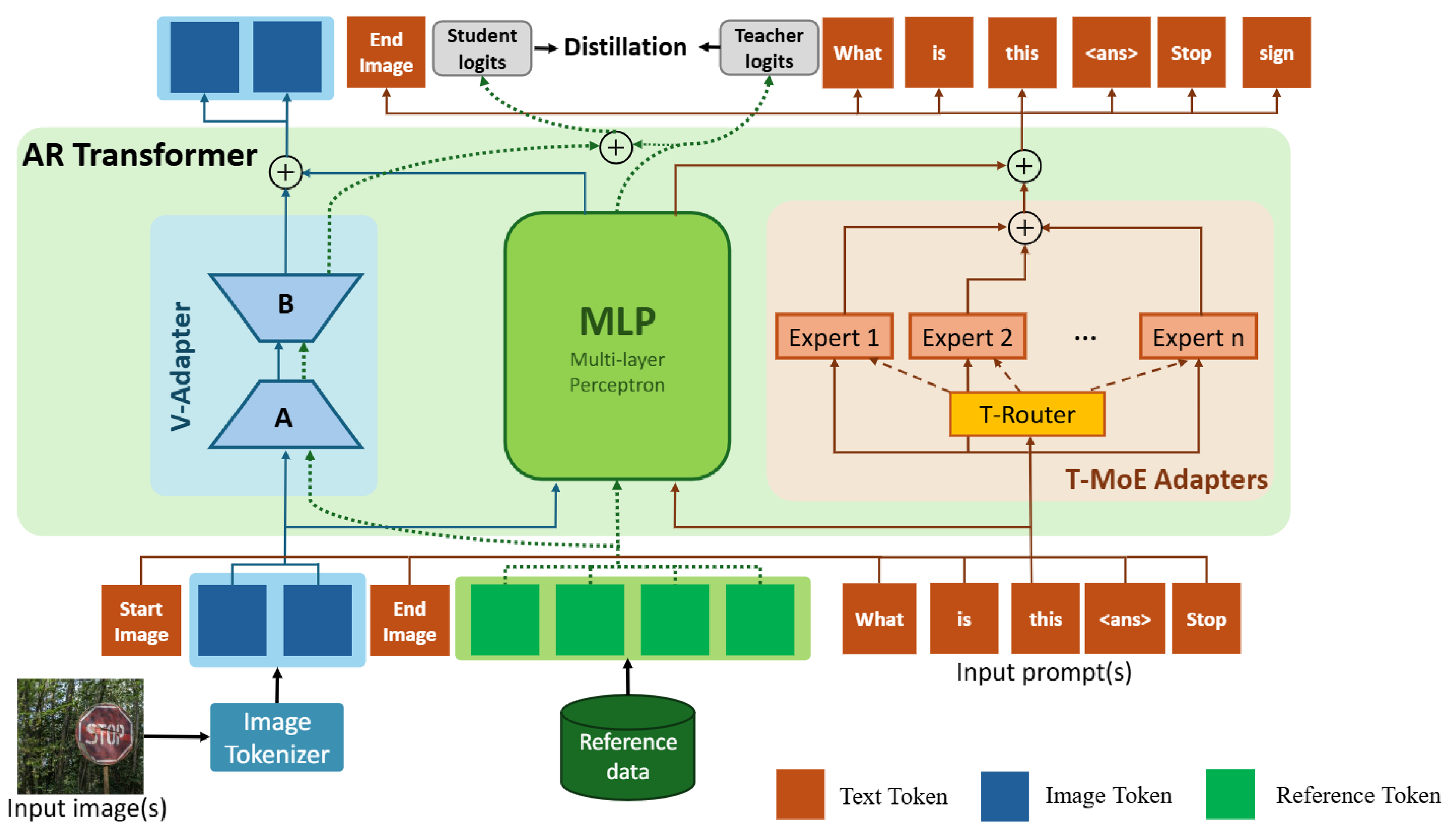
Prompts: *"A photo of a barn" "A photo of a cat" "A photo of a plushie" "A photo of a car"*



Inter-modal forgetting: improving multimodal understanding degrades multimodal generation quality.

- When trained sequentially on multiple VQA tasks, the visual generation fidelity decreases.
- Underexplored in existing work.

MoDE: Modality-decoupled Experts



Key Components

- *T-MoE (Text Mixture-of-Experts)*

A sparse mixture of LoRA experts dynamically routed per task to reduce intra-modal forgetting in multimodal understanding tasks.

- *V-Adapter (Visual LoRA Adapter)*

A single image adapter regularized via logit-level knowledge distillation from the pre-trained model to prevent inter-modal forgetting and maintain image generation fidelity.

- *Knowledge distillation (KD)*

Teacher: pre-trained model; student: currently training MoDE-adapted model.

To maintain visual generation capability of the pre-trained model.

Training Objectives

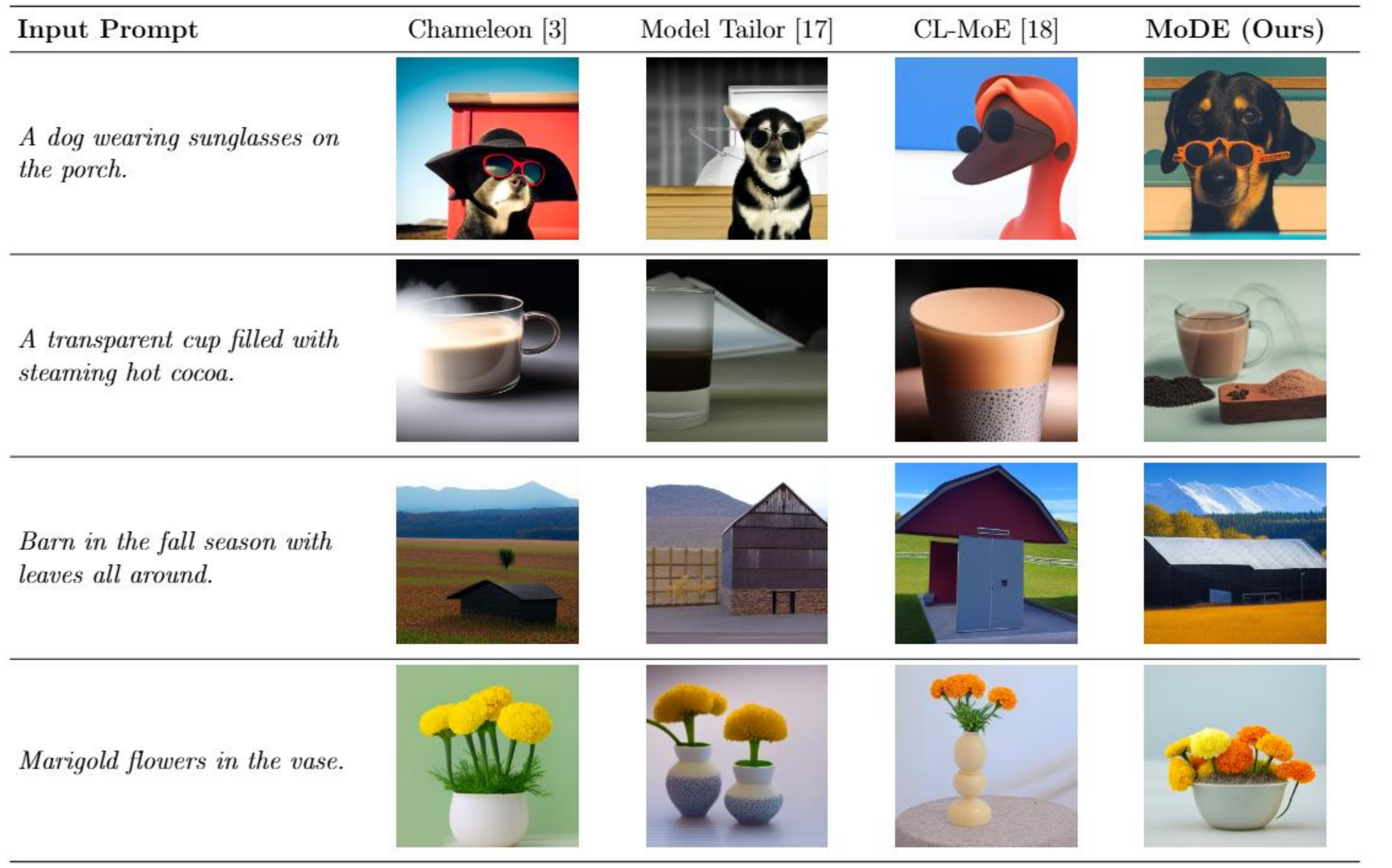
$$L_{total} = L_{T\text{-MoE}} + L_{V\text{-Adapter}} = L_{CE} + \lambda L_{KD}$$

where λ balances instruction-following accuracy and visual generation consistency.

* Only MoDE components are updated; the UMGM backbone remains frozen

Experimental Results

- Qualitative results: mitigate inter-modal forgetting



- Quantitative results: mitigate intra- and inter-modal forgetting

Method	Image Generation			Multimodal Understanding		
	Text alignment (↑)	Image alignment (↑)	FID (↓)	Accuracy (↑)	Forgetting (↓)	Δ (↓)
Zero-shot	0.2592	0.5205	52.13	22.48	-	34.84
Seq LoRA	0.2162	0.5150	56.12	28.43	35.33	28.57
Model tailor [17]	0.2384	0.5093	55.47	32.62	27.66	24.70
DualPrompt [16]	**0.2648**	0.5083	56.08	31.92	**6.82**	25.40
MoELoRA [44]	0.2248	0.5095	65.16	33.01	30.77	24.31
CL-MoE [18]	0.2081	0.5150	65.87	32.86	30.95	24.46
MoDE (Ours)	0.2458	**0.5170**	**53.74**	**33.47**	25.99	**22.78**

NSF
THE UNIVERSITY OF TEXAS AT AUSTIN
amazon
iMAGiNE INTELLIGENT MACHINE ENGINEERING CONSORTIUM