MiCo: Multi-image Contrast for Reinforcement Visual Reasoning

Xi Chen[1] Mingkang Zhu[3] Shaoteng Liu[3] Xiaoyang Wu[1] Xiaogang Xu[3]
Yu Liu[2] Xiang Bai[4] Hengshuang Zhao[1]
[1] HKU [2] Tongyi Lab, Alibaba Group [3] CUHK [4] HUST

NEURAL INFORMATION PROCESSING SYSTEMS

Demonstrations

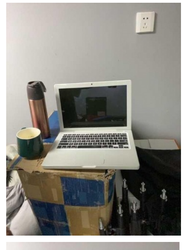
Question: How many different cups could you find?

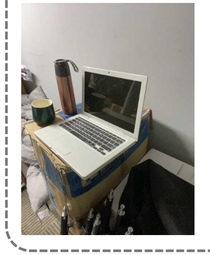
MiCo: To determine how many different cups can be found, let's analyze each image:
Image 1: The person is holding a cup with a simple design of coffee beans and steam. This appears to be one type of cup.
Image 2: The person is holding a cup with a more intricate design featuring swirls and coffee beans. This looks like a different style of cup compared to the first one.
Image 3: The cup here has a similar design to the second image, with swirls and coffee beans…This suggests that this cup might be the same as the one in Image 2.
Image 1 shows one type of cup. Image 2 shows another distinct type of cup. Image 3 shows the same cup as in Image 2.
Therefore, there are two distinct types of cups visible across all three images.

Question: How to transform the first image to the second?

MiCo: Let's analyze the differences between the two images:.
Oven Door: In the first image, the oven door is open, revealing the interior where the bread is baking. In the second image, the oven door is closed, and we can see the bread through the glass window.
Lighting: The lighting inside the oven appears brighter in the first image compared to the second image, which might be due to the open door allowing more light to enter.
Background Elements: The second image has additional elements in the background, such as a sign that reads "SUB" and a pink cloth hanging on the wall, which are not present in the first image.
Focus: The focus seems to be on the bread in both images, but the second image provides a clearer view of the bread through the closed door.
Hence, close the door of the oven.

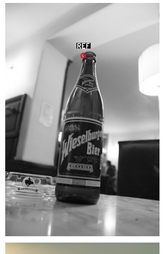
Q: How does the camera move when shooting the video?





MiCo: The differences between the two images suggest that there has been a change in the position of the laptop relative to the rest of the scene, particularly concerning its angle and the position of the items on the table. In the first image, the laptop appears to be more angled towards the viewer, while in the second image, it seems to be more upright and centered.
Given these observations, it's likely that the camera moved to capture this change in perspective. Since the the first image is from the beginning of the video and the second image is from the end, and considering the changes described.
Hence, the camera must have moved to the right to achieve this new framing.

Question: Which point is corresponding to the reference point?



MiCo: To determine the corresponding point for the reference point (REF) in the first image, we need to identify a similar point in the second image that could be analogous.
In the first image: The reference point (REF) is located near the top center of the bottle, close to the cap area.
In the second image:
- ***Point A*** is at the very top of the bottle's cap.
- ***Point B*** is slightly below the cap, near the neck of the bottle.
- ***Point C*** is around the middle of the bottle, near the label.
- ***Point D*** is at the bottom of the bottle.

The point that is closest to the top center of the bottle, similar to where the reference point is located in the first image, is ***Point A***.
Therefore, the answer is *Point A*.

Method

Algorithm 1 `MiCo`: Reinforcement Multi-image Reasoning

Input: Policy π_θ, old policy $\pi_{\theta_{\text{old}}}$, image triplet dataset $\mathcal{D} = \{(I_1, I_2, I_3)\}$, training steps $T_{\max}$, group size G, clip parameter ϵ, weak augment operators $\mathcal{T}^w$, strong augment operators $\mathcal{T}^s$
for $t = 1$ to $T_{\max}$ **do**
  Sample triplet $(I_1, I_2, I_3) \sim \mathcal{D}$
  Apply weak augmentation: $(I_1^w, I_2^w, I_3^w) = \mathcal{T}^w(I_1, I_2, I_3)$
  Apply strong augmentation: $(I_1^s, I_2^s, I_3^s) = \mathcal{T}^s(I_1, I_2, I_3)$
  Construct prompts $\mathbf{q}^w$ and $\mathbf{q}^s$ from the weak and strong augmented triplets, respectively
  Sample G CoT responses $\{\mathbf{o}_i\}_{i=1}^G$ from $\pi_{\theta_{\text{old}}}(\cdot \mid \mathbf{q}^w)$ ▷ Rollouts from weak prompt
  Evaluate reward $R_i = R(I^w, \mathbf{q}^w, \mathbf{o}_i)$ for each $i = 1, \ldots, G$
  Compute group baseline $\bar{R} = \frac{1}{G}\sum_{i=1}^G R_i$, and advantages $\hat{A}_i = \frac{R_i - \bar{R}}{\sigma(R)}$
  Optimize π_θ on the strong prompt $\mathbf{q}^s$ using the group rollouts:
  $L(\theta) = \frac{1}{G}\sum_{i=1}^G \min\left(r_i \hat{A}_i, \text{clip}(r_i, 1-\epsilon, 1+\epsilon)\hat{A}_i\right)$, where $r_i = \frac{\pi_\theta(\mathbf{o}_i|\mathbf{q}^s)}{\pi_{\theta_{\text{old}}}(\mathbf{o}_i|\mathbf{q}^s)}$
  $\theta \leftarrow \theta - \nabla_\theta L(\theta)$
  $\theta_{\text{old}} \leftarrow \theta$
end for

The core components of MiCo

- **Self-Supervised RL:** we construct scalable training samples by comparing "same/different" among video frames and image editing data.
- **Augmented-GRPO:** we apply weak augmentations to get more correct rollouts and use them to optimize harder examples with stronger augmentations.

Experiment Results

Baselines or Models	General		Object			Person				Overall*	
	Mat	Trk	Cpr	Cnt	Grp	Cpr	Cnt	Grp	VID	Avg	Δ_{human}
Chance-Level	25.00	25.00	50.00	34.88	25.00	50.00	34.87	25.00	-	32.73	-61.44
Human-Level	95.06	98.11	96.02	94.23	91.29	97.08	92.87	91.17	100.00	95.16	0.00
◦ LLaVA-OneVision[17]	16.60	13.70	47.22	56.17	27.50	62.00	46.67	37.00	47.25	39.35	-55.81
◦ LLaVA-Video-7B [43]	18.53	12.79	54.72	62.47	28.50	62.00	66.91	25.00	59.00	45.65	-49.51
◦ LongVQA-7B [42]	14.29	12.98	46.53	49.47	29.00	58.00	41.56	25.00	45.00	37.10	-58.06
◦ mPLUG-Owl2-7B [38]	17.37	18.26	49.17	62.97	31.00	63.00	58.06	29.00	43.00	40.87	-54.31
◦ Qwen2-VL-7B [2]	18.07	19.18	68.08	61.84	37.50	72.00	67.92	47.00	55.25	49.76	-45.40
◦ InternVL2.5-8B [8]	41.24	26.53	72.22	67.65	40.00	**85.00**	66.67	52.25	50.25	55.41	-39.75
◦ InternVL2.5-26B [8]	30.50	30.59	43.33	51.48	52.50	59.50	59.67	61.25	45.25	45.59	-49.57
◦ Qwen2.5-VL-7B [2]	35.91	43.38	71.39	41.72	47.50	80.00	59.76	69.00	45.00	54.82	-40.34
◦ GPT-4o [15]	37.45	39.27	74.17	**80.62**	57.50	50.00	**90.50**	47.00	**66.75**	60.36	-34.80
• MM-Eureka-7B [23]	55.60	47.03	74.10	52.50	54.00	77.50	60.00	51.00	43.50	57.24	-37.91
• NoisyRollout-7B [20]	40.93	43.83	63.33	50.83	34.50	70.50	63.33	47.00	36.50	50.08	-45.08
• ThinkLite-VL-7B [34]	40.45	46.58	75.56	62.50	49.50	77.50	62.50	51.00	36.50	55.79	-39.37
• VLAA-Thinker-7B [4]	47.49	63.03	72.20	61.40	55.00	71.00	57.50	51.00	47.75	58.49	-36.67
◦ Qwen2.5-VL-7B-CoT[2]	43.24	42.92	66.39	50.56	36.00	62.50	55.83	39.00	36.75	48.91	-46.24
• MiCo-7B-CoT	**57.14**	**67.12**	**81.94**	56.67	**58.00**	65.00	57.50	62.00	44.25	**61.06**	**-34.09**
Δ Improvement	+13.90	+24.20	+15.55	+6.11	+22.00	+2.50	+1.67	+23.00	+7.50	+12.93	+12.93

Results on VLM2-Bench

	MuirBench [32]	**BLINK** [10]	Hallusion [11]	MMStar [6]	MMMU [39]	MathVistas [22]
MM-Eureka-7B [23]	60.57	54.39	68.45	65.73	54.11	72.00
NoisyRollout-7B [20]	59.61	56.07	66.66	65.66	54.55	71.60
VLAA-Thinker-7B [4]	61.00	54.81	69.08	63.60	54.44	70.80
ThinkLite-VL-7B [34]	57.62	55.81	72.97	66.80	53.55	71.89
Qwen2.5VL-7B [2]	58.43	55.54	69.50	64.06	54.11	67.10
MiCo-7B	60.53	57.23	69.61	65.60	54.77	67.90
Δ Improvement	+2.10	+1.69	+0.11	+1.54	+0.66	+0.80

Results on General Vision Benchmarks

(a) Learning Paradigm

	General	Object	Person
Qwen2.5-VL [2]	39.64	53.53	63.44
SFT	42.90	51.15	55.98
No-CoT RL	45.36	50.01	55.23
CoT RL	62.13	65.53	57.18

(b) Data Source

	General	Object	Person
Edit Data[1] [35]	61.23	65.33	56.35
Edit Data[2] [45]	60.88	64.33	55.27
Video Data	60.29	64.50	55.68
Edit[1] + Video	62.13	65.53	57.18

(c) Rollout Augmentation

	General	Object	Person
Qwen2.5-VL [2]	39.64	53.53	63.44
(Strong, Strong)	59.41	64.00	56.98
(Weak, Weak)	55.58	62.03	54.81
(Weak, Strong)	62.13	65.53	57.18

(d) Sample Formulation

	General	Object	Person
Qwen2.5-VL [2]	39.64	53.53	63.44
Image Pairs	56.41	66.98	55.68
Image Triplets	60.64	65.33	55.81
Pairs + Triplets	62.13	65.53	57.18

(e) Prompt Diversity

	General	Object	Person
Qwen2.5-VL [2]	39.64	53.53	63.44
Single Prompt	55.53	64.16	51.50
20 Variations	62.13	65.53	57.18
50 Variations	63.13	65.29	54.93

(f) Image Augmentations

	General	Object	Person
Base (Crop, Resize)	62.13	65.53	57.18
Base + Flip	61.13	63.98	56.77
Base + Rotat.	62.58	65.03	55.86
Base + Color.	60.15	64.26	54.86

Ablation Studies

	Action	Attr. Sim.	Cartoon	Counting	Diagram	Diff. Spot.
Qwen2.5VL-7B	40.85	58.67	46.15	34.19	77.89	54.41
MiCo-7B	40.85	57.65 -	46.15	34.19	**79.90** +	**55.29** +
	Geo. Und.	Img-Text	Ordering	Scene Und.	Vis. Grnd.	Vis. Ret.
Qwen2.5VL-7B	49.00	72.63	14.06	61.83	33.33	63.70
MiCo-7B	**53.00** +	**74.14** +	**20.31** +	**63.98** +	**35.71** +	**71.23** +
	ArtStyle	Counting	Forensic	FuncCorr	IQTest	Jigsaw	MultiView
Qwen2.5VL-7B	69.23	70.83	48.48	27.69	18.00	59.33	54.89
MiCo-7B	**72.65** +	70.00 -	47.27 -	**30.77** +	**26.00** +	**69.33** +	42.11 -
	ObjLoc	RelDepth	RelReflect.	SemCorr	SpatialRel	VisCorr	VisSim
Qwen2.5VL-7B	54.10	81.45	40.30	33.09	88.81	52.33	86.67
MiCo-7B	**54.92** +	76.61 -	31.34 -	**41.73** +	**90.21** +	**61.05** +	85.19 -

Analysis in Specific Tasks