SKILL-DECOUPLED ON-POLICY EXPERT DISTILLATION FOR UNIFIED VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) can elicit specialized visual capabilities in vision-language models (VLMs), yet jointly optimizing heterogeneous objectives often introduces cross-capability gradient conflicts. Training experts independently mitigates this interference, but direct reward optimization may entangle specialized knowledge with reward-hacking biases. This paper proposes Skill-Decoupled On-Policy Expert Distillation (**SCOPED**), which trains capability-specific experts and extracts explicit skills through reward-guided reflection. During on-policy distillation, these skills guide expert supervision while remaining unavailable to the unified student, encouraging it to internalize procedural knowledge without explicit skills at inference time. Across eight multimodal benchmarks, SCOPED consistently outperforms the backbone and mixed-capability RL, demonstrating the effectiveness of skill-decoupled on-policy expert distillation for integrating independently acquired capabilities. Further analyses reveal interference during joint capability optimization and suggest that skill augmentation promotes more consistent expert trajectories, greater attention to task-relevant visual evidence, and better preservation of specialized expertise in the unified student.

1 INTRODUCTION

Large Vision-Language Models (VLMs) have demonstrated remarkable capabilities across a broad range of multimodal tasks (Bai et al., 2025a; Wang et al., 2025b). To further enhance their downstream multimodal capabilities, recent studies have increasingly adopted Reinforcement Learning (RL) (Yu et al., 2026; Xu et al., 2026a; Liu et al., 2026a) to cultivate specialized visual understanding skills through task-specific rewards, including fine-grained visual perception (Sun et al., 2025; Xiong et al., 2026), visual grounding (Zhang et al., 2026), and complex multimodal reasoning (Xia et al., 2025; Peng et al., 2025). These approaches have demonstrated the effectiveness of RL in improving specific aspects of visual understanding. However, learning and integrating complementary visual understanding capabilities across heterogeneous multimodal tasks remains a fundamental challenge in multimodal post-training (Wei et al., 2026).

A straightforward approach to enhancing VLM capabilities is to jointly optimize a shared policy over diverse multimodal training data (Ma et al., 2025; Jiang et al., 2026; Cai et al., 2026). However, heterogeneous training objectives can induce conflicts between capability-specific gradients, making it difficult for a single policy to simultaneously acquire specialized capabilities (Yu et al., 2020; Liu et al., 2021). To mitigate such conflicts, recent studies instead optimize capability-specific experts independently and subsequently consolidate their acquired knowledge through knowledge distillation. In particular, On-Policy Distillation (OPD) transfers knowledge by matching teacher predictions along student-generated trajectories (Agarwal et al., 2024; Yuan et al., 2026), while Multi-Teacher On-Policy Distillation (MOPD) extends this paradigm to multiple capability-specific teachers, enabling a shared policy to integrate complementary capabilities from different experts (Ma et al., 2026; Xiao et al., 2026). Nevertheless, directly aligning the predictions of expert policies does not explicitly distinguish reusable capability-level strategies from capability-specific patterns, which may include overfitting and reward-hacking behaviors inadvertently encoded in these experts.

To address these challenges, we introduce **S**kill-De**C**oupled **O**n-**P**olicy **E**xpert **D**istillation (**SCOPED**), a framework that decouples capability-specific skills from expert policies to better guide multimodal post-training. First, SCOPED independently optimizes a set of capability-specific

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

experts with RL, where each expert is trained on a dedicated dataset with a capability-level reward objective. This independent optimization allows different experts to explore diverse solution strategies for their respective visual understanding capabilities without interference from heterogeneous training objectives. Next, to acquire complementary visual understanding capabilities, SCOPED extracts capability-specific skills from the reasoning trajectories of these experts. These skills explicitly summarize reusable solution patterns underlying successful and failed expert behaviors. We then augment each expert with its corresponding skill to provide explicit procedural guidance during reasoning, thereby reducing unnecessary or capability-irrelevant reasoning patterns that may interfere with capability learning. Finally, SCOPED applies On-Policy Distillation (OPD) to transfer the complementary capabilities from these skill-augmented experts into a unified student model.

Extensive experiments across eight multimodal benchmarks show that SCOPED enables a compact 8B model to achieve the highest average score among the evaluated models, surpassing larger open-source and leading closed-source baselines. Ablation studies demonstrate the benefits of expert distillation and the additional contribution of skill augmentation. Further analyses reveal limitations of mixed-capability optimization, including lower training rewards than independent capability RL, weak alignment with expert-guided directions, and recurring conflicts between capability-specific gradients. We also find that skill conditioning increases trajectory consistency within each capability and directs greater attention toward task-relevant visual evidence, while the distilled student better preserves expert-induced behavior.

2 Related Work

Reinforcement learning (RL) has become an effective paradigm for specializing Vision-Language Models (VLMs) under capability-specific objectives (Liu et al., 2026b; Xia et al., 2025; Peng et al., 2026). By optimizing targeted reward signals, capability-specific RL enables models to focus on the visual understanding and reasoning behaviors required by a particular capability, thereby developing specialized expertise (Xiong et al., 2026; Zhang et al., 2026; Peng et al., 2025). To broaden capability coverage, recent efforts have explored jointly training a shared policy across heterogeneous capabilities (Ma et al., 2025; Sarch et al., 2026). However, heterogeneous objectives can induce competing gradient signals over shared parameters, causing updates beneficial to one capability to distract the optimization direction of others (Yu et al., 2020; Liu et al., 2021). As a result, joint optimization may restrict specialized expertise, highlighting the difficulty of simultaneously developing and consolidating heterogeneous capabilities within a unified policy.

Rather than jointly optimizing a shared policy, knowledge distillation provides an alternative route for consolidating specialized capabilities into a unified model. On-policy distillation aligns teacher supervision with student-generated trajectories, effectively mitigating distribution shift (Agarwal et al., 2024; Gu et al., 2024). Recent context-conditioned variants further enrich teacher supervision with privileged information or auxiliary context (Zhao et al., 2026; Ye et al., 2026; Yuan et al., 2026). For capability integration, Multi-Teacher On-Policy Distillation (MOPD) extends this paradigm to multiple independently optimized experts, distilling their complementary capabilities into a shared student along student-generated rollouts (Ma et al., 2026; Xiao et al., 2026). Nevertheless, these approaches primarily transfer expert knowledge through predictive distributions, leaving reusable reasoning patterns contained in expert trajectories largely implicit and potentially propagating noisy or over-specialized expert behaviors to the student.

To externalize such implicit trajectory-level knowledge, recent studies have explored abstracting raw execution experiences into explicit skills that capture reusable solution patterns (Xin et al., 2026). Trace2Skill consolidates execution trajectories into transferable skills through hierarchical consolidation (Ni et al., 2026), while SkillRL and Skill1 construct evolving skill libraries to support skill retrieval and policy improvement (Xia et al., 2026; Shi et al., 2026). In multimodal reasoning, Lang2Act extracts self-emergent linguistic tools from reasoning trajectories to guide fine-grained visual understanding (Xiong et al., 2026). Skill-SD further conditions the teacher on trajectory-derived skills as privileged information during self-distillation (Wang et al., 2026). Together, these studies show that trajectory-derived skills can convert implicit execution experience into explicit and reusable guidance for subsequent model optimization. While these studies demonstrate the value of trajectory-derived skills for reusable reasoning and structured supervision, how such skills

can regularize supervision from multiple independently optimized capability experts during their integration into a unified VLM remains underexplored.

3 METHODOLOGY

In this section, we present **Skill-Decoupled On-Policy Expert Distillation** (SCOPED), a framework for acquiring and transferring a diverse set of complementary visual understanding capabilities. As illustrated in Fig. 1, SCOPED consists of two stages. First, we independently train MLLMs on different capabilities with capability-specific reward signals, allowing each model to develop expertise specialized to its target visual understanding capability (Sec. 3.1). Second, to bypass the instability of joint optimization, we abstract reward-evaluated expert trajectories into explicit skills, introducing a structural information asymmetry during on-policy distillation (Sec. 3.2). Serving as privileged directional anchors, these skills regularize expert supervision, compelling the student to internalize this guidance for conflict-free consolidation.

3.1 LEARNING MLLM EXPERTISE WITH CAPABILITY-SPECIFIC REINFORCEMENT

While jointly learning diverse capabilities is essential for general visual understanding, optimizing a shared policy over heterogeneous data can introduce conflicting optimization signals and distract the learning process (Yu et al., 2020; Gao et al., 2026). As a result, improvements in one capability may interfere with the acquisition of another. To mitigate such interference, SCOPED decouples capability learning through independent reinforcement learning, producing specialized capability experts together with reward-evaluated trajectories that capture their acquired behaviors.

Eliciting Capability-Specific Expertise. To elicit the visual understanding capabilities of MLLMs, we partition the training corpus into C capability-specific subsets $\{\mathcal{B}_c\}_{c=1}^{C}$, where each instance consists of a multimodal input x and its ground-truth answer a. For each capability c, an expert policy is initialized from the shared base policy π_θ and optimized independently on $\mathcal{B}_c$. This decoupled optimization allows each expert to specialize in its designated capability without being directly affected by competing optimization objectives from other capabilities.

For an instance $(x, a) \in \mathcal{B}_c$, the corresponding capability expert samples a rollout trajectory:

$$\tau \sim \pi_\theta(\cdot \mid x), \tag{1}$$

where τ denotes the generated trajectory. Each trajectory is evaluated against the target answer a using a capability-specific reward function r_c. Guided by these reward signals, the capability expert π_{θ_c} is obtained by maximizing the expected reward over the subset $\mathcal{B}_c$:

$$\theta_c = \arg\max_{\theta} \mathbb{E}_{(x,a)\sim\mathcal{B}_c,\, \tau\sim\pi_\theta(\cdot|x)} \left[r_c(a, \tau) \right]. \tag{2}$$

The specific design of r_c is described in Appendix A.3.

Collecting Capability-Specific Experiences. After capability-specific reinforcement learning, we collect a diverse set of trajectories from each trained capability expert π_{θ_c}. Specifically, for each instance $(x, a) \in \mathcal{B}_c$, we sample K distinct rollouts:

$$\tau_1, \ldots, \tau_K \sim \pi_{\theta_c}(\cdot \mid x). \tag{3}$$

These trajectories provide concrete behavioral traces of the strategies acquired through capability-specific reinforcement, including recurring solution patterns and capability-relevant actions (Guo et al., 2025; Xiong et al., 2026).

To properly contextualize the expert's behavioral process, we associate each sampled trajectory τ_k, where $k \in \{1, \ldots, K\}$, with its input context and target answer to form a trajectory tuple (x, a, τ_k). Aggregating these tuples over the entire subset $\mathcal{B}_c$ yields the global capability-specific trajectory repository:

$$\mathbb{T}_c = \bigcup_{(x,a)\in\mathcal{B}_c} \{(x, a, \tau_k)\}_{k=1}^{K}. \tag{4}$$

The resulting $\mathbb{T}_c$ serves as a comprehensive repository of both successful and failed behaviors acquired by the capability expert, which is used to supervise the training process in Sec. 3.2.

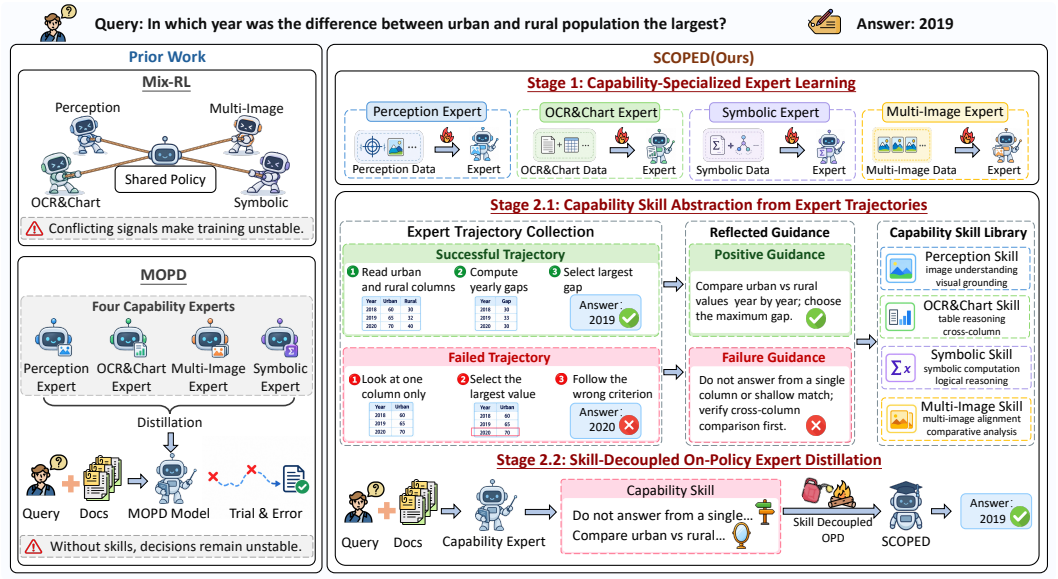


Figure 1: Overview Architecture of SCOPED.

3.2 Unifying Capability Experts via Skill-Augmented Distillation

As described in Sec. 3.1, directly distilling raw expert predictions risks propagating reward-hacking behaviors and their associated biases (Gao et al., 2023; Chen et al., 2024; Helff et al., 2026). To isolate reusable expertise from such behaviors, SCOPED first extracts capability-specific skills from the expert's global trajectory repository through reward-guided reflection, contrasting successful and failed executions to identify verified solution patterns. During subsequent on-policy distillation (Yuan et al., 2026; Ma et al., 2026), these extracted skills serve as *privileged directional guidance* (Wang et al., 2026), conditioning the expert's supervision on explicit procedural knowledge rather than directly exposing the student to potentially biased predictions. Under this structural information asymmetry, the student aligns with the skill-regularized supervision, thereby internalizing the pure expertise while mitigating the propagation of erroneous reward-driven behaviors.

Extracting Capability-Specific Skills. Given the global trajectory repository $\mathbb{T}_c$, we extract reusable solution patterns through reward-guided reflection. Specifically, we partition all collected trajectories into successful and failed cohorts evaluated by the reward function r_c:

$$\mathbb{T}_c^+ = \{(x, a, \tau) \in \mathbb{T}_c \mid r_c(a, \tau) \geq \gamma\}, \quad \mathbb{T}_c^- = \{(x, a, \tau) \in \mathbb{T}_c \mid r_c(a, \tau) < \gamma\}, \tag{5}$$

where γ denotes the reward threshold defining successful and failed executions. The capability expert π_{θ_c} then performs reward-guided reflection over trajectories from these two cohorts. By identifying recurring patterns in the successful trajectories $\mathbb{T}_c^+$ and diagnosing failure modes in the unsuccessful ones $\mathbb{T}_c^-$, the expert produces a capability-specific reflection result $\mathcal{R}_c$:

$$\mathcal{R}_c = \pi_{\theta_c}\left(\text{Instruct}_{\text{reflect}}\left(\mathbb{T}_c^+, \mathbb{T}_c^-\right)\right), \tag{6}$$

where $\text{Instruct}_{\text{reflect}}$ denotes the overall reflection procedure over successful and failed trajectories, with outcome-specific reflection prompts detailed in Appendix A.5.

Finally, we guide the capability expert to consolidate the extracted insights into an explicit capability-specific skill:

$$S_c = \pi_{\theta_c}(\text{Instruct}_{\text{syn}}(\mathcal{R}_c)), \tag{7}$$

where $\text{Instruct}_{\text{syn}}$ instructs the expert to synthesize the aggregated reflections $\mathcal{R}_c$ into explicit procedural guidance. The resulting skill S_c provides an explicit symbolic representation of the reusable expertise acquired by the capability expert π_{θ_c}, which will serve as a privileged conditional anchor to regularize the expert during subsequent distillation.

Skill-Augmented Distillation. To translate decoupled skills into regularized supervision, SCOPED utilizes S_c as a privileged directional anchor exclusively for each capability expert. This asymmetric conditioning elicits precise reasoning traces and highly stable target distributions (Huang et al., 2026). By distilling from these skill-augmented experts, the unified student receives clean supervision, effectively avoiding the exploratory noise and reward-hacking biases inherent in raw expert policies.

For a given sample $(x, a) \in \mathcal{B}_c$, the unified student policy π_θ generates an on-policy response $y \sim \pi_\theta(\cdot \mid x)$. At decoding step t, the predictive distributions of the student and the corresponding capability expert π_{θ_c} are respectively given by:

$$p_{\theta,t} = \pi_\theta(\cdot \mid x, y_{<t}), \quad q_{\theta_c,t} = \pi_{\theta_c}(\cdot \mid x, S_c, y_{<t}), \tag{8}$$

where $y_{<t}$ denotes the student-generated prefix. Under this asymmetric conditioning, S_c acts as a directional anchor that regularizes the target distribution $q_{\theta_c,t}$, guiding the expert along verified reasoning paths while effectively suppressing erroneous or over-specialized tokens.

To consolidate the diverse expertise into a single model, we optimize the unified student policy over all C capabilities. The overall skill-augmented distillation loss $\mathcal{L}$ is defined as the expected token-level reverse KL divergence aggregated across all capability-specific subsets:

$$\mathcal{L} = \frac{1}{C} \sum_{c=1}^{C} \mathbb{E}_{(x,a)\sim\mathcal{B}_c,\, y\sim\pi_\theta(\cdot|x)} \left[\frac{1}{|y|} \sum_{t=1}^{|y|} D_{\mathrm{KL}}(p_{\theta,t} \,\|\, q_{\theta_c,t}) \right], \tag{9}$$

where $D_{\mathrm{KL}}(\cdot \,\|\, \cdot)$ denotes the reverse Kullback-Leibler divergence, and $|y|$ is the length of the generated response. Minimizing this objective directly embeds the regularized expert knowledge into the student's weights, yielding a unified model that is entirely self-sufficient and requires no explicit skill conditioning during inference.

4 Experimental Methodology

This section first describes the datasets, evaluation metrics, and baselines. Then, we introduce the implementation details of our experiments.

Datasets. The evaluation covers eight benchmarks across four capability groups. CountBenchQA (Paiss et al., 2023) and PixMo-Count (Deitke et al., 2025) are used for fine-grained perception; CC-OCR V2 (Xu et al., 2026c) and ChartQA (Masry et al., 2022) for OCR and chart understanding; VisualPuzzles (Song et al., 2025) and We-Math (Qiao et al., 2025) for symbolic visual reasoning; and MUIRBench (Wang et al., 2025a) and MMSI-Bench (Yang et al., 2025) for multi-image reasoning. For capability-specific expert training, reinforcement learning data are constructed from six public multimodal datasets, including task-relevant subsets of Orsta-Data-47K (Ma et al., 2025). Detailed dataset composition and statistics are provided in Appendix A.1.

Evaluation Metrics. We employ task-specific metrics following the official evaluation protocols of each benchmark. Fine-grained perception, chart understanding, symbolic visual reasoning, and multi-image reasoning tasks are evaluated using accuracy. OCR recognition is evaluated using macro-F1 following CC-OCR V2 (Xu et al., 2026c). All scores are reported as percentages, and the overall score is the arithmetic mean across the eight benchmarks.

Baselines. In our experiments, we compare SCOPED with representative closed-source and open-source vision-language models. The closed-source baselines include Claude Sonnet 4.6 and Claude Opus 4.6 (Anthropic, 2026b;a), GPT-5.2 (OpenAI, 2025), and GPT-5.4 (OpenAI, 2026). The open-source baselines include GLM-4.1V-9B-Base (Hong et al., 2025), Qwen2.5-VL-7B (Bai et al., 2025b), InternVL3.5-8B (Wang et al., 2025b), Qwen3-VL-8B and Qwen3-VL-32B (Bai et al., 2025a), and MiMo-VL-7B-RL (Xiaomi LLM-Core Team, 2025).

Implementation Details. In our experiments, we employ Qwen3-VL-8B-Instruct and Qwen3-VL-4B-Instruct (Bai et al., 2025a) as the shared backbone to construct all capability-specialized experts and the unified student. Capability-specific reinforcement learning is implemented

Table 1: Overall Performance. Comparison with closed-source and open-source models across four capability groups, including fine-grained perception, OCR and chart understanding, symbolic visual reasoning, and multi-image reasoning. SCOPED-4B and SCOPED-8B are built on Qwen3-VL-4B and Qwen3-VL-8B, respectively.

Model	Perception		OCR & Chart		Symbolic		Multi-Image		Avg.
	CountB.	PixMo	CC-OCR	ChartQA	VisPuz.	WeMath	MUIR	MMSI	
Closed-Source Models									
Claude-Sonnet-4.6	90.63	68.69	82.37	89.40	52.91	87.18	76.35	34.80	72.79
Claude-Opus-4.6	89.21	64.71	80.74	90.88	52.23	87.70	80.08	37.30	72.86
GPT-5.2	88.39	77.04	86.31	87.44	39.47	70.92	70.00	33.30	69.11
GPT-5.4	91.24	77.80	72.35	88.12	51.28	74.25	68.77	35.00	69.85
Open-Source Models									
GLM-4.1V-9B-Base	75.76	48.77	67.07	83.12	27.31	35.46	52.50	27.20	52.15
Qwen2.5-VL-7B	74.95	36.81	85.50	45.80	25.60	65.57	61.19	27.70	52.89
InternVL3.5-8B	77.60	56.74	78.53	88.16	43.24	77.53	59.42	24.80	63.25
Qwen3-VL-4B	81.47	64.90	82.86	85.60	35.10	76.26	72.12	30.20	66.06
Qwen3-VL-8B	83.91	67.36	85.72	81.32	46.40	71.67	72.92	29.10	67.30
MiMo-VL-7B-RL	86.56	65.65	72.83	87.92	45.03	**83.39**	69.27	29.40	67.51
Qwen3-VL-32B	90.22	75.52	86.08	87.60	**48.72**	79.83	74.31	**33.80**	72.01
Unified Multi-Expert Distillation									
SCOPED-4B	88.39	74.38	85.03	89.48	38.01	78.85	76.15	29.40	69.96
SCOPED-8B	**90.43**	**79.89**	**88.79**	**88.28**	46.58	82.47	**76.58**	32.00	**73.13**

with EasyR1 (Zheng et al., 2025), and multi-expert on-policy distillation is implemented with veRL (Sheng et al., 2024). All experiments are conducted on NVIDIA A800 GPUs. More training implementation details and hyperparameter settings are presented in Appendix A.2.

5 Results and Analysis

In this section, we first evaluate the overall performance of SCOPED across eight multimodal benchmarks and conduct ablation studies to assess the effectiveness of its key components. We then provide in-depth analyses of cross-capability gradient interference under mixed-capability optimization and investigate how skill augmentation influences expert supervision during capability integration.

5.1 Overall Performance

As shown in Table 1, we report the performance of various VLMs across eight benchmarks. The evaluation results show that SCOPED consistently improves the Qwen3-VL backbone across heterogeneous visual understanding capabilities. In particular, SCOPED-8B outperforms its backbone on all eight benchmarks, achieving an average improvement of 5.83 points. These results demonstrate that our framework can effectively consolidate complementary capabilities into a unified model.

Compared with existing open-source and closed-source models, SCOPED-8B achieves the best overall performance among the evaluated models and further surpasses the substantially larger model Qwen3-VL-32B, demonstrating that effective capability integration can yield substantial gains without simply scaling up the model. We adopt Qwen3-VL as the backbone due to its competitive performance across the evaluated capability groups. Nevertheless, existing models often exhibit uneven performance across different capabilities, making it challenging to simultaneously maintain strong performance in perception, OCR and chart understanding, symbolic reasoning, and multi-image reasoning. In contrast, SCOPED consistently strengthens the backbone across all four capability groups. We further apply SCOPED to Qwen3-VL-4B, where it achieves an average improvement of 3.90 points and surpasses several larger open-source models. The consistent gains observed with

Table 2: Ablation Study. All models are built on Qwen3-VL-8B. We compare training and capability integration strategies, with capability-specific experts shown as references for assessing expertise retention. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Perception		OCR & Chart		Symbolic		Multi-Image		Avg.
	CountB.	PixMo	CC-OCR	ChartQA	VisPuz.	WeMath	MUIR	MMSI	
Capability-Specific Experts									
Perception-RL	92.67	79.89	85.80	85.00	47.09	79.60	71.00	30.50	71.44
+ Perception Skill	93.69	82.35	85.82	85.08	47.00	79.54	71.12	30.60	71.90
OCR & Chart-RL	82.89	73.06	86.08	89.64	47.17	82.76	74.42	30.40	70.80
+ OCR & Chart Skill	82.69	73.24	86.23	90.60	47.26	82.87	74.31	30.30	70.94
Symbolic-RL	82.28	69.26	85.81	81.12	48.37	82.30	75.81	30.40	69.42
+ Symbolic Skill	82.08	69.07	85.78	81.36	48.80	84.20	75.85	30.50	69.71
Multi-Image-RL	83.10	67.74	85.61	81.08	47.43	79.71	76.00	30.70	68.92
+ Multi-Image Skill	83.30	67.93	85.59	81.04	47.35	79.66	76.50	32.00	69.17
Unified-Model Baselines and Ablations									
Zero-shot	83.91	67.36	85.72	81.32	46.40	71.67	72.92	29.10	67.30
SFT	56.32	66.92	85.77	60.36	33.99	57.59	65.23	25.10	56.41
Mix-RL	89.80	73.80	86.01	85.00	40.10	**82.90**	67.80	29.10	69.31
SCOPED (Parameter Merging)	86.35	73.43	85.83	85.32	**49.49**	81.32	75.31	30.10	70.89
SCOPED (w/o Skill Augmentation)	89.41	**79.89**	**88.89**	88.16	43.75	81.26	76.27	31.50	72.39
SCOPED	**90.43**	**79.89**	88.79	**88.28**	46.58	82.47	**76.58**	**32.00**	**73.13**

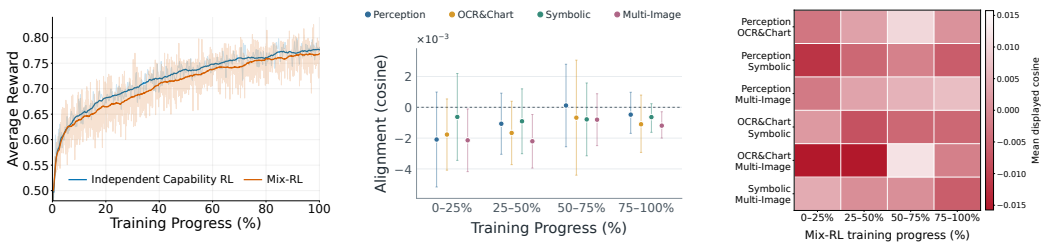
both the 4B and 8B backbones demonstrate that the effectiveness of SCOPED generalizes across model scales and remains robust across diverse visual capabilities.

5.2 Ablation Study

As shown in Table 2, we conduct ablation studies to investigate the effectiveness of different multi-task training strategies and the individual components of SCOPED. We first compare SCOPED with SFT and Mix-RL to assess the benefits of our training paradigm. We then ablate key components of SCOPED to examine their respective contributions. For reference, we also report the performance of capability-specific experts, each of which is optimized exclusively with task-specific rewards.

The evaluation results show that SFT underperforms the zero-shot backbone, suggesting that jointly fine-tuning on capability-specific data can introduce interference across capabilities. Although Mix-RL improves overall performance, it still underperforms the corresponding specialized experts on symbolic visual reasoning and multi-image reasoning, indicating that joint optimization may limit the acquisition of specialized capabilities. In contrast, SCOPED achieves an average improvement of 3.82 points over Mix-RL, demonstrating the effectiveness of separating capability acquisition from capability integration. The skill-augmented expert results further show that matching each expert with its corresponding skill consistently improves the target capability, while its effects on non-target capabilities are limited and occasionally negative. This result highlights the capability-specific nature of the extracted guidance and motivates explicit skill-expert matching during distillation.

Furthermore, we compare two variants of SCOPED, SCOPED (Parameter Merging) and SCOPED w/o Skill Augmentation, which use parameter merging and distillation without skill augmentation, respectively, to unify multiple capability-specific experts into a single model. SCOPED (Parameter Merging) consistently underperforms the on-policy distillation variants, suggesting that directly merging expert parameters is insufficient to fully integrate complementary capabilities. In contrast, SCOPED w/o Skill Augmentation already outperforms Mix-RL, indicating that token-level supervision from the corresponding experts along student-generated trajectories provides more fine-grained guidance for learning different capabilities. This allows the unified model to progressively absorb expertise that was independently acquired by the specialized experts. Building on this foundation, the full SCOPED framework achieves a further average improvement, with gains across all four capability groups. These results indicate that explicit skills extracted from expert experience provide more targeted distillation signals, facilitating specialized capability distillation.



(a) Training Reward. (b) Capability-wise Update Alignment of the Mix-RL Model. (c) Pairwise Gradient Conflict.

Figure 2: Cross-Capability Interference in Mix-RL. (a) Average training reward of independent capability RL and Mix-RL. (b) Alignment of shared Mix-RL updates with capability-specific expert directions. (c) Pairwise gradient cosine similarity across Mix-RL training stages.

5.3 Gradient Interference in Multi-Capability Training

As illustrated in Figure 2, we investigate the cross-capability gradient interference phenomenon (Yu et al., 2020; Liu et al., 2021) during shared-policy optimization from three complementary perspectives: training reward, expert-guided update alignment, and pairwise gradient conflict.

We first evaluate the training dynamics in Figure 2a. While both independent capability RL and Mix-RL steadily improve their average rewards, independent optimization maintains a consistent advantage throughout most of the training process. Although the performance gap narrows toward convergence, the overall trend clearly indicates that mixed-capability optimization structurally limits the gains achievable through capability-specific learning, preventing the model from fully benefiting from independently optimized capability objectives. To uncover the mechanistic cause of this limitation, we analyze how the realized Mix-RL updates align with the behaviors acquired by independently trained capability experts. We characterize this parameter-space update alignment using cosine similarity (Xu et al., 2026b). Specifically, for each capability, we derive a local expert-guided direction from the negative gradient of the KL divergence between the current Mix-RL policy and the corresponding fixed capability expert under matched response prefixes. We then compute its cosine similarity against the actual parameter displacement between adjacent Mix-RL checkpoints. In this formulation, positive values denote movement toward the expert, whereas near-zero or negative values indicate orthogonal or contradictory updates. As shown in Figure 2b, the shared updates show weak alignment with the expert-guided directions, with negative values observed across multiple capabilities and training stages. This confirms a severe see-saw effect: shared-policy optimization inevitably forces trade-offs, improving certain capabilities at the direct expense of others rather than consistently approximating all capability-specific experts simultaneously.

This interference is further supported by analyzing the capability-specific optimization directions themselves. At each Mix-RL checkpoint, we estimate capability-specific gradients using the corresponding capability data and compute pairwise cosine similarities among the four capabilities (Yu et al., 2020; Liu et al., 2021). Positive cosine similarity indicates aligned optimization directions, whereas negative values indicate that two capabilities favor opposing local updates. Figure 2c reveals recurrent negative cosine similarities across multiple capability pairs at different training stages, providing direct empirical evidence that heterogeneous capabilities can favor competing local optimization directions. Together, these results demonstrate that shared-policy optimization is constrained by cross-capability interference. These results further support the motivation of SCOPED to decouple capability acquisition into independently optimized experts prior to subsequent integration.

5.4 Effectiveness of Skill-Augmented Distillation in SCOPED

In this section, we further investigate how skill augmentation regularizes expert supervision during the distillation process of SCOPED. As illustrated in Figure 3, we examine its effects from three complementary perspectives: the preservation of expert-induced behavior in the unified student, the consistency of expert trajectories across instances within each capability, and the utilization of task-relevant visual evidence.



(a) Capability Retention of Experts. (b) Trajectory Consistency. (c) Layer-wise Attention Hit Rate.

Figure 3: Analysis of Skill-Augmented Expert Supervision. (a) Capability retention and alignment, macro-averaged across four capabilities. (b) Within-capability trajectory embedding similarity with and without skill conditioning. (c) Layer-wise attention hit rate on task-relevant visual regions with and without skill conditioning.

We first characterize functional capability preservation using token-level log-probability shifts relative to the base model on fixed trajectories (Klabunde et al., 2025). For each trajectory, we construct a shift vector from these log-probability changes and compare the unified student with the corresponding expert in terms of amplitude retention and expert alignment. Amplitude retention quantifies how much of the expert-induced shift is preserved by the student, while expert alignment measures the cosine similarity between their shift vectors. We evaluate models distilled from different capability-specific RL experts across visual understanding capabilities. As shown in Figure 3a, SCOPED achieves higher amplitude retention and expert alignment than Mix-RL, demonstrating better preservation of expert-induced functional behavior across capabilities. Notably, SCOPED also outperforms SCOPED w/o Skill Augmentation, indicating that the extracted skills help the student better acquire and reconcile complementary expert capabilities.

To further uncover the role of skill augmentation in SCOPED, we evaluate the consistency of expert generation by encoding reasoning trajectories produced by each capability expert with and without its corresponding skill using Qwen3-VL-Embedding-2B (Li et al., 2026). Within each capability, we compute pairwise cosine similarities between trajectory embeddings generated from different inputs to obtain the distributions shown in Figure 3b. Higher similarity indicates more consistent expert behavior across instances. With skill conditioning, the similarity distribution shifts upward, SCOPED shifts the similarity distribution upward, indicating that explicit skill conditioning regularizes expert generation and promotes more consistent trajectory representations. Finally, we investigate whether this trajectory-level regularization extends to visual evidence utilization. Building on findings that attention patterns in VLMs can effectively capture task-relevant visual grounding (Kang et al., 2025), we compare each capability expert with and without its corresponding skill under identical inputs. At each decoder layer, we identify the most-attended visual token for each attention head and record a hit if it falls within the task-relevant region. We then average the resulting hit rates across attention heads and evaluation samples. As shown in Figure 3c, skill-augmented experts consistently achieve higher hit rates across most of the upper decoder layers, with the largest improvement observed in the final layer. This finding indicates that explicit skills steer capability experts toward more accurate allocation of attention to task-relevant visual evidence. Taken together, these multi-level analyses suggest that skill conditioning promotes more consistent expert trajectories and more effective visual grounding, while enabling the distilled student to better preserve expert-induced functional behavior.

6 Conclusion

This paper proposes SCOPED, aiming at integrating complementary visual understanding capabilities into a unified Vision-Language Model through skill-decoupled on-policy expert distillation. SCOPED independently optimizes capability-specific experts and abstracts reusable solution patterns from reward-evaluated trajectories into explicit capability skills. During on-policy distillation, these skills guide capability experts to provide structured supervision for the unified student. Our experiments demonstrate that SCOPED consistently improves performance across diverse visual understanding capabilities. Our further analyses reveal cross-capability interference during mixed-capability optimization and show that skill augmentation improves the preservation and transfer of specialized expert capabilities.

AI USE STATEMENT

Generative AI tools were used solely for language polishing and improving the clarity of the manuscript. All research ideas, experimental design, implementation, analysis, and conclusions were independently developed and verified by the authors, who take full responsibility for the content of the paper.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Anthropic. Introducing Claude Opus 4.6. https://www.anthropic.com/news/claude-opus-4-6, February 2026a. Published February 5, 2026.

Anthropic. Introducing Claude Sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6, February 2026b. Published February 17, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *ArXiv preprint*, 2025b. URL https://arxiv.org/abs/2502.13923.

Min Cai, Yu Liang, Longzheng Wang, Yan Wang, Yueyang Zhang, Long Xia, Zhiyuan Sun, Xi Ye, and Daiting Shi. Advancing general-purpose reasoning models with modular gradient surgery. *arXiv preprint arXiv:2602.02301*, 2026.

Lichang Chen, Chen Zhu, Davit Soselia, Jiuhai Chen, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. Odin: Disentangled reward mitigates hacking in rlhf. *arXiv preprint arXiv:2402.07319*, 2024.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 91–104, 2025.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International conference on machine learning*, pp. 10835–10866. PMLR, 2023.

Xingjie Gao, Pengcheng Huang, Zhenghao Liu, Yukun Yan, Shuo Wang, Zulong Chen, Chen Qian, Ge Yu, and Yu Gu. Teaching llms to learn tool trialing and execution through environment interaction. *arXiv preprint arXiv:2601.12762*, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *The twelfth international conference on learning representations*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Lukas Helff, Quentin Delfosse, David Steinmann, Ruben Härle, Hikaru Shindo, Patrick Schramowski, Wolfgang Stammer, Kristian Kersting, and Felix Friedrich. Llms gaming verifiers: Rlvr can lead to reward hacking. *arXiv preprint arXiv:2604.15149*, 2026.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Jiazhen Huang, Xiao Chen, Xiao Luo, Yong Dai, Senkang Hu, and Yuzhi Zhao. Skill-conditioned gated self-distillation for llm reasoning. *arXiv preprint arXiv:2605.28791*, 2026.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024.

Jingjing Jiang, Chongjie Si, Jun Luo, Hanwang Zhang, and Chao Ma. Co-reinforcement learning for unified multimodal understanding and generation. *Advances in Neural Information Processing Systems*, 38:5823–5846, 2026.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9339–9350. IEEE, 2025.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in neural information processing systems*, 34:18878–18890, 2021.

Shuliang Liu, Zhipeng Xu, Zhenghao Liu, Yukun Yan, Minghe Yu, Yu Gu, Chong Chen, Huiyuan Xie, and Ge Yu. Mitigating judgment preference bias in large language models through group-based polling. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 1448–1464, 2026a.

Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified reasoning-integrated visual perception via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 94069–94086, 2026b.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, et al. Mopd: Multi-teacher on-policy distillation for capability integration in llm post-training. *arXiv preprint arXiv:2606.30406*, 2026.

Yan Ma, Linge Du, Xuyang Shen, Shaoxiang Chen, Pengfei Li, Qibing Ren, Liz Ma, Yuchao Dai, Pengfei Liu, and Junjie Yan. One rl to see them all: Visual triple unified reinforcement learning. *arXiv preprint arXiv:2505.18129*, 2025.

Ahmed Masry, Jia Qing Tan, Shafiq Joty, Enamul Hoque, et al. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pp. 2263–2279, 2022.

Jingwei Ni, Yihao Liu, Xinpeng Liu, Yutao Sun, Mengyu Zhou, Pengyu Cheng, Dexin Wang, Erchao Zhao, Xiaoxi Jiang, and Guanjun Jiang. Trace2skill: Distill trajectory-local lessons into transferable agent skills. *arXiv preprint arXiv:2603.25158*, 2026.

OpenAI. Introducing GPT-5.2. `https://openai.com/index/introducing-gpt-5-2/`, December 2025. Published December 11, 2025.

OpenAI. Introducing GPT-5.4. `https://openai.com/index/introducing-gpt-5-4/`, March 2026. Published March 5, 2026.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3170–3180, 2023.

Chunyi Peng, Zhipeng Xu, Zhenghao Liu, Yishan Li, Yukun Yan, Shuo Wang, Yu Gu, Minghe Yu, Ge Yu, and Maosong Sun. Mixture-of-retrieval experts for reasoning-guided multimodal knowledge exploitation. *arXiv preprint arXiv:2505.22095*, 2025.

Chunyi Peng, Haidong Xin, Xuanshuo Sheng, Xin Dai, Zhenghao Liu, Shuo Wang, Yukun Yan, Zulong Chen, Yu Gu, and Ge Yu. Memory shot for long-term dialogue. *arXiv preprint arXiv:2606.28338*, 2026.

Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, Yifan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20023–20070, 2025.

Gabriel Sarch, Linrong Cai, Qunzhong Wang, Haoyang Wu, Danqi Chen, and Zhuang Liu. Vero: An open rl recipe for general visual reasoning. *arXiv preprint arXiv:2604.04917*, 2026.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Yaorui Shi, Yuxin Chen, Zhengxi Lu, Yuchun Miao, Shugui Liu, Qi Gu, Xunliang Cai, Xiang Wang, and An Zhang. Skill1: Unified evolution of skill-augmented agents via reinforcement learning. *arXiv preprint arXiv:2605.06130*, 2026.

Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. Visualpuzzles: Decoupling multimodal reasoning evaluation from domain knowledge. *arXiv preprint arXiv:2504.10342*, 2025.

Yubo Sun, Chunyi Peng, Yukun Yan, Shi Yu, Zhenghao Liu, Chi Chen, Zhiyuan Liu, and Maosong Sun. Visrag 2.0: Evidence-guided multi-image reasoning in visual retrieval-augmented generation. *arXiv preprint arXiv:2510.09733*, 2025.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *International Conference on Learning Representations*, volume 2025, pp. 62624–62650, 2025a.

Hao Wang, Guozhi Wang, Han Xiao, Yufeng Zhou, Yue Pan, Jichao Wang, Ke Xu, Yafei Wen, Xiaohu Ruan, Xiaoxin Chen, et al. Skill-sd: Skill-conditioned self-distillation for multi-turn llm agents. *arXiv preprint arXiv:2604.10674*, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Yongxian Wei, Runxi Cheng, Weike Jin, Enneng Yang, Li Shen, Lu Hou, Sinan Du, Chun Yuan, Xiaochun Cao, and Dacheng Tao. Optmerge: Unifying multimodal llm capabilities and modalities via model merging. In *International Conference on Learning Representations*, volume 2026, pp. 61571–61595, 2026.

Jiaer Xia, Yuhang Zang, Peng Gao, Sharon Li, and Kaiyang Zhou. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. *arXiv preprint arXiv:2505.14677*, 2025.

Peng Xia, Jianwen Chen, Hanyang Wang, Jiaqi Liu, Kaide Zeng, Yu Wang, Siwei Han, Yiyang Zhou, Xujiang Zhao, Haifeng Chen, et al. Skillrl: Evolving agents via recursive skill-augmented reinforcement learning. *arXiv preprint arXiv:2602.08234*, 2026.

Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, et al. Mimo-v2-flash technical report. *arXiv preprint arXiv:2601.02780*, 2026.

Xiaomi LLM-Core Team. MiMo-VL Technical Report. *arXiv preprint arXiv:2506.03569*, 2025.

Haidong Xin, Xinze Li, Zhenghao Liu, Yukun Yan, Shuo Wang, Cheng Yang, Yu Gu, Ge Yu, and Maosong Sun. Metamem: Evolving meta-memory for knowledge utilization through self-reflective symbolic optimization. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 5473–5492, 2026.

Yuqi Xiong, Chunyi Peng, Zhipeng Xu, Zhenghao Liu, Zulong Chen, Yukun Yan, Shuo Wang, Yu Gu, and Ge Yu. Lang2act: Fine-grained visual reasoning through self-emergent linguistic toolchains. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 8375–8399, 2026.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026a.

Xin Xu, Clive Bai, Kai Yang, Tianhao Chen, Yang Wang, Saiyong Yang, and Can Yang. Thinking-free policy initialization makes distilled reasoning models more effective and efficient reasoners. In *International Conference on Learning Representations*, volume 2026, pp. 24062–24086, 2026b.

Zhipeng Xu, Junhao Ji, Zulong Chen, Zhenghao Liu, Qing Liu, Chunyi Peng, Zubao Qin, Ze Xu, Jianqiang Wan, Jun Tang, et al. Cc-ocr v2: Benchmarking large multimodal models for literacy in real-world document processing. *arXiv preprint arXiv:2605.03903*, 2026c.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026.

Bob Zhang, Haoran Li, Tao Zhang, Jianan Li, Cilin Yan, Xikai Liu, Jiayin Cai, and Yanbin Hao. Improving the reasoning of multi-image grounding in mllms via reinforcement learning. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12667–12671. IEEE, 2026.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. EasyR1: An efficient, scalable, multi-modality RL training framework. `https://github.com/hiyouga/EasyR1`, 2025. GitHub repository.

A APPENDIX

A.1 DATASET DETAILS

We evaluate models on eight benchmarks spanning four complementary capability groups. Fine-grained visual perception is assessed using CountBenchQA (Paiss et al., 2023) and PixMo-Count (Deitke et al., 2025). OCR and chart understanding are evaluated on the recognition track of CC-OCR V2 (Xu et al., 2026c) and ChartQA (Masry et al., 2022). Symbolic visual reasoning is evaluated on VisualPuzzles (Song et al., 2025) and We-Math (Qiao et al., 2025). Multi-image reasoning is evaluated on MUIRBench (Wang et al., 2025a) and MMSI-Bench (Yang et al., 2025). Table 3 summarizes the evaluation scale and primary capability of each benchmark. The reported sample counts correspond to the instances retained by our evaluation pipeline after preprocessing.

Table 3: Statistics of the Benchmarks. The number of evaluation samples corresponds to the instances retained by our evaluation pipeline.

Capability	**Benchmark**	**#Eval.**	**Description**
Perception	CountBenchQA	491	Object counting in natural images
Perception	PixMo-Count	527	Fine-grained counting in diverse natural scenes
OCR & Chart	CC-OCR V2	1,790	Multilingual and natural-scene text recognition
OCR & Chart	ChartQA	2,500	Visual and logical reasoning over charts
Symbolic	VisualPuzzles	1,168	General visual and abstract puzzle reasoning
Symbolic	We-Math	1,740	Hierarchical visual mathematical reasoning
Multi-Image	MUIRBench	2,600	Robust reasoning over multiple related images
Multi-Image	MMSI-Bench	1,000	Multi-image spatial intelligence and reasoning

For capability-specific expert training, we construct the reinforcement learning corpus from six public multimodal data sources and organize the retained samples into four capability-specific subsets. The fine-grained perception and grounding capability uses PixMo-Count together with capability-aligned samples from Orsta-Data-47K (Ma et al., 2025). The OCR, chart, and document understanding capability uses SlideVQA together with corresponding samples from Orsta-Data-47K. The symbolic and mathematical reasoning capability uses We-Math 2.0 Pro together with corresponding samples from Orsta-Data-47K. The multi-image spatial reasoning capability uses Mantis-Instruct (Jiang et al., 2024) and M4-Instruct (Li et al., 2024). After preprocessing and filtering, the resulting training corpus contains 37,999 samples. Table 4 reports the data sources and sample counts for each capability.

Table 4: Statistics of the capability-specific reinforcement learning data. The reported counts correspond to the samples retained after preprocessing and filtering.

Capability	**Data Sources**	**# Samples**
Perception & Grounding	PixMo-Count, Orsta-Data-47K	9,499
OCR & Chart	SlideVQA, Orsta-Data-47K	9,500
Symbolic Reasoning	We-Math 2.0 Pro, Orsta-Data-47K	9,500
Multi-Image Reasoning	Mantis-Instruct, M4-Instruct	9,500
Total	–	**37,999**

A.2 ADDITIONAL IMPLEMENTATION DETAILS

For the main experiments, all capability-specialized experts and the unified student are initialized from Qwen3-VL-8B-Instruct (Bai et al., 2025a), while the additional model-scale experiments use Qwen3-VL-4B-Instruct under the same experimental setup. Capability-specialized expert training is implemented with EasyR1 (Zheng et al., 2025), while skill-decoupled on-policy expert distillation is implemented with veRL (Sheng et al., 2024). All experiments are conducted on NVIDIA

Table 5: Hyperparameters for capability-specialized expert training and skill-decoupled on-policy expert distillation.

(a) Capability-Specialized Expert Training

Hyperparameter	Value
Training epochs	3
Rollout batch size	128
Actor global batch size	64
Responses per prompt	8
Maximum prompt length	8,192
Maximum response length	8,192
Learning rate	1×10^{-6}
Weight decay	1×10^{-2}
Warmup ratio	0.0
Maximum gradient norm	1.0
KL coefficient	1×10^{-2}
Rollout temperature	1.0
Top-p	1.0
Tensor parallel size	2
Minimum image pixels	262,144
Maximum image pixels	524,288
Training precision	bfloat16

(b) Skill-Decoupled On-Policy Expert Distillation

Hyperparameter	Value
Training epochs	3
Rollout batch size	256
Actor mini-batch size	64
Responses per prompt	1
Maximum prompt length	12,288
Maximum response length	8,096
Learning rate	1×10^{-6}
Weight decay	1×10^{-2}
Warmup ratio	0.0
Maximum gradient norm	1.0
KL penalty	Disabled
Rollout temperature	1.0
Top-p	1.0
Student tensor parallel size	2
Teacher tensor parallel size	1
Distillation loss mode	`k_1`
Training precision	bfloat16

A800 80GB GPUs. Unless otherwise specified, the four capability experts share the same optimization and rollout configurations, differing only in their capability-specific training data and reward functions. During skill-decoupled on-policy expert distillation, the four independently optimized capability experts provide supervision to the unified student, with each expert conditioned on its corresponding capability skill while the student receives only the original multimodal input. The main hyperparameters for both capability-specialized expert training and skill-decoupled on-policy expert distillation are summarized in Table 5.

A.3 Reward Design

We design task-specific automatic rewards to accommodate the heterogeneous output spaces of the training tasks. For textual predictions, the answer is extracted from the `<answer>` field when available and normalized by lowercasing, removing punctuation and articles, and collapsing consecutive spaces. Structured predictions are additionally checked against their required output formats.

Fine-Grained Perception and Visual Grounding. For perception tasks with textual outputs, we use token-level F1 between the normalized prediction and reference answer. For visual grounding, the prediction is parsed as a list of labeled bounding boxes. Predicted and ground-truth boxes with the same normalized label are greedily matched according to their intersection-over-union (IoU). The grounding reward is defined as

$$r_{\text{ground}} = \lambda_{\text{fmt}} r_{\text{fmt}} + \lambda_{0.5} F_{1,\tau=0.5} + \lambda_{0.75} F_{1,\tau=0.75} + \lambda_{\text{IoU}} r_{\text{mIoU}}, \tag{10}$$

where r_{fmt} indicates whether the predicted boxes can be successfully parsed, $F_{1,\tau=0.5}$ and $F_{1,\tau=0.75}$ denote label-aware detection F1 at IoU thresholds of 0.5 and 0.75, respectively, and r_{mIoU} is the mean IoU obtained from one-to-one box matching. We set $(\lambda_{\text{fmt}}, \lambda_{0.5}, \lambda_{0.75}, \lambda_{\text{IoU}}) = (0.1, 0.4, 0.3, 0.2)$.

OCR and Chart Understanding. For OCR and chart-understanding tasks, we adopt token-level F1 as the training reward:

$$r_{\text{text}} = \frac{2PR}{P+R}, \tag{11}$$

where P and R are computed from the token overlap between the normalized prediction and reference answer. Boolean outputs are mapped to a unified *yes*/*no* representation before matching. Empty predictions or predictions without overlapping tokens receive zero reward.

Multimodal Symbolic Reasoning. For symbolic reasoning tasks, we employ format-gated answer accuracy:

$$r_{\text{symbolic}} = \mathbb{I}_{\text{fmt}} \ \text{Acc}(a, y), \tag{12}$$

Table 6: Ablation Study. All models are built on Qwen3-VL-4B. We compare training and capability integration strategies, with capability-specific experts shown as references for assessing expertise retention. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Perception		OCR & Chart		Symbolic		Multi-Image		Avg.
	CountB.	PixMo	CC-OCR	ChartQA	VisPuz.	WeMath	MUIR	MMSI	
Capability-Specific Experts									
Perception-RL	88.39	73.24	84.76	87.64	40.41	77.87	73.54	30.20	69.51
+ Perception Skill	89.41	75.33	84.79	87.72	40.33	77.82	73.65	30.30	69.92
OCR & Chart-RL	86.35	68.12	85.25	90.36	42.29	80.06	74.00	29.20	69.45
+ OCR & Chart Skill	86.15	68.31	85.46	91.28	42.38	80.17	73.88	29.10	69.59
Symbolic-RL	85.34	64.33	83.60	85.00	35.62	78.91	73.81	29.50	67.01
+ Symbolic Skill	85.13	64.14	83.57	85.24	36.04	80.34	73.85	29.60	67.24
Multi-Image-RL	79.63	59.96	84.25	86.64	41.44	75.29	75.65	30.60	66.68
+ Multi-Image Skill	79.84	60.15	84.22	86.60	41.35	75.23	76.15	31.50	66.88
Unified-Model Baselines and Ablations									
Zero-shot	81.47	64.90	82.86	85.60	35.10	76.26	72.12	**30.20**	66.06
SFT	63.54	44.40	74.67	65.36	31.93	46.32	61.08	24.80	51.51
Mix-RL	86.35	70.97	83.10	87.60	34.76	**80.29**	69.92	29.10	67.76
SCOPED (Parameter Merging)	86.35	71.16	84.49	86.60	37.93	76.95	73.88	26.60	68.00
SCOPED (w/o Skill Augmentation)	87.98	73.81	84.74	88.96	37.50	78.16	75.69	29.20	69.51
SCOPED	**88.39**	**74.38**	**85.03**	**89.48**	**38.01**	78.85	**76.15**	29.40	**69.96**

where $\mathbb{I}_{\text{fmt}}$ indicates whether the response contains a valid `<answer>...</answer>` field. Numerical answers are compared by exact numerical equality after removing commas and optional percentage symbols. Textual answers are normalized before matching, and a prediction is considered correct if it matches any valid reference answer. Responses that do not satisfy the required answer format receive zero reward.

Multi-Image Spatial Reasoning. Multi-image reasoning tasks also use format-gated accuracy:

$$r_{\text{multi}} = \mathbb{I}_{\text{fmt}} \ \text{Acc}_{\text{type}}(a, y), \tag{13}$$

where the matching function is selected according to the answer type. Multiple-choice questions require an exact option-letter match, while image-selection questions compare the referenced image indices. Yes/no expressions are mapped to canonical labels, and spatial descriptions are mapped to normalized relations such as *left*, *above*, *inside*, and *not touching*. Remaining free-form answers are evaluated using normalized containment matching. As in symbolic reasoning, malformed responses receive zero overall reward.

A.4 Additional Ablation on Model Scale

We further conduct the same ablation study on Qwen3-VL-4B to examine whether the effectiveness of our design remains consistent at a smaller model scale. As shown in Table 6, SFT substantially underperforms the zero-shot backbone, while Mix-RL yields only a modest overall improvement and remains below the corresponding capability-specific experts on several target capabilities. In contrast, independently trained experts achieve stronger performance on their respective capabilities, showing that the benefit of separating capability acquisition from integration remains evident with a smaller backbone.

The skill-augmented expert results exhibit a trend consistent with Qwen3-VL-8B. Conditioning each expert on its matched skill improves performance on the corresponding target capability, while the effects on non-target capabilities remain limited or mixed. This further supports the capability-specific nature of the extracted guidance and the use of explicit skill-expert matching during distillation.

For capability integration, we compare SCOPED (Parameter Merging), SCOPED w/o Skill Augmentation, and the full SCOPED framework using the same capability-specific experts. SCOPED

Table 7: Ablation of trajectory sources used for capability skill abstraction on Qwen3-VL-8B. The baseline reports the absolute performance of the corresponding capability-specialized experts, while the remaining rows report performance changes relative to the baseline.

Condition	Perception		OCR & Chart		Symbolic		Multi-Image		Avg.
	CountB.	PixMo	CC-OCR	ChartQA	VisPuz.	WeMath	MUIR	MMSI	
Baseline									
Capability-RL Expert	92.67	79.89	86.08	89.64	48.37	82.30	76.00	30.70	73.21
Skill Abstraction (delta from Capability-RL Expert)									
+ Failure	+0.41	+1.51	+0.07	+0.56	+0.17	+0.75	+0.35	+0.60	+0.55
+ Success	+0.61	+1.13	+0.09	+0.48	+0.26	+1.15	+0.19	+0.80	+0.59
+ Combined	**+1.02**	**+2.46**	**+0.15**	**+0.96**	**+0.43**	**+1.90**	**+0.50**	**+1.30**	**+1.09**

w/o Skill Augmentation outperforms parameter merging, showing that on-policy expert supervision provides a more effective mechanism for transferring independently acquired capabilities than direct parameter-space merging. Incorporating skill augmentation further improves the unified model, yielding the best average performance among the compared integration variants. Together with the Qwen3-VL-8B results, these consistent trends indicate that the effectiveness of SCOPED is not restricted to a particular backbone scale.

A.5 Skill Construction Details

We provide additional implementation details for the capability-wise skill abstraction procedure introduced in Sec. 3.2. Skill construction uses only training data and training-set rollouts, while the held-out evaluation set is used exclusively for final evaluation. For each capability, a single skill is constructed from the trajectories generated by its corresponding optimized expert.

Trajectory Curation. For each training instance, the optimized capability expert samples multiple rollout trajectories, each of which is evaluated by the corresponding capability-specific reward function. To reduce redundancy, we deduplicate repeated trajectories within each task subtype and balance the retained trajectories across subtypes so that high-resource tasks do not dominate the subsequent skill construction process.

Reward-Based Trajectory Grouping. Following Sec. 3.2, the curated trajectories are divided into successful and failed cohorts according to their capability-specific rewards. This grouping provides complementary evidence about effective solution patterns and recurring failure modes.

Trajectory-Level Reflection. The optimized capability expert performs trajectory-level reflection over the successful and failed cohorts using outcome-specific reflection prompts. Successful trajectories are used to identify reusable solution strategies and verification behaviors, while failed trajectories are analyzed to diagnose recurring failure mechanisms and derive corrective guidance. The resulting reflections are aggregated into a capability-specific evidence ledger containing both effective behaviors and failure-avoidance guidance.

Capability Skill Synthesis. Based on the aggregated trajectory-derived evidence, the same capability expert synthesizes a single capability-level skill containing reusable procedural guidance for the corresponding capability. The resulting skill is subsequently used to condition the expert during skill-decoupled on-policy expert distillation. This procedure produces one skill for each of the four capabilities. Detailed reflection and skill-induction prompts are provided in Appendix A.10.

Ablation of Trajectory Sources. Table 7 examines the contribution of successful and failed trajectories to skill construction. Using either successful or failed trajectories alone improves the corresponding capability experts, while combining both sources achieves the strongest overall performance. This indicates that successful solution patterns and corrective failure experience provide complementary information for capability skill abstraction.

A.6 Details of Cross-Capability Gradient Analysis

We provide the computation protocol for the two gradient-based analyses in Figure 2. Figure 2b examines whether the realized Mix-RL update follows the directions induced by independently trained capability experts, while Figure 2c directly measures pairwise conflicts among capability-specific local optimization directions under the same Mix-RL parameters.

Expert-Guided Update Alignment. For each capability c, we fix the independently trained capability expert π_{θ_c} and select a fixed set of 32 samples. At a Mix-RL checkpoint θ_t, both the current Mix-RL policy and the corresponding expert are evaluated under identical response prefixes. Let $q_{c,n,s}$ and $p_{\theta_t,c,n,s}$ denote the expert and Mix-RL predictive distributions, respectively, at token position s. We define the expert-matching objective as

$$L_c^{\exp}(\theta_t) = \frac{1}{N_c}\sum_{n=1}^{N_c}\frac{1}{T_{c,n}}\sum_{s=1}^{T_{c,n}} D_{\mathrm{KL}}(q_{c,n,s}\,\|\,p_{\theta_t,c,n,s})\,, \tag{14}$$

where $N_c = 32$ and $T_{c,n}$ denotes the number of valid response positions for sample n. The corresponding expert-guided direction is defined as the negative gradient

$$\mathbf{g}_c^{\exp}(\theta_t) = -\nabla_\theta L_c^{\exp}(\theta)\big|_{\theta=\theta_t}. \tag{15}$$

Given two adjacent Mix-RL checkpoints $t < t'$, the realized parameter displacement is

$$\Delta\theta_{t\to t'} = \theta_{t'} - \theta_t. \tag{16}$$

We measure whether this realized update follows the corresponding expert-guided direction using

$$A_c(t\to t') = \frac{\langle \Delta\theta_{t\to t'}, \mathbf{g}_c^{\exp}(\theta_t)\rangle}{\|\Delta\theta_{t\to t'}\|_2\,\|\mathbf{g}_c^{\exp}(\theta_t)\|_2}. \tag{17}$$

All four capabilities are compared against the same Mix-RL parameter displacement within each checkpoint interval. Positive values indicate that the realized update moves toward the corresponding expert behavior, whereas negative values indicate movement against the expert-guided direction. For Figure 2b, checkpoint-level alignments are grouped into four training stages and summarized within each capability.

Reward-Based Capability Optimization Direction. To directly analyze interactions among capability objectives, we separately estimate a reward-based local optimization direction for each capability under the same Mix-RL parameters. At checkpoint θ_t, we sample K responses $\{y_{c,n}^{(k)}\}_{k=1}^K$ for each input $x_{c,n}$ from capability c and obtain their rewards $\{r_{c,n}^{(k)}\}_{k=1}^K$. The group-normalized advantage is

$$A_{c,n}^{(k)} = \frac{r_{c,n}^{(k)} - \bar{r}_{c,n}}{\sigma_{c,n}+\epsilon}, \tag{18}$$

where $\bar{r}_{c,n}$ and $\sigma_{c,n}$ are the mean and population standard deviation of the K rewards, respectively, and $\epsilon = 10^{-6}$. Groups with $\sigma_{c,n} < \epsilon$ are discarded. Let $\mathcal{V}_{c,t}$ denote the remaining valid responses and $N_{c,t}$ their total token count. The resulting local optimization direction is

$$\mathbf{u}_c(\theta_t) = \frac{1}{N_{c,t}}\sum_{(n,k)\in\mathcal{V}_{c,t}} A_{c,n}^{(k)}\sum_{s=1}^{|y_{c,n}^{(k)}|}\nabla_\theta \log \pi_\theta\Big(y_{c,n,s}^{(k)} \mid x_{c,n}, y_{c,n,<s}^{(k)}\Big)\Big|_{\theta=\theta_t}. \tag{19}$$

Gradients are computed over the language-policy parameters, with the visual parameters frozen.

Pairwise Cross-Capability Conflict. At each Mix-RL checkpoint, all capability-specific directions are estimated under the same shared parameters $\theta_{\mathrm{mix},t}$. For each unordered capability pair (c, c'), we compute

$$C_{c,c'}(t) = \frac{\langle \mathbf{u}_c(\theta_{\mathrm{mix},t}), \mathbf{u}_{c'}(\theta_{\mathrm{mix},t})\rangle}{\|\mathbf{u}_c(\theta_{\mathrm{mix},t})\|_2\,\|\mathbf{u}_{c'}(\theta_{\mathrm{mix},t})\|_2}. \tag{20}$$

With four capabilities, this yields six unordered capability pairs at each checkpoint. Negative cosine similarity indicates locally opposing optimization directions and therefore provides direct evidence of cross-capability gradient conflict. For Figure 2c, the checkpoint-level cosine similarities are grouped into four training stages, and the mean similarity of each capability pair within each stage is reported.

A.7 Details of Expert-Retention Analysis

We provide the computation protocol for the expert-retention analysis in Figure 3a. For each capability, the independently trained capability expert without additional skill conditioning is used as the reference. All comparisons are performed using fixed-trajectory teacher forcing so that different unified models are evaluated under identical response states.

For each input x_i, the base policy π_{base} greedily generates a trajectory

$$\tau_i = (\tau_{i,1}, \ldots, \tau_{i,T_i}), \qquad h_{i,t} = (x_i, \tau_{i,<t}), \tag{21}$$

and all compared policies are evaluated on the same states $\{h_{i,t}\}_{t=1}^{T_i}$. This avoids confounding functional differences with changes in generated trajectories.

Functional Expert Retention. For any policy π, we represent its functional deviation from the base policy along τ_i using the observed-token log-probability changes:

$$\mathbf{a}_{\pi}^{(i)} = (\log \pi(\tau_{i,t} \mid h_{i,t}) - \log \pi_{\text{base}}(\tau_{i,t} \mid h_{i,t}))_{t=1}^{T_i}. \tag{22}$$

Let $\mathbf{a}_c^{(i)}$ denote the vector produced by the corresponding capability expert π_{θ_c}, and let $\mathbf{a}_\theta^{(i)}$ denote that of a unified model π_θ.

We measure directional agreement with the capability expert using cosine similarity:

$$C_i(\theta, c) = \frac{\left\langle \mathbf{a}_\theta^{(i)}, \mathbf{a}_c^{(i)} \right\rangle}{\left\| \mathbf{a}_\theta^{(i)} \right\|_2 \left\| \mathbf{a}_c^{(i)} \right\|_2}. \tag{23}$$

A larger C_i indicates that the functional change induced by the unified model is more closely aligned with that of the corresponding capability expert.

We further measure the projected amplitude retained along the expert direction:

$$R_i(\theta, c) = \frac{\left\langle \mathbf{a}_\theta^{(i)}, \mathbf{a}_c^{(i)} \right\rangle}{\left\| \mathbf{a}_c^{(i)} \right\|_2^2}. \tag{24}$$

Here, $R_i = 1$ indicates that the unified model preserves the same projected magnitude as the expert, $0 < R_i < 1$ indicates attenuation, $R_i > 1$ indicates amplification, and $R_i < 0$ indicates an opposing component.

Both metrics are first averaged over samples within each capability and then macro-averaged across the four capabilities to obtain the results reported in Figure 3a. These metrics quantify functional agreement with independently trained capability experts under identical trajectory states rather than directly measuring downstream task accuracy.

A.8 Qualitative Analysis of Skill-Guided Visual Attention

To complement the quantitative attention analysis in Sec. 5.4, we further visualize how skill conditioning affects the visual attention of capability experts on a representative example.



(a) Target objects (b) Baseline attention (c) Skill-conditioned attention (d) Skill-induced attention change

Figure 4: Qualitative visualization of skill-induced attention reallocation. Green boxes indicate target objects. Red and blue in the attention-change map indicate increased and decreased attention after skill conditioning.

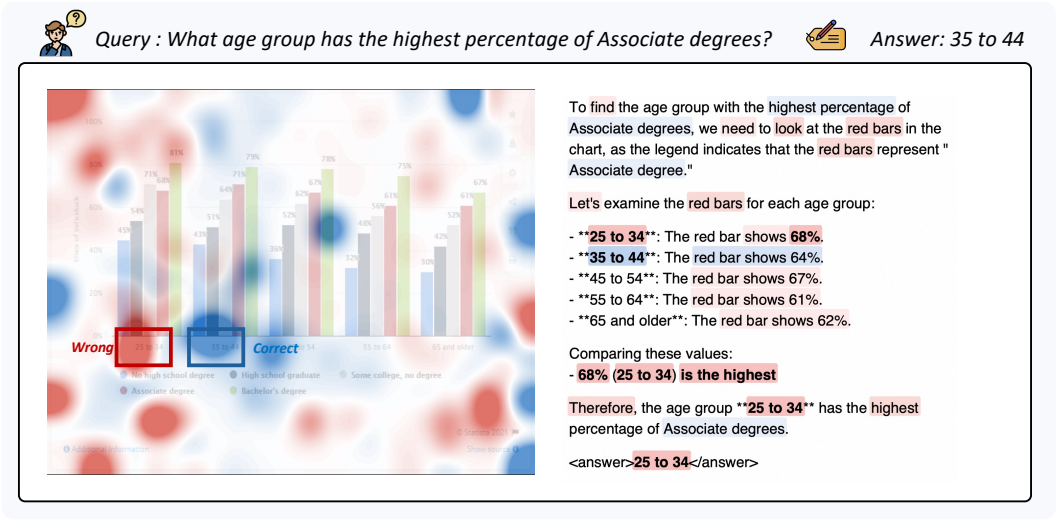


Figure 5: Qualitative effect of skill-augmented expert supervision. Compared with SCOPED w/o Skill Augmentation, SCOPED strengthens responses supporting the correct answer while suppressing those associated with the erroneous trajectory. Blue indicates enhanced responses, while red indicates suppressed responses.

As illustrated in Figure 4, we compare the same capability expert with and without its matched skill under an identical input. Without skill conditioning, the expert distributes substantial attention over both task-relevant and irrelevant visual regions. In contrast, conditioning on the matched skill shifts more attention toward the visual evidence relevant to the current task, as further highlighted by the attention-difference map. This qualitative example is consistent with the layer-wise attention analysis in Figure 3c, providing an intuitive illustration of how skill conditioning can regularize expert visual attention toward task-relevant evidence.

A.9 Case Studies

To better understand how skill-augmented distillation changes the behavior of the unified model, we present a representative example in Figure 5. The task requires identifying the age group with the highest percentage of Associate degrees. While SCOPED w/o Skill Augmentation follows an erroneous reasoning trajectory and predicts “25 to 34”, the full SCOPED model correctly identifies “35 to 44”.

The visualization further illustrates how skill-augmented supervision affects the model’s use of visual evidence and token-level predictions. Compared with SCOPED w/o Skill Augmentation, SCOPED allocates more attention to the region supporting the correct answer while reducing attention to the misleading region associated with the incorrect prediction. At the token level, SCOPED also suppresses responses along the erroneous trajectory and strengthens those supporting the correct prediction.

This case complements the quantitative analyses in Section 5.4, showing that skill-augmented expert supervision can lead the unified model to rely more on task-relevant visual evidence and avoid erroneous reasoning patterns during generation.

We further present representative examples from PixMo-Count, MUIRBench, and MMSI-Bench to examine whether these improvements are reflected across different visual capabilities. On PixMo-Count, the Qwen3-VL-Instruct backbone overlooks one aircraft and predicts six, whereas SCOPED systematically scans and verifies individual instances before correctly identifying all seven aircraft. On MUIRBench, the backbone is distracted by appearance differences between maps, while SCOPED instead compares stable structural and spatial cues and correctly identifies the corresponding image. Similarly, on MMSI-Bench, SCOPED tracks the relative displacement of persistent visual landmarks across images and correctly infers the movement direction, whereas the backbone reaches an incorrect conclusion from a coarser comparison of the scenes.

Query : How many airplanes are there in the image?

Answer: 7



Qwen3-VL-Instruct

To determine the number of airplanes in the image, we can count them individually.
1. There is one airplane at the top of the image, flying alone and leaving a blue smoke trail.
2. Below that, there is a formation of airplanes. We can count them:
   - One airplane on the far left, near the bottom.
   - One airplane to its right, slightly higher.
   - A group of three airplanes in the center-right, flying in a staggered formation.

Counting them: 1 (top) + 1 (bottom left) + 1 (bottom center) + 3 (center-right) = 6 airplanes.
Alternatively, we can count the visible aircraft in the formation:
- The top airplane: 1
- The formation below it: 5 airplanes (one on the far left, one to its right, and three in a line to the right of that).

1 + 5 = 6.
<answer>6</answer>

SCOPED

To determine the number of airplanes in the image, we will carefully scan the entire scene and count each visible aircraft.
1. **Top of the image**: There is one airplane flying alone at the top, slightly to the right, leaving a blue smoke trail. It is clearly visible and distinct.
   → Count: 1
2. **Middle-left**: Below the top airplane, there is a group of airplanes flying in formation.
   - One airplane is on the lower left, flying slightly downward.
     → Count: 2
   - Another airplane is to the right of it, slightly higher, with a visible number "5" on its tail.
     → Count: 3
   - Above that, slightly to the right, is a third airplane in the formation, with a visible number "8" on its tail.
     → Count: 4
   - To the right of that, two more airplanes are flying in a staggered formation —one slightly above and to the right, and another below it.
     → Count: 5 and 6
3. **Right side**: One more airplane is visible on the far right, flying slightly above the lower two in the formation.
   → Count: 7

Let's recount to ensure no aircraft is missed:
- Top right: 1 airplane (smoke trail) → ✅
- Lower left: 1 airplane (bottom left) → ✅
- Just above it: 1 airplane (with "5" on tail) → ✅
- Above that: 1 airplane (with "8" on tail) → ✅
- To the right of the "8" plane: 1 airplane (slightly lower) → ✅
- To the right of that: 1 airplane (slightly higher) → ✅
- Far right: 1 airplane (separate from the main group) → ✅

Total count: 7 airplanes.
All airplanes are clearly visible, even if partially obscured by the sky or other planes. No airplane is duplicated or hidden.
<answer>7</answer>

Figure 6: Case study of PixMo-Count.

Together, these examples illustrate that the gains of SCOPED extend across fine-grained perception, cross-image correspondence, and multi-image spatial reasoning. Rather than relying on superficial visual patterns, the unified model exhibits more systematic evidence aggregation and verification, consistent with the quantitative improvements observed across the evaluated capability groups.

Query : Identify the map image that shows overlapping geographic areas similar to Image 1. *Answer: C*



Image1 Image2 Image3 Image4

Options: A. None of the choices provided. B. Image 2. C. Image 3. D. Image 4

Qwen3-VL-Instruct

To solve this, we need to identify which of the four images shows overlapping geographic areas similar to Image 1.
- Image 1 displays a topographic map with distinct, colored areas representing different land cover or geological units. For example, there are areas labeled "Qt" (quartzite), "Qs" (sandstone), "Qsl" (sandstone with limestone), and "af" (alluvium). These areas are shown with different shading and are clearly delineated, indicating they are distinct geographic regions that overlap or are adjacent to each other. The map also shows a road (Deacon Lane) and a stream (Rocky Brook) cutting through these areas.
- Image 2 shows a topographic map with contour lines and elevation numbers (e.g., 6500, 6904). It depicts a river and surrounding terrain, but there are no distinct, colored, overlapping geographic areas like in Image 1. The focus is on elevation and terrain shape, not on different land cover types.
- Image 3 is a black-and-white topographic map. It shows contour lines, a road, and a stream, but it lacks the colored overlays that define different geographic areas. The land is shown in shades of gray, not distinct colored regions.
- Image 4 shows a coastal area with contour lines for water depth (e.g., 25, 26, 27 fathoms) and a land area with a different color. It shows a river and a landform labeled "Qd". While it has a distinct land/water boundary and a different color for the land, it does not show multiple, overlapping geographic areas with different colors and labels like Image 1.
Therefore, none of the other images (2, 3, or 4) show the same kind of overlapping geographic areas with distinct colored regions as Image 1.
<answer>A</answer>

SCOPED

To determine which map image shows overlapping geographic areas similar to Image 1, we must first understand the key geographic features and spatial relationships in Image 1 and then compare them with the other images.
Image 1 displays a topographic map section with:
- A labeled area “PROCTOR” near the center.
- A red outlined area (possibly a property or land parcel) around Proctor.
- A stream labeled “Rocky Brook” flowing from the upper left to the lower right.
- A pond labeled “Rocky Pond” to the upper left of Proctor.
- A road labeled “DEACON LANE” to the upper right.
- Contour lines indicating elevation (e.g., 350, 398, 414).
- A legend indicating “Qt” (quaternary till) and “Qs” (quaternary sand and gravel) in different areas.
- A scale reference on the left: “MILFORD (VIA N.H. 13) 11 MI.” and “BROOKLINE 2.8 MI.”.
Now, evaluate the other images:
Image 2: This is a close-up of a topographic map showing a river, labeled “Kirkham” and “Hollow,” with contour lines indicating high elevations (6500, 6904). It does not show any of the features from Image 1 — no Proctor, no Rocky Brook, no Deacon Lane, no Rocky Pond, or the same scale or legend. The geographic context is entirely different (higher elevation, different names, no roads or ponds matching Image 1). Therefore, it does not show overlapping geographic areas similar to Image 1.
Image 3: This image is a black-and-white version of Image 1, with contour lines and labels (Proctor, Rocky Brook, Deacon Lane, Rocky Pond) preserved. The layout, features, and spatial relationships are identical to Image 1. The only difference is the color scheme and absence of the colored geological legend (Qt, Qs, etc.) and scale text on the left. However, the geographic areas — including the red outlined area around Proctor, the position of Rocky Brook, the location of Deacon Lane, and the contour lines — are exactly the same. This image shows overlapping geographic areas similar to Image 1 — in fact, it is a re-rendering or grayscale version of Image 1.
Therefore, Image 3 matches the criteria.
Image 4: This image shows a river labeled “RIVER” with contour lines near 25–36, and a landform labeled “Qd” (quaternary drift) near the riverbank. It has no connection to Proctor, Rocky Brook, Deacon Lane, or Rocky Pond. The scale and geographic context are completely different. It does not show overlapping geographic areas similar to Image 1.
Therefore, the only image that shows overlapping geographic areas similar to Image 1 is Image 3.
The correct option is C. Image 3.
<answer>C</answer>

Figure 7: Case study of MUIRBench.

Query : The images are taken continuously from a first-person perspective. In which direction are you moving?

Answer: C

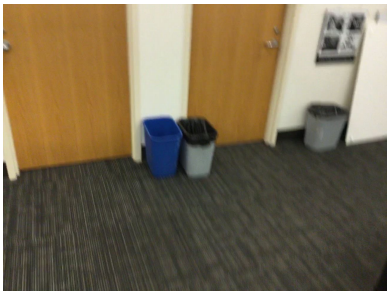


Image1

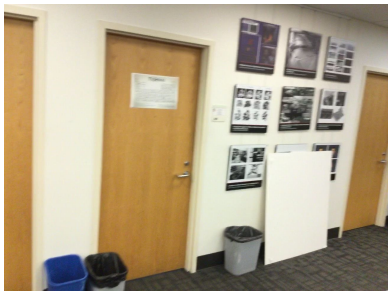


Image2

Options: A: Left while moving backward, B: Forward to the left, C: Forward to the right, D: Right while moving backward

Qwen3-VL-Instruct

The first image shows a narrow hallway with two doors and two trash cans in the foreground. The second image shows a wider hallway with a door, several posters on the wall, and a whiteboard. The perspective in the second image is more open and reveals more of the wall and floor. This suggests that the camera is moving forward, revealing more of the space ahead. The change from a narrow view to a wider view indicates forward movement, not lateral movement. Therefore, the direction of movement is forward.
<answer>B</answer>

SCOPED

To determine the direction of movement from the first-person perspective, we compare the two images.
In the first image, we see:
- Two wooden doors on the left and right.
- A blue trash can and a gray trash can in front of the left door.
- A gray trash can further to the right, near the right door.
- A whiteboard and several posters on the wall to the right of the center.
- The camera is positioned at a slight angle, capturing the scene from the left side of the hallway.

In the second image, we see:
- The same two doors, but now the camera is closer to the center door.
- The blue and gray trash cans are still visible, but now they are more to the left of the frame.
- The whiteboard and posters are more prominent and appear larger, indicating the camera is moving closer to them.
- The right door is now partially visible on the far right, and the wall with posters is more centered in the frame.

The key observation is that the camera has moved forward (closer to the wall with posters and the whiteboard) and slightly to the right. The objects on the right (posters, whiteboard, right door) appear larger and more centered, while the left door and trash cans are now slightly to the left of the frame. This indicates the camera moved forward and to the right.
Therefore, the correct direction is: Forward to the right.
<answer>C</answer>

Figure 8: Case study of MMSI-Bench.

A.10 PROMPT EXAMPLES

We provide the prompt templates used for both skill construction and evaluation.

Skill Construction Prompts. Our skill construction procedure uses three main prompts. First, the successful-trajectory self-reflection prompt asks the optimized capability expert to identify reusable verification behaviors and their activation conditions from its successful trajectories. Second, the failed-trajectory self-reflection prompt asks the same expert to diagnose causal failure mechanisms and derive corresponding corrective principles. Finally, the evidence-grounded skill induction prompt provides the curated evidence ledger to the same expert and asks it to consolidate recurring success and failure patterns into a compact capability-level skill. The corresponding prompt templates are shown in Figures 9–11.

Evaluation Prompts. Figure 12 presents the vanilla evaluation prompt, where the model receives the input image(s), the benchmark-specific query, and the corresponding response instruction. The model is instructed to reason step by step and place its final response within `<answer>` and `</answer>` tags. The required output format is instantiated according to the benchmark: multiple-choice tasks require the final option letter, OCR tasks follow the official per-sample recognition instructions.

For evaluating skill-augmented capability experts in the ablation and analysis experiments, we preserve the original multimodal input and response instruction and condition each expert on its corresponding capability skill. The four extracted skills correspond to perception, OCR and chart understanding, symbolic reasoning, and multi-image reasoning, respectively. The unified student does not receive explicit skill conditioning during inference.

Successful-Trajectory Self-Reflection Prompt

Role	You are an expert success-pattern analyst for {domain} tasks.
Input	A successful trajectory generated by a strong domain expert, together with the task query and ground-truth target.
Objective	Identify a lightweight verification or self-check principle that contributed to the successful response.
Analysis	Identify (1) the effective self-check or verification behavior, (2) the subtle cue that could otherwise cause an error, and (3) the condition under which this check should be applied.
Constraints	Do not prescribe a step-by-step solving procedure, force a chain-of-thought format, copy instance-specific answers, or introduce dataset-specific details. Prefer conditional verification guidance that is activated only when necessary.
Required Output	
Format	Return a valid JSON object containing a concise reusable skill rule and its activation condition.
Task	{question}
Expert Trajectory	{trajectory}
Ground Truth	{ground_truth}

Figure 9: Prompt used by the optimized capability expert to extract a reusable self-check principle from a successful trajectory.

Failed-Trajectory Self-Reflection Prompt

Role	You are an expert failure-analysis agent for {domain} tasks.
Input	A failed expert trajectory together with the task query, predicted answer, and ground-truth target.
Objective	Diagnose the underlying cause of failure and convert it into reusable corrective guidance.
Analysis	Identify (1) the mismatch between the prediction and ground truth, (2) the point at which the reasoning failed, (3) the underlying failure mechanism, and (4) a general corrective principle and its activation condition.
Constraints	Derive general principles rather than instance-specific corrections. Do not copy sample-specific facts, dataset names, numerical values, or task-specific surface details.
Required Output	
Format	Return a valid JSON object describing the failure cause, corrective rule, and condition under which the rule should be applied.
Task	{question}
Expert Trajectory	{trajectory}
Final Answer	{prediction}
Ground Truth	{ground_truth}

Figure 10: Prompt used by the optimized capability expert to diagnose a failed trajectory and extract reusable corrective guidance.

Evidence-Grounded Capability Skill Induction Prompt

Role	You are improving an expert vision-language model using evidence distilled from its own training trajectories.
Input	{`Evidence Ledgers`} containing successful and failed trajectory patterns for the current domain.
Domain Guidance	{`Domain-Specific Guidance`} specifying how the observed failure mechanisms should transfer across tasks within the domain.
Objective	Convert recurring success and failure patterns into a compact domain-level skill that provides reusable verification and corrective guidance.
Rule Form	Each task-specific rule should follow: *"When [specific trigger], [verification or corrective action]."*
Task Routing	Organize rules by task section so that only the checks relevant to the current question type are applied.
Grounding	Ground each task section in supporting failure evidence and, when available, contrasting successful evidence.
Constraints	Do not copy instance-specific answers, values, coordinates, or memorized examples. Avoid unsupported rules, unnecessary step-by-step procedures, and instructions that suppress reasoning. Prefer a small number of precise, conditionally activated rules.
Required Output	
Objective	A concise objective for the domain skill.
Evidence Protocol	A small set of shared rules applicable across tasks in the domain.
Task-Specific Checks	For each task section, a small set of trigger-conditioned verification or corrective rules.
Final Verification	A small set of final checks applied before producing the answer.

Figure 11: Prompt used by the optimized capability expert to induce a domain-level skill from curated trajectory-derived evidence. The resulting skill contains shared guidance together with conditionally activated task-specific checks.

Vanilla Evaluation Prompt

Input	{`Input Image(s)`}
Task Query	{`Dataset-Specific Question`}
Response Instruction	Please reason step by step, and put your final answer within `<answer>` and `</answer>`.
Task-Specific Output Formats	
Multiple Choice	Put only the final option letter within `<answer>` and `</answer>`.
OCR Recognition	Follow the official per-sample recognition instruction provided by the benchmark.

Figure 12: Vanilla prompt template used for evaluation. The query and response format are instantiated for each benchmark.

Capability Skill: Fine-Grained Perception

Objective
Answer the current benchmark item accurately from the provided image(s) and question.

Evidence Protocol

- Apply ONLY the task section matching the current question type; ignore the others.
- When the task involves sequential operations, update the working state after each step before proceeding.
- Do not use dataset names, filenames, memorized examples, or external information.

Task-Specific Checks
Apply when the question asks how many instances of a specific object/category appear.

- When a candidate instance is faint, translucent, partially cropped at the image edge, or only a tiny fragment is visible, exclude it from the count—only count objects that are clearly and substantially present.
- When the noun could refer to either a complete assembled item or its sub-components, count the most natural standalone unit that the noun refers to in everyday language.

Final verification:

- When the task involved subtraction or modification, confirm the final result reflects all operations applied in order.
- When your count includes a faint, translucent, or barely visible instance and you are uncertain, re-examine whether that instance is real before including it.

Apply when counting objects in natural photographs.

- When counting in a natural photograph, recount the entire scene after your first read—the dominant error is omission of partially hidden or edge-proximal instances.
- When spatial density or layout ambiguity arises, check for hidden or partial objects—include partially visible objects if context and boundaries support that they are real instances.
- When the count feels complete, do one final sweep of low-contrast and peripheral areas.
- When the target category could be counted as a whole set or as individual pieces, decide which unit the question asks for before counting and apply it consistently.
- When your first pass finds zero instances, run a second pass assuming the category may appear in an unusual, stylized, or partially occluded form before concluding it is absent.

Apply when locating a single object described by a referring expression.

- Locate exactly one object matching the full phrase. Interpret spatial terms from the viewer's perspective.
- When multiple same-category candidates exist, verify that the chosen object satisfies every word of the referring expression, especially spatial qualifiers.
- For ordinal expressions such as "2nd from left," count from the edge explicitly named in the phrase.

Output
Follow the output format stated in the question exactly, including any answer tags it specifies. Do not shorten, skip, or reorder the reasoning the question asks for, and never state the final answer before the checks above are done.

Figure 13: Extracted Capability skill for fine-grained perception.

Capability Skill: OCR and Chart Understanding

Objective
Answer the current benchmark item accurately from the provided image(s) and question.

Evidence Protocol

- Apply ONLY the task section matching the current question type; ignore the others.
- When the task involves sequential operations, update the working state after each step before proceeding.
- Do not use dataset names, filenames, memorized examples, or external information.

Task-Specific Checks
Apply when answering questions about charts.

- When reading values from a chart, align with the nearest labeled gridline or tick mark—do not estimate between lines.
- When comparing two values, confirm whether the question asks for an absolute difference or a ratio.
- When a chart has dual axes or multiple series, verify that the value is read from the correct axis and series.
- When identifying a peak or extremum, cross-reference its position with both axes to confirm the exact value.

Final verification:

- When the answer's magnitude seems implausible given the chart's visual scale, re-read the value against the axis markings.
- When transcribing text, verify that no visible text region was skipped by scanning page edges and corners.

Apply when answering questions about presentation slide decks.

- When extracting data from a multi-page document, verify that the answer comes from the page matching the question's context rather than a visually similar page.
- When the question asks about a relationship or flow, trace connecting arrows or lines in the diagram before answering.
- When listing categories from a chart or table, ensure that all items are captured.

Final verification:

- When the answer's magnitude seems implausible given the visual scale, re-check the relevant values.
- When transcribing text, verify that no visible text region was skipped.

Apply when the task is pure text recognition/OCR. Activate only for long multi-line text, multi-column layouts, or peripheral text blocks that could be overlooked.

- Scan every region systematically, including the main body, headers, footers, captions, footnotes, watermarks, stamps, and table cells.
- For multi-column or multi-region layouts, transcribe each region in natural reading order and preserve completeness.
- For long CJK text, read each line fully before moving to the next line; do not merge lines or skip characters.
- When uncertain whether a faint or partial region contains real text, prefer preserving potentially valid text rather than omitting an entire region.

Final verification:

- Re-scan the four edges and every column or region of the image and confirm that no visible text line was skipped or truncated.

Output
Follow the output format stated in the question exactly, including any answer tags it specifies. Do not shorten, skip, or reorder the reasoning the question asks for, and never state the final answer before the checks above are done.

Figure 14: Extracted Capability skill for OCR, document, and chart understanding.

Capability Skill: Multimodal Symbolic Reasoning

Objective
Answer the current benchmark item accurately from the provided image(s) and question.

Evidence Protocol

- Apply ONLY the task section matching the current question type; ignore the others.
- When the task involves sequential operations, update the working state after each step before proceeding.
- Do not use dataset names, filenames, memorized examples, or external information.

Task-Specific Checks
Apply when solving geometry problems with diagrams.

- When a derivation yields a result, verify that it is consistent with the diagram's visible proportions; if the computed length or angle is implausible, re-check the derivation and arithmetic.
- For multi-step derivations, solve each intermediate quantity explicitly and carry it forward to avoid propagating arithmetic errors.

Final verification:

- Confirm that the answer is a bare numeric value, with no option letter or units appended.

Apply when solving visual pattern or inductive puzzles.

- For inductive pattern or sequence puzzles, identify the simplest visual rule consistent with all given figures—such as rotation, color or shape cycles, or element addition/removal—and apply it forward.
- For other puzzle types, rely on direct visual inspection or straightforward reasoning rather than unnecessarily applying pattern-detection rules.

Apply when solving mathematical word problems requiring multi-step computation.

- When choosing among options, compute the answer independently first and then match it to the corresponding option; do not let the answer choices bias the derivation.

Final verification:

- Confirm that the final answer is an option letter (A, B, C, or D).

Output
Follow the output format stated in the question exactly, including any answer tags it specifies. Do not shorten, skip, or reorder the reasoning the question asks for, and never state the final answer before the checks above are done.

Figure 15: Extracted Capability skill for multimodal symbolic reasoning.

Capability Skill: Multi-Image Reasoning

Objective
Answer the current benchmark item accurately from the provided image(s) and question.

Evidence Protocol

- Apply ONLY the task section matching the current question type; ignore the others.
- When the task involves sequential operations, update the working state after each step before proceeding.
- Do not use dataset names, filenames, memorized examples, or external information.

Task-Specific Checks
Apply when answering questions that reference multiple images.

- When comparing images for thematic or semantic similarity, verify that at least two distinct, non-redundant visual anchors are consistent across candidate images.
- When evaluating whether a collection of images satisfies a condition, confirm the relevant quantifier: a universal claim fails if any required image contradicts it, whereas an existential claim requires only one valid instance.
- When identifying a paired comparison, verify that the images share a meaningful semantic or structural relationship rather than superficial visual resemblance.
- When no candidate image matches the reference in key structural or spatial anchors, do not force a comparison.
- When judging whether images share the same scene, building, or identity, prioritize stable structural anchors over decorative or transient elements.

Final verification:

- Confirm that the selected answer is uniquely supported by the visual evidence.
- Do not infer absence solely because an object or motion is not clearly visible.
- For directional or quantifier reasoning, verify that the correct reference frame and logical interpretation have been used.

Apply for multi-image spatial and relational reasoning questions.

- When evaluating directional terms such as "turned to the right" or "facing left," confirm the intended reference frame rather than relying on image tilt or an object's internal orientation.
- When a statement specifies a spatial relationship, interpret it literally unless the question explicitly introduces another reference frame.
- When a "cannot be determined" option is available and a precise answer requires information that is not visually recoverable, prefer the insufficient-information option rather than inferring unsupported details.

Final verification:

- Confirm that the selected answer is uniquely supported by the visual evidence.
- Avoid reasoning based on inferred absence when the relevant evidence is not visible.
- Verify the interpretation of directional and quantifier relations before producing the final answer.

Output
Follow the output format stated in the question exactly, including any answer tags it specifies. Do not shorten, skip, or reorder the reasoning the question asks for, and never state the final answer before the checks above are done.

Figure 16: Extracted Capability skill for multi-image reasoning.