R2-T2: Re-Routing in Test-Time for Multimodal Mixture-of-Experts

Zhongyang Li[1], Ziyue Li[2], Tianyi Zhou[2]

Johns Hopkins University[1], University of Maryland, College Park[2]

ICML International Conference On Machine Learning

◆ Introduction

R2-T2 improves expert selection in multimodal Mixture-of-Experts models by locally optimizing routing weights at test time — using nearby successful examples and without changing any model parameters. Our key contributions are:

- **Test-Time Re-Routing Framework**: We formalize adjusting routing outputs at inference via reference examples.
- **Three Optimization Methods**: We propose Neighborhood Gradient Descent, Kernel Regression, and Mode Finding for per-input weight optimization.
- **Significant Performance Gains**: We demonstrate consistent, significant gains across eight benchmarks, nearing oracle performance without any retraining.

◆ Reference Set & Experts

Our reference set spans three tasks—visual understanding, reasoning, and OCR—and uses six experts:

$\mathbf{I}_{\mathbf{AUX}}$ cross-attends visual features to structured CV outputs

$\mathbf{I}_{\mathbf{LANG}}$ aligns visual features with language semantics

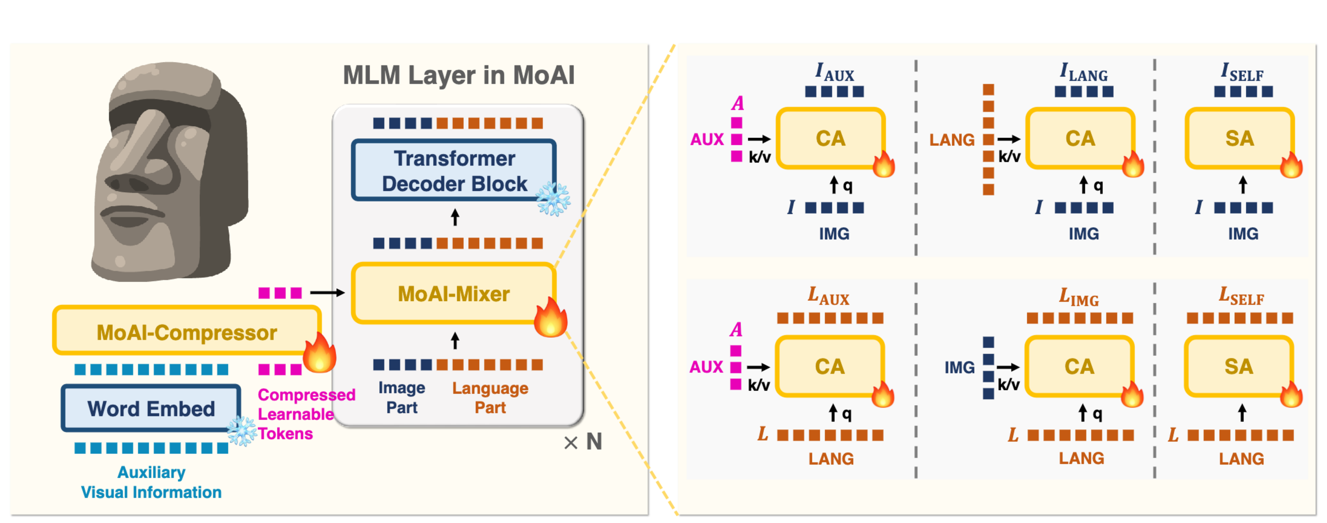
$\mathbf{I}_{\mathbf{SELF}}$ preserves spatial detail via self-attention

$\mathbf{L}_{\mathbf{AUX}}$ integrates CV outputs into language understanding

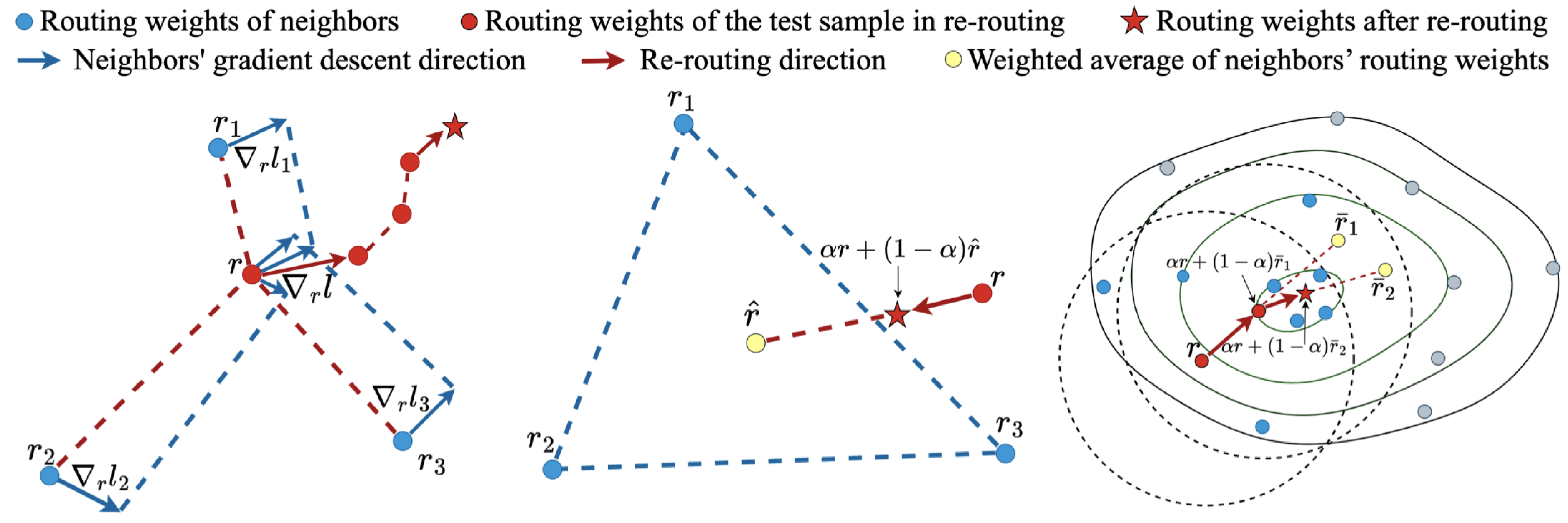
$\mathbf{L}_{\mathbf{IMG}}$ grounds language in visual context

$\mathbf{L}_{\mathbf{SELF}}$ ensures coherent text generation.

Task Type	Reference	Size	Evaluation	Size
General Visual Understanding	VQA-V2 Visual7W COCO-QA CLEVR	5,000 5,000 5,000 5,000	MMBench MME-P CVBench$^{2D/3D}$ GQA	2,374 2,114 2,638 1,590
Knowledge-Based Reasoning	A-OKVQA TQA MathVista	5,000 5,000 5,000	SQA-IMG AI2D	2,017 3,087
Optical Character Recognition	ST-VQA DocVQA	5,000 5,000	TextVQA	5,734



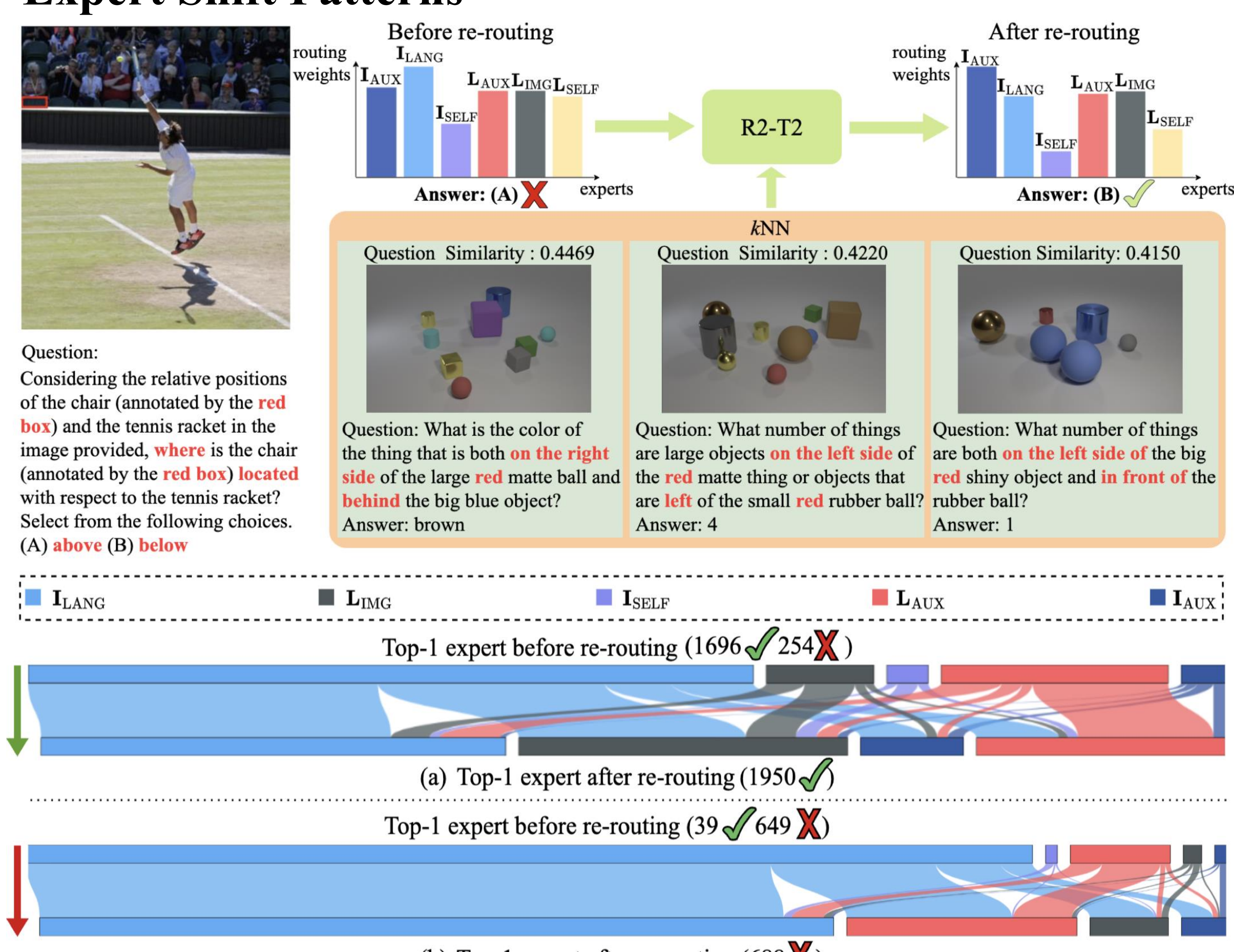
◆ Method: Test-Time Re-Routing



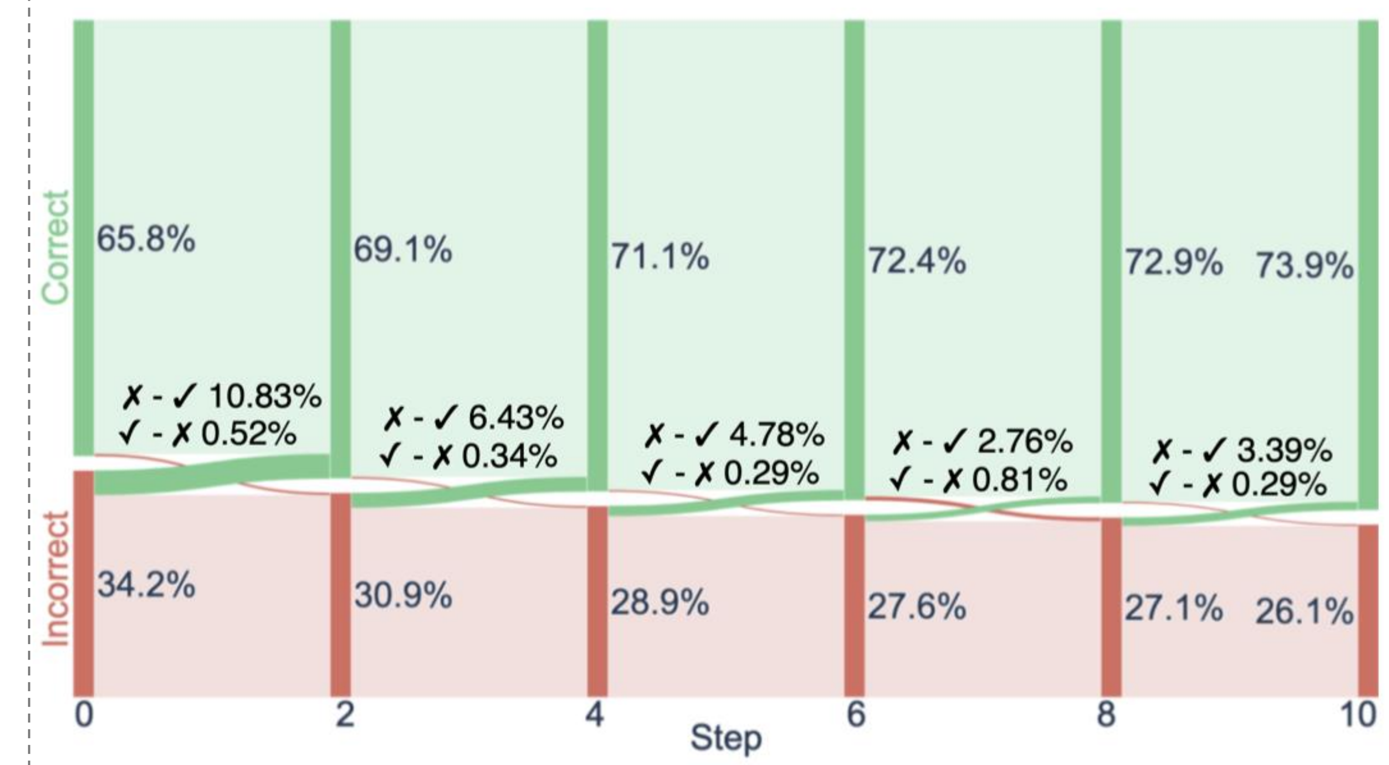
(a) Neighborhood Gradient Descent (b) Kernel Regression (c) Mode Finding

- **Neighborhood Gradient Descent** estimates the gradient of r using the loss function of the nearest neighbors in reference set and take gradient steps on r to minimize this loss.
- **Kernel Regression** computes a kernel-weighted average $\hat{r}$ of neighbors' routing vectors. Then interpolate between the original r and $\hat{r}$, using the binary search to find α that maximizes model confidence.
- **Mode Finding** identifies the high-density "mode" of neighbors' routings via a mean-shift update in the routing-weight space. Iteratively move r toward this dense region.

◆ Expert Shift Patterns



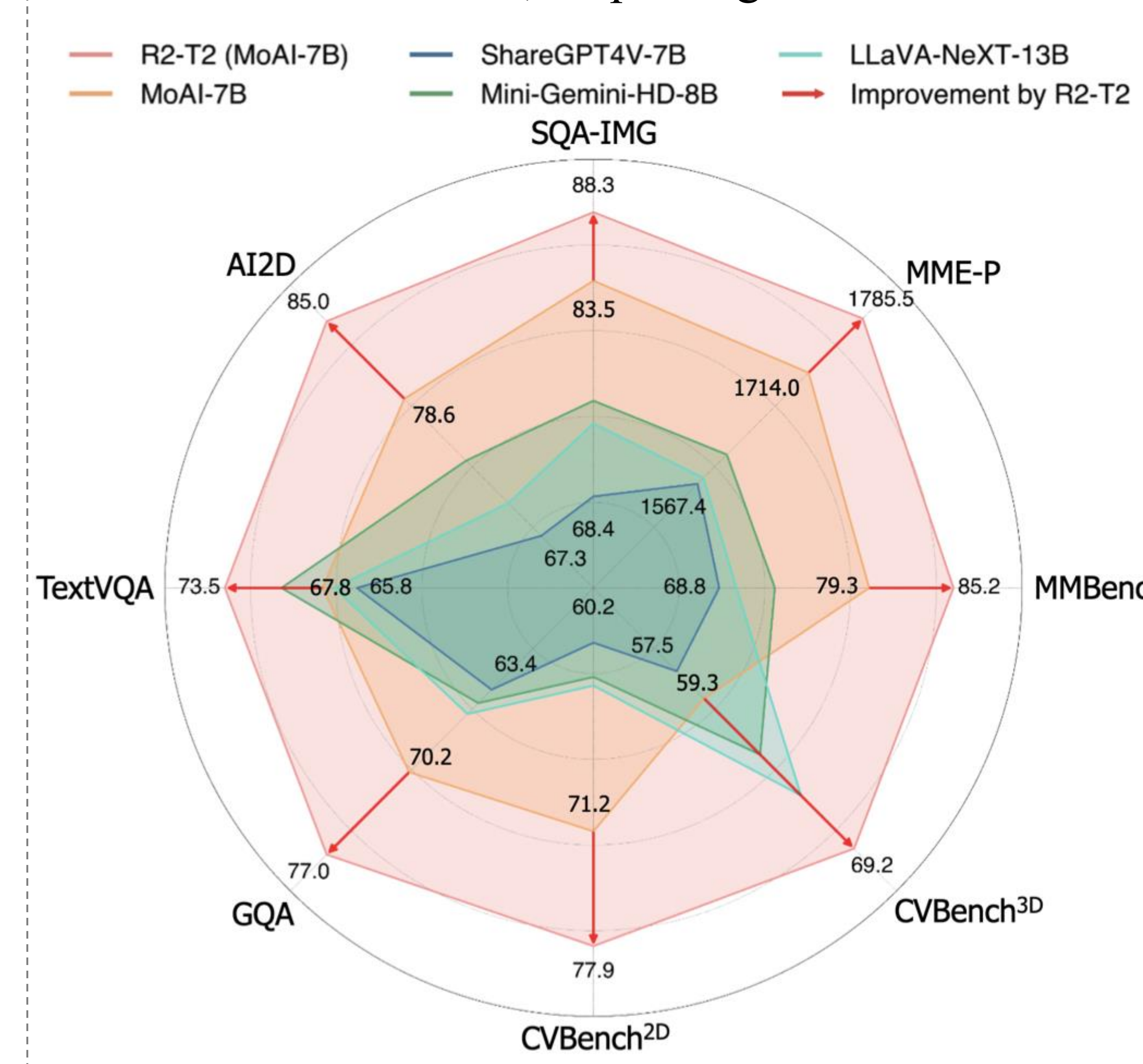
◆ Accuracy Transition Analysis



This figure illustrates the transition of predictions as NGD progresses over ten steps. During Step 0 to Step 10, a total of 28.12% of incorrect predictions have been converted to correct ones.

◆ Results

R2-T2 applied to MoAI-7B compared against 7/8/13B VLMs on 8 benchmarks, surpassing a recent 13B VLM.



Method	MMBench	MME-P	SQA-IMG	AI2D	TextVQA	GQA	CVBench^{2D}	CVBench^{3D}
MoVA (base model)	74.3	1579.2	74.4	74.9	76.4	64.8	61.6	62.3
Mode Finding	75.2	1587.1	74.9	75.8	77.3	65.7	62.5	63.2
Kernel Regression	77.9	1610.6	76.4	78.5	79.9	68.3	65.2	65.9
NGD	**81.2**	**1645.3**	**79.1**	**81.8**	**83.2**	**71.5**	**68.3**	**68.9**
Oracle (upper bound)	87.6	1735.4	87.3	88.4	89.5	76.2	72.5	73.2
MoAI (base model)	79.3	1714.0	83.5	78.6	67.8	70.2	71.2	59.3
Mode Finding	80.8	1725.2	84.1	79.8	66.5	71.4	70.0	60.1
Kernel Regression	83.7	1756.7	86.2	82.6	71.2	74.5	74.6	64.5
NGD	**85.2**	**1785.5**	**88.3**	**85.0**	**73.5**	**77.0**	**77.9**	**69.2**
Oracle (upper bound)	92.1	1860.2	93.8	91.2	79.6	83.2	84.0	76.8