SSL4RL: REVISITING SELF-SUPERVISED LEARNING AS INTRINSIC REWARD FOR VISUAL-LANGUAGE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) have shown remarkable abilities by integrating large language models with visual inputs. However, they often fail to utilize visual evidence adequately, either depending on linguistic priors in vision-centric tasks or resorting to textual shortcuts during reasoning. Although reinforcement learning (RL) can align models with desired behaviors, its application to VLMs has been hindered by the lack of scalable and reliable reward mechanisms. To overcome this challenge, we propose **SSL4RL**, a novel framework that leverages self-supervised learning (SSL) tasks as a source of verifiable rewards for RL-based fine-tuning. Our approach reformulates SSL objectives—such as predicting image rotation or reconstructing masked patches—into dense, automatic reward signals, eliminating the need for human preference data or unreliable AI evaluators. Experiments show that SSL4RL substantially improves performance on both vision-centric and vision-language reasoning benchmarks, with encouraging potentials on open-ended image-captioning tasks.. Through systematic ablations, we identify key factors—such as data volume, model scale, model choice, task difficulty, and semantic alignment with the target domain—that influence the effectiveness of SSL4RL tasks, offering new design principles for future work. We also demonstrate the framework's generality by applying it to graph learning, where it yields significant gains. SSL4RL establishes a versatile and effective paradigm for aligning multimodal models using verifiable, self-supervised objectives.

1 INTRODUCTION

Vision–Language Models (VLMs) have rapidly advanced multimodal understanding by leveraging the expert-level reasoning capabilities of Large Language Models (LLMs). This synergy has enabled broad applications, from visual question answering to interactive dialogue. Yet the reliance on linguistic priors introduces systematic weaknesses. For *vision-centric tasks*—where answers must be derived solely from image content, such as classification—VLMs often lag behind specialist vision models (Fu et al., 2025; Tong et al., 2024; Fu et al., 2024). Conversely, in *vision–language reasoning tasks*, VLMs tend to exploit textual knowledge rather than grounding their reasoning in visual evidence, a tendency amplified in long-form generation (Jian et al., 2025). These limitations underscore the need for training methods that reinforce visual grounding and robust reasoning simultaneously.

Reinforcement learning (RL) has emerged as the dominant paradigm for post-training large models, demonstrating that preference-based signals—collected from humans or distilled from AI feedback—can substantially improve helpfulness and alignment (Ouyang et al., 2022; Rafailov et al., 2023; Bai et al., 2022). More recently, *verifier-driven RL* has shown particular promise: models trained with automatically checkable rewards achieve striking gains in domains such as mathematics and programming (Le et al., 2022; Shao et al., 2024). However, these successes expose a fundamental bottleneck: outside domains with explicit programmatic verifiers, scalable and reliable rewards are scarce. In such cases, training pipelines often revert to *LLM-as-a-judge* heuristics, which are biased, noisy, and prone to adversarial manipulation (Raina et al., 2024; Chen et al., 2024b). This raises a key question: *how can we obtain abundant, verifiable reinforcement signals for VLMs in domains where external verifiers are absent?*

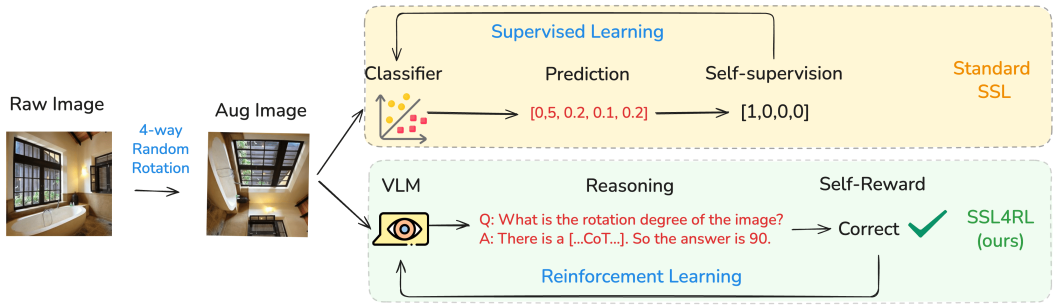


Figure 1: Overview of the SSL4RL framework. A corruption function transforms an input into a context–target pair. The model conditions on the context, generates predictions, and receives a verifiable reward by comparing against the target. The reward is then used to optimize the model via Reinforcement Learning (RL).

Self-supervised learning (SSL) offers a natural but underexplored answer. SSL has been central to representation learning across modalities: masked language modeling and next-token prediction for text (Devlin et al., 2019; Lewis et al., 2020; Raffel et al., 2020), contrastive learning and masked autoencoders for vision (Gidaris et al., 2018; Noroozi & Favaro, 2016; Doersch et al., 2015; Chen et al., 2020; He et al., 2020; Grill et al., 2020; Caron et al., 2021; He et al., 2022). The principle is simple yet powerful: perturb data, and require the model to reconstruct or discriminate. Crucially, SSL tasks define *intrinsically verifiable targets*. Given an image and its rotated variant, the ground-truth angle is unambiguous. Unlike preference-based signals, these targets are properties of the data itself, providing dense, reliable supervision at scale.

In this paper, we propose **SSL4RL**, a general framework that repurposes SSL tasks as verifiable reward functions for RL-based post-training. Instead of treating SSL solely as a pre-training tool (Tong et al., 2024), we reinterpret it through the lens of RL: corrupted inputs define trajectories, correctness defines rewards, and policy optimization drives updates. SSL4RL requires no human labels, external verifiers, or heuristic judges, yet produces dense and scalable reinforcement signals. Importantly, unlike conventional SSL, SSL4RL emphasizes generating natural language reasoning paths to solve vision tasks, thereby bridging perceptual learning and reasoning alignment.

We systematically evaluate SSL4RL on both vision-centric and multimodal reasoning benchmarks. On ImageNet-1K, SSL4RL significantly improves classification accuracy over the base model. For vision–language reasoning tasks, it delivers consistent gains, with average improvements of 9% on MMBench and 8% on SEED-Bench. On open-ended image-captioning platform CapArena (Cheng et al., 2025), SSL4RL consistently improves over the base model, with the largest performance gain of 8.14 points. A key finding is that the effectiveness of SSL tasks in SSL4RL differs from their role in traditional SSL: contrastive objectives, while dominant in pre-training, show limited benefit unless paired with stronger data augmentation, whereas position prediction—often deemed too trivial for SSL—emerges as highly effective in the SSL4RL setting. The ablation study reveals that the efficacy of an SSL4RL task is influenced by model capacity, its semantic alignment with downstream tasks, and the inherent difficulty of the task itself. These findings offer initial insights into what makes a good SSL4RL task.

We demonstrate that the SSL4RL paradigm generalizes beyond vision by extending it to the graph domain. Through three graph-related SSL tasks—attribute masking, neighbor prediction, and link prediction (Hou et al., 2022; Hu et al., 2020)—we achieve marked improvements on node classification and link prediction benchmarks. These findings highlight SSL4RL as a versatile recipe for extracting verifiable rewards with self-supervised tasks.

Contributions. Our work makes two key contributions: **(1)** we introduce SSL4RL, a unified framework bridging self-supervised learning and RL-based post-training through verifiable rewards; **(2)** we provide a comprehensive cross-domain study identifying which SSL tasks best transfer to reasoning and which do not. Together, these results challenge the assumption that all self-supervision is equally useful, and emphasize that the abundance of verifiable signals in SSL can be harnessed not only for representation learning, but also to drive alignment and reasoning in VLMs.

2 Related Work

RL training with external verifiers. Reinforcement learning has become a dominant paradigm for aligning large models. RLHF aligns LLMs with human intent through preference data (Ouyang

et al., 2022), while Direct Preference Optimization (DPO) reframes preference learning as a contrastive loss without explicit reward models (Rafailov et al., 2023). Constitutional AI replaces human raters with rule-based AI feedback (Bai et al., 2022). More recently, *verifier-driven RL* has achieved notable success in domains such as code and math, where correctness can be mechanically checked (Le et al., 2022; Shao et al., 2024). However, in domains lacking such verifiers, many systems fall back on *LLM-as-a-judge* signals. While convenient, these rewards are biased, noisy, and adversarially manipulable (Raina et al., 2024; Chen et al., 2024b), undermining their reliability. This limitation motivates the search for scalable alternatives where correctness is intrinsic to the data.

RL training with self-reward. Several methods attempt to reduce dependence on external labels or verifiers by enabling models to generate their own training signals. Self-Instruct (Wang et al., 2023b) bootstraps fine-tuning data through synthetic instructions, while STaR (Zelikman et al., 2022) and Reflexion (Shinn et al., 2023) improve reasoning via self-generated rationales and feedback. Self-Consistency (Wang et al., 2023a) aggregates multiple sampled rationales to increase reliability. Reinforced Pre-Training (RPT) scales this idea by using RL objectives such as next-token prediction at the pretraining stage (Liu et al., 2025a). While these methods reduce the need for human labels, they still optimize toward approximating *original* task accuracy, often requiring correctness evaluation or model-judging heuristics. In contrast, our approach seeks verifiable, abundant signals outside the original task by reinterpreting SSL tasks as reinforcement rewards.

Self-supervised learning across modalities. Self-supervision has been the foundation of representation learning across domains. In language, pretext tasks include masked language modeling (Devlin et al., 2019; Lewis et al., 2020; Raffel et al., 2020) and next-token prediction. Variants that mask reasoning steps show promise in enhancing mathematical reasoning (Chen et al., 2024a). In vision, early tasks include rotation (Gidaris et al., 2018), jigsaw (Noroozi & Favaro, 2016), and context prediction (Doersch et al., 2015), while modern SSL emphasizes contrastive learning (Chen et al., 2020; He et al., 2020; Grill et al., 2020; Caron et al., 2021) and generative masking (He et al., 2022). In multimodal learning, CLIP (Radford et al., 2021) and ALBEF (Li et al., 2021) demonstrate the power of contrastive and distillation objectives for aligning vision and language. Graph SSL extends these principles to structural data, with node/edge masking (Hu et al., 2020; Hou et al., 2022) and contrastive augmentations (You et al., 2020). A unifying property of all these objectives is their *intrinsically verifiable targets*, making them natural candidates for repurposing as RL rewards. The most relevant work is Jigsaw-R1 (Wang et al., 2025), which first establishes jigsaw puzzles as an effective pretext task for MLLM reinforcement learning. We generalize this concept into a unified SSL4RL framework beyond one specific task and can be applied to other domains like graphs. Through comprehensive studies, we systematically investigate the scaling of data volume, model size, task difficulty, and task combinations, offering broader insights into this learning paradigm.

Building on these threads, our work unifies SSL and RL post-training. By treating SSL objectives as reinforcement rewards, SSL4RL supplies dense, scalable, and verifiable signals without human annotations, model-judging heuristics, or handcrafted verifiers. This perspective highlights that the supervision already embedded in SSL tasks can be harnessed to drive reasoning improvements.

3 SSL4RL Framework

We formalize SSL4RL as a general recipe for converting self-supervised learning (SSL) objectives into reinforcement learning (RL) rewards for post-training large language models and related architectures. This section introduces the notation, shows how SSL tasks are reinterpreted under the RL formalism, and describes the optimization strategy. A high-level illustration of the framework is shown in Figure 1.

Problem Setup. Let π_θ denote a parametric model with parameters θ, defined over sequences of actions. In vision–language tasks, actions may include discrete classification labels (*e.g.*, rotations, patch indices) or text tokens. We assume access to a data distribution $\mathcal{D}$ of inputs $x \in \mathcal{X}$, which can be text and images. In the standard RL formalism, a trajectory τ is generated by rolling out π_θ in an environment $\mathcal{E}$, and receives a scalar reward $R(\tau)$. In SSL4RL, the "environment" is defined by a corruption function $c(x) = (\tilde{x}, y)$, which maps an input x into a corrupted context $\tilde{x}$ and a ground-truth target y. The policy π_θ conditions on $\tilde{x}$ to produce an output $\hat{y}$, and a reward $r(\hat{y}, y)$ is computed based on agreement with the ground truth. Thus, every SSL task induces an RL task:

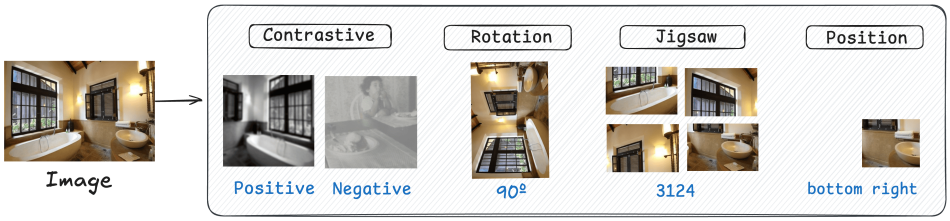


Figure 2: Four SSL4RL tasks considered in our study. *Rotation*: An image is rotated by a predefined angle, and the task is to predict this angle. *Jigsaw*: After dividing an image into a grid and permuting the patches, the goal is to predict the correct permutation index. *Contrastive*: Two augmented views are generated from an image, and the objective is to identify whether two given views originate from the same source image. *Position*: Given an image and a single patch cropped from it, the task is to predict the patch's original spatial position.

the corruption defines the environment, the target defines the verifiable ground truth, and the reward function provides supervision.

From SSL to RL Rewards. Our work revisits four representative SSL pretext tasks, depicted in Figure 2. Each SSL task is defined by a tuple (c, y, r), where c is a corruption function applied to the inputs x (specifically referring to images), y is the self-supervised target derived from the corruption, and r is the reward signal computed based on the model's prediction $\hat{y}$. Specifically,

- ***Rotation Prediction*** (Gidaris et al., 2018): $c(x, y)$ rotates x by an angle y. The reward is $r = \mathbb{1}[\hat{y} = y]$, where $\hat{y}$ is the predicted angle.
- ***Jigsaw Puzzles*** (Noroozi & Favaro, 2016): $c(x, y)$ partitions x into a grid and permutes patches by index y. The reward is $r = \mathbb{1}[\hat{y} = y]$, where $\hat{y}$ is the predicted permutation.
- ***Contrastive Learning*** (Chen et al., 2020; Radford et al., 2021): $c(x)$ generates augmented views. The reward r is a binary classification, imitating the InfoNCE similarity score that encourages high similarity for positive pairs and low similarity for negatives.
- ***Patch Position Prediction*** (Doersch et al., 2015): $c(x, y)$ extracts a patch from location y. The reward is $r = \mathbb{1}[\hat{y} = y]$, where $\hat{y}$ is the predicted location.

These rewards are *verifiable* as they are computed against unambiguous ground-truth targets y.

Optimization via GRPO. We adopt Grouped Reinforcement Policy Optimization (GRPO) (Shao et al., 2024), an efficient policy-gradient method designed for large-scale LLM training. Given a reference policy π_0 (e.g., the supervised fine-tuned model before RL), GRPO optimizes π_θ by maximizing the following regularized objective:

$$\mathcal{J}(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}[R(\tau) - \beta \, \mathrm{KL}(\pi_\theta(\cdot|\tau) \,\|\, \pi_0(\cdot|\tau))] \,, \tag{1}$$

where $R(\tau)$ is the SSL-derived reward, and β controls the strength of KL regularization. The KL penalty prevents divergence from the reference distribution and stabilizes training. GRPO performs updates by sampling rollouts, normalizing rewards across groups, and applying clipped policy-gradient updates similar to PPO (Schulman et al., 2017), but in a manner more compute-efficient for large batch LLM training. All experiments in this work apply the same GRPO configuration across tasks, ensuring that differences in outcomes are attributable to the choice of SSL reward.

4 What Makes a Good SSL4RL Task?

In this section, we examine the core principles for designing effective SSL4RL tasks through comprehensive experiments on vision-language reasoning (Section 4.1), open-ended image-captioning (Section 4.2), and vision-centric benchmarks (Section 4.3). Our results indicate that standard SSL strategies do not directly transfer to the SSL4RL setting. We provide a robustness analysis in Section 4.4. In Section 4.5, we present ablation studies on data volume, model size, base model choice, task difficulty scaling, and task combinations to elucidate these design principles further.

4.1 VISION-LANGUAGE REASONING TASKS

SSL Task Settings. For each SSL pretext task, we implement the following configurations. In Rotation, images are rotated counterclockwise by a randomly selected angle from 0°, 90°, 180°, 270°. In Jigsaw, each image is partitioned into a 2×2 grid, and the patches are randomly shuffled. In Contrastive, we apply the standard augmentation pipeline from Chen et al. (2020), including color jittering, grayscale conversion, Gaussian blur, horizontal flipping, and random resized cropping, each with an application probability of 0.2. In Position, the image is divided into four equal quadrants, and the target is to identify which quadrant (upper-left, upper-right, lower-left, lower-right) contains a specified patch.

Training and Evaluation Settings. Our experiments primarily adopt Qwen-2.5-VL-3B/7B-Instruct (Bai et al., 2025) for their moderate sizes and strong reasoning performance. For GRPO training, we set the group size to be 5, the KL loss coefficient to be 0.01, the entropy loss coefficient to be 0, and the context length to be 2048. We train the models on 8xA800 GPUs with a batch size of 512. For evaluation, we adopt the third-party evaluation tool VLMEvalKit (Duan et al., 2024), with the default sampling configured with a temperature of 0.01, top-p of 0.001, and top-k of 1. We visualize the rewards, entropy, and response length trajectories during RL training in Appendix K.

Benchmarks. We assess our approach on six prominent Vision Questioning Answer (VQA) benchmarks: MMBench (Liu et al., 2024), SEED-Bench (Li et al., 2023), V* (Wu & Xie, 2024), RealWorldQA (xAI, 2024), BLINK (Fu et al., 2024), and MME-RealWorld-Lite (Zhang et al., 2024), featuring challenging VLMs' ability on visual perception, spatial understanding, detail capturing, real-world applications, and so on. Detailed datasets descriptions are provided in Appendix I.

Results. Overall, the proposed SSL4RL paradigm leads to substantial performance gains on vision questioning answer tasks. Shown in Table 1 and Appendix Table 23, on average, SSL4RL models outperform the base model by 7.39% on MMBench and 8.94% on SEED-Bench. Notably, SSL4RL achieves a remarkable improvement of 39.00 percentage points (80.54% vs. 41.54%) on the Relation Reasoning task in MMBench. On the Visual Reasoning task of SEED-Bench, the improvement reaches up to 19.63 percentage points (73.41% vs. 53.78%). Results of other benchmarks are presented in Appendix Table 24, where the SSL4RL strategy brings significant improvements, especially on V* (+8.90%) and RealWorldQA (+9.55%). These consistent improvements validate the effectiveness of SSL4RL in enhancing vision-language reasoning. To contextualize the performance of our SSL4RL method, we include a strong baseline *VLM-R1* (Shen et al., 2025). We fine-tune the base model on 60% of the benchmark and evaluate on the held-out test set. The RL reward is computed directly based on the *golden* answer. We denote the tuned model as **Golden-3B** (see more details in Appendix G). As shown in Appendix Table 20 and 21, the performance gap between our SSL4RL variants and the Golden-3B oracle is relatively small (*e.g.*, 81.35% vs. 84.93% on MMBench, and 69.80% vs 73.21% on SEED-Bench), compared to our improvements over base models. This demonstrates that our self-supervised objectives, which require *no labeled downstream data*, can effectively close the gap to the performance driven by idealized, task-specific reward signals.

Table 1: Test performance (%) on MMBench downstream tasks. Logical: Logical Reasoning, Relation: Relation Reasoning, Attribute: Attribute Reasoning, Coarse: Coarse Perception, Cross Inst.: Cross-Instance Fine-grained Perception, Single-Inst.: Single-Instance Fine-grained Perception.

Category	**Model**	Logical	Relation	Attribute	Coarse	Cross-Inst.	Single-Inst.	***Average***
Base	Qwen2.5-VL-3B	61.77	41.54	76.62	73.55	64.32	82.06	72.99
SSL4RL	Rotation	65.84	**80.54**	**83.89**	80.21	**71.53**	84.76	**80.38**
	Jigsaw	62.86	74.51	80.35	77.92	67.82	84.31	77.82
	Contrastive	61.12	73.42	71.81	65.38	58.39	78.50	69.27
	Position	**67.65**	77.19	82.22	**82.15**	66.51	**85.39**	80.08
Maximal Improvement		**↑ 5.88**	**↑ 39.00**	**↑ 6.77**	**↑ 8.60**	**↑ 7.21**	**↑ 3.33**	**↑ 7.39**

Analysis. Our experimental analysis reveals that the Rotation and Position pretext tasks consistently yield the strongest performance gains. The success of the Position task is intuitive, as localizing a patch within the global image compels the model to integrate fine-grained local details with the overall scene layout, fostering integrated spatial understanding. In contrast, the remarkable effec-

tiveness of the Rotation task presents a more nuanced insight. Despite its perceptual simplicity for humans, we find the task poses a considerable challenge to VLMs, as evidenced by the base model's near-chance accuracy. We interpret this as evidence that rotation prediction facilitates "anti-commonsense" learning. The model's pre-training instills a strong prior for canonical orientations, *e.g.*, people are typically depicted upright. Rotated images violate this prior, forcing the model to reconcile anomalous inputs with its existing knowledge, thereby sharpening its relational reasoning and visual comprehension. Conversely, the standard Contrastive task leads to performance degradation on some downstream tasks. We hypothesize that its default augmentations are insufficiently challenging, causing the model to overfit to superficial invariances without learning semantically meaningful structures. Subsequent ablation studies confirm this: employing a more aggressive augmentation strategy recovers significant performance. This finding highlights the critical need to align SSL task difficulty with model capacity to elicit generalizable feature learning, as explored in Section 4.5.2. Through a sub-task specific analysis, we find that Rotation exhibits superior performance in Relation Reasoning and Cross-Instance Perception, which can be attributed to its inherent requirement for a structural comprehension of object orientation and spatial relationships. Position demonstrates leading results in Logical Reasoning and holistic Scene Understanding, as its objective of reconstructing a coherent scene from disparate patches fosters robust integrative scene modeling. Jigsaw shows consistent utility in domains reliant on contextual reasoning. Limited by the insufficient augmentations, Contrastive's optimization for global image similarity appears to inadequately cultivate the fine-grained and relational reasoning capabilities.

Observations. Through a qualitative analysis of model responses, we observe two key improvements attributable to SSL4RL training. (1) **Sharper Attention**: The trained models exhibit more precise attention alignment with text queries. For instance, when queried about "hair" (Figure 3), the base model's attention is diffuse, whereas our model accurately focuses on the relevant region. This is not merely an effect of sparsity, as evidenced by the "sky" query (Appendix Figure 13): our model attends to the entire sky, while the base model activates only scattered pixels. (2) **Reduced Language Bias**: SSL4RL mitigates over-reliance on linguistic priors, fostering greater dependence on visual evidence. For example, when asked about a chandelier's color (Appendix Figure 12), the base model defaults to a common-sense response (*e.g.*, a typical decorative color), while our model localizes the object and answers based on the actual appearance.

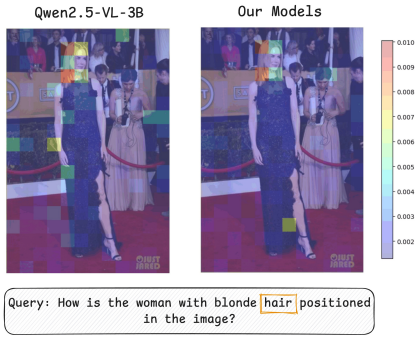


Figure 3: Cross-Attention Heatmap Comparison. More instances are shown in Appendix L.

Takeaway

SSL as Intrinsic Reward Sharpens VLM Reasoning. The SSL4RL paradigm demonstrably enhances vision-language reasoning by repurposing SSL tasks as intrinsic rewards. It deepens the perception and understanding of the image itself, leading towards more precise visual attention and less language bias.

Task Choice is Critical. SSL tasks show effectiveness when their inherent semantic aligns with core reasoning skills (*e.g*, Position and Rotation), while an inappropriate task may induce negative transfer and hinder downstream performance.

4.2 Open-ended Vision-Language Task: Image Captioning

SSL4RL requires no human labels, external verifiers, or heuristic judges, yet produces dense and scalable reinforcement signals. Intuitively, it has potential for open-ended tasks where ground truth is ill-defined and human annotations are expensive. To evaluate SSL4RL's potential on open-ended tasks, we leverage a recent image captioning platform, CapArena (Cheng et al., 2025), where captions from a test model are compared against those from strong baseline models using GPT-4o as a judge, with human references provided for context. See Appendix I.2 for more details. As shown in Table 2. **SSL4RL consistently improves over the base model**, with the largest performance gain of **8.14 points** (56.45 vs. 48.31). This clearly shows that the self-supervised rewards provide a meaningful learning signal even in the absence of a verifiable ground truth.

Qualitative Insights. We conduct a qualitative comparison with examples in Appendix Table 26, shed light on *how* SSL4RL enhances open-ended generation: (1) **Enhanced Detail Capture.** SSL4RL models correctly capture more details in the image. For instance, our SSL4RL model correctly identifies a "scoreboard" displaying "1 0 0" and a "disabled persons' sign," which the base model either misinterprets or omits. (2) **Improved Spatial Reasoning.** SSL4RL models tend to use more precise spatial descriptors like "*behind the fence*," "*to the left of*," and "*the center of the court*", aligning well with the spatial reasoning ability required by SSL pretext tasks. In summary, by providing a dense, automated learning signal derived from the data's intrinsic structure, our method successfully improves model performance on the complex, open-ended task of image captioning.

Table 2: Scores on CapArena Platform. GPT-Score: the score against GPT-4o. Cog-Score: the score against CogVLM-19B. CPM-Score: the score against MiniCPM-8B.

Category	Model	GPT-Score	Cog-Score	CPM-Score	*Average*
Base	Qwen2.5-VL-3B	0.00	6.48	92.96	48.31
SSL4RL-3B	Rotation	19.15	49.30	96.48	55.28
	Jigsaw	8.87	54.93	98.59	55.64
	Contrastive	4.96	57.95	100.00	56.45
	Position	0.00	59.15	92.25	51.04

4.3 Vision-Centric Tasks: ImageNet Classification

Benchmarks. We further evaluate the SSL4RL paradigm on vision-centric tasks using the ImageNet-1K dataset (Deng et al., 2009), which comprises approximately 1.3 million images across 1,000 categories. From this dataset, we construct a balanced subset of 100,000 training and 10,000 test images. To probe reasoning capabilities at varying difficulty levels, we design three question types: (1) ***Completion***: directly answer with the species name; (2) ***Choice-20***: select the correct species from 20 candidates; (3) ***Choice-200***: select the correct species from 200 candidates.

Results. As shown in Table 3, models fine-tuned with SSL4RL consistently outperform the base model across all question types on the ImageNet-1K classification task. Consistent with findings on reasoning benchmarks, the Position task leads to the largest performance gains, *e.g.*, 67.14% vs. 57.20% on Choice-200. However, a key divergence emerges with the Contrastive task. While it underperformed on vision-language reasoning, it shows competitive results on ImageNet classification. We attribute this result to the nature of the downstream benchmark. As an instance discrimination task, ImageNet benefits from learning strong semantic representations through invariance to augmentations—precisely the strength of contrastive learning. This result also verifies that task selection for SSL4RL must consider the specific capabilities required by the target application.

Table 3: Test performance (%) on ImageNet downstream tasks.

Category	Model	Completion	Choice-20	Choice-200
Base Model	Qwen2.5-VL-3B	24.93	85.22	57.10
SSL4RL	Rotation	**29.19**	87.26	58.48
	Jigsaw	28.75	87.55	60.80
	Contrastive	26.84	<u>89.51</u>	<u>61.78</u>
	Position	<u>28.76</u>	**92.35**	**67.14**

4.4 Robustness Analysis

Settings. To evaluate our model's robustness, we conducte a comprehensive evaluation under various image perturbations. We designed two levels of perturbation—weak and strong—to assess the models' resilience to visual corruptions. The weak perturbation applies a series of common image transformations, each with a probability of 0.5, including color jittering, random conversion to grayscale, Gaussian blur, and horizontal flipping. The strong perturbation builds upon the weak one by incorporating an additional multi-crop strategy. This strategy generates two 224×224 pixel crops per image via randomly resized cropping (scale range [0.08, 1.0]), presenting a more significant challenge to the model's perception. Examples of these perturbed images are provided in Figure 6.

Results. As shown in Table 18 and Table 19, our method consistently demonstrates superior robustness across both perturbation levels. For instance, under weak perturbation, our SSL4RL-Rotation model achieves an average score of 72.11%, outperforming the base model at 66.80%. This performance gap is maintained under strong perturbation, where our SSL4RL-Position model scores 64.37% compared to the base model's 59.62%. These results confirm that our SSL4RL strategy effectively enhances the model's robustness to photometric variations.

4.5 Ablation Study

4.5.1 Data Volume Scaling

A key question for data-driven methods is the ability to leverage increasing amounts of data. To characterize the relationship between performance and data scaling for SSL4RL, we progressively expand our training dataset, obtaining three regimes of increasing data volumes: (1) **Base Set**: ∼4,000 samples from MMBench. (2) **Extended Set**: ∼18,000 samples from a mixture of MMBench and SEED-Bench. (3) **Full Set**: ∼118,000 samples from MMBench, SEED-Bench, and ImageNet.

We select the representative Position for evaluation. The results, detailed in Table 4, demonstrate a clear positive scaling relationship across almost all subtasks. When scaling from Base Set to Extended Set, the average performance on MMBench improved from 80.08% to 81.38%, a gain of +1.30%. A further expansion to the Full Set yielded an additional +1.04% improvement, reaching 82.42%. This consistent, monotonic improvement suggests that the effectiveness of SSL4RL continues to benefit from larger, more diverse datasets. We provide more discussions in Appendix E.

Table 4: Performance (%) on MMBench with increasing training data volumes.

Model	**Training Volumes**	Logical	Relation	Attribute	Coarse	Cross-Inst.	Single-Inst.	***Average***
Qwen2.5-VL-3B	–	61.77	41.54	76.62	73.55	64.32	82.06	72.99
SSL4RL-Position	4,000	67.65	77.19	82.22	82.15	66.51	85.39	80.08
	18,000	68.71	79.26	**85.03**	81.65	70.16	86.47	81.38
	118,000	**68.77**	**79.29**	84.26	**83.66**	**72.87**	**88.31**	**82.42**

4.5.2 Model Size Scaling

To investigate the scalability of the SSL4RL paradigm, we apply it to the larger Qwen2.5-VL-7B-Instruct model. The results (Table 5 and Appendix Table 9) confirm that SSL4RL still yields improvements over the base model, notably enhancing Logical Reasoning on MMBench by 3.70% and Text Understanding on SEED-Bench by 3.57%. However, the gains are less pronounced than those observed with the 3B model. Even after increasing the task difficulty to a 5×5 grid for Jigsaw and Position tasks, no significant improvement was observed (see Appendix C).

We attribute these diminishing returns to a fundamental ceiling effect imposed by the predefined SSL tasks. The *absolute* difficulty of the four SSL tasks is fixed, presenting a suitable challenge for the 3B model but potentially failing to engage the full capabilities of a 7B model. The larger model's enhanced abilities could render the tasks trivial, thereby weakening their effectiveness as a learning signal. This underscores a key insight: the efficacy of an SSL task is contingent on its ability to present a non-trivial challenge commensurate with the model's capacity. Consequently, a primary challenge for future work is the design of adaptive or inherently more complex SSL objectives that can continue to provide a learning signal for large-scale models. In Appendix J, we provide a preliminary exploration of harder SSL4RL tasks on 7B models.

Table 5: Test performance (%) of 7B models on MMBench downstream tasks.

Category	**Model**	Logical	Relation	Attribute	Coarse	Cross-Inst.	Single-Inst.	***Average***
Base	Qwen2.5-VL-7B	76.49	84.68	85.69	84.66	84.49	89.15	86.37
SSL4RL-7B	Rotation	78.70	**86.16**	87.13	**85.91**	**88.47**	88.92	87.50
	Jigsaw	**80.19**	84.64	**88.02**	85.70	84.53	**90.93**	**87.73**
	Contrastive	77.73	83.19	84.58	85.89	85.29	87.74	86.25
	Position	79.06	82.13	85.29	85.02	85.42	88.95	86.25
Maximal Improvement		**↑ 3.70**	**↑ 1.48**	**↑ 2.33**	**↑ 1.25**	**↑ 3.98**	**↑ 0.98**	**↑ 1.36**

4.5.3 BASE MODEL CHOICE

To evaluate whether the benefits of SSL4RL extend to architectures beyond the Qwen2.5-VL series, we select Gemma3-4B (Team et al., 2025), a recent and powerful model from Google with a distinct architecture. To ensure a fair comparison under computational constraints, we adapted our training setup by halving the batch size to 256 while meticulously preserving all other hyperparameters and training procedures. Results presented in Appendix Table 22 provide strong evidence for the general applicability of our method. On the Gemma3 base model, our SSL4RL models yield a consistent and notable average performance improvement of 2.88% on MMBench. **The gain is particularly pronounced in Cross-instance Perception (5.76%), Logical Reasoning (4.35%), and Attribute Reasoning (4.13%).** Moreover, the relative efficacy of the individual SSL4RL tasks is remarkably consistent with our prior findings. As with Qwen2.5-VL, tasks like Rotation and Position confer the most significant benefits, while the simpler Contrastive task shows more modest gains. The consistent results validate that SSL4RL is a general-purpose principle for VLMs, not an artifact of a particular model family.

4.5.4 TASK DIFFICULTY SCALING

To investigate the role of task difficulty, we design more challenging self-supervised learning (SSL) tasks by intensifying their corruption strategies. Specifically, for the Position and Jigsaw tasks, we increase the crop granularity from 2 to 3, resulting in a finer 3×3 grid of patches. For the Contrastive task, we enhance the augmentation strength by raising the application probability from 0.2 to 0.8 and reducing the maximum crop scale from 1.0 to 0.3, thereby generating positive samples that differ more substantially from the anchor image. For the Rotation task, we refine the rotation angle interval from 90° to 45°, increasing the complexity of the angle prediction.

The impact of task difficulty varies considerably across different SSL tasks, as evidenced by the results in Figure 4 (full results in Appendix C). The performances for the Contrastive task shows a marked improvement upon increasing its difficulty, elevated from 69.27% to 77.89% on MMBench and from 61.90% to 65.00% on SEED-Bench. Conversely, the benefits are marginal for the Position task and even detrimental for Rotation and Jigsaw. A plausible explanation is that the Contrastive task's inherent simplicity as a binary discrimination problem means that raising the difficulty compels the model to learn more robust and informative features. On the other hand, exacerbating the difficulty of already challenging tasks like Jigsaw might induce over-specialization to the SSL objective, resulting in a negative transfer where the learned representations are less transferable or even counterproductive for downstream reasoning.

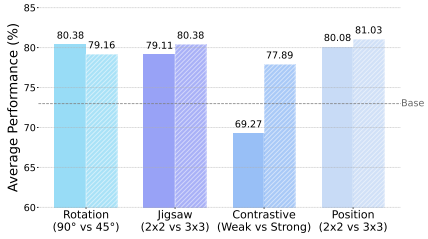


Figure 4: Test accuracy (%) on MMBench, varying the SSL task difficulty.

4.5.5 TASK COMBINATION

The preceding sections primarily investigate the effect of individual SSL rewards. A natural subsequent question is whether combining them during training can yield better performance compared to any single reward. To explore this, we train the Qwen2.5-VL-3B-Instruct model using a combination of all four SSL rewards. As illustrated in Figure 5, the combined approach, somewhat counterintuitively, does not yield significant improvements over the best single-reward setups. We hypothesize that this lack of additive improvement stems from several potential factors. First, different SSL tasks may encourage the model to learn complementary yet potentially conflicting feature representations. Simultaneously optimizing for multiple, distinct objectives could create a difficult optimization landscape where the model struggles to find a unified representation that satisfies all constraints at once, leading to interference rather than synergy. Second, the individual reward signals might vary in scale and dynamics, making it non-trivial to balance their contributions effectively without

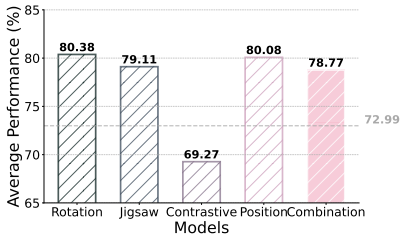


Figure 5: Test accuracy of SSL tasks.

careful reward shaping or weighting. A simple averaging of rewards might drown out the most useful learning signals. This finding suggests that a naive combination of SSL rewards is insufficient for achieving cumulative gains. It points to the need for more sophisticated integration strategies, such as dynamic reward weighting, curriculum learning that schedules different tasks, or even a meta-learner that selects the most beneficial task at different training stages.

Takeaway

Goldilocks Principle of Task Difficulty. The effectiveness of an SSL task is contingent on its difficulty being appropriately matched to the model's capacity. Insufficient challenge provides a weak learning signal, while excessive difficulty leads to negative transfer.

Diminishing Returns with Model Size. The performance gains from the four SSL tasks diminish as model size increases (from 3B to 7B), suggesting designing SSL tasks with inherently higher complexity for large-scale models.

Non-additivity of Rewards. A naive combination of multiple SSL rewards does not yield cumulative improvements, indicating potential optimization conflicts and underscoring the need for sophisticated integration strategies rather than simple averaging.

5 Extension to Other Domains: An Example on Graph

Having established SSL4RL for vision-language reasoning, we now explore its broader applicability. The paradigm's core principle—generating rewards from data transformations—naturally extends to any domain with rich structural information. Beyond images, graph-structured data presents a compelling candidate, given its explicit relational semantics that are amenable to various pretext tasks. In this section, we empirically validate this potential by adapting SSL4RL to the graph domain. We provide details of tasks, reward definitions, and benchmarks in Appendix D. The results in Table 6 and Appendix Table 16 demonstrate the successful application of SSL4RL to graph-structured data. The 3B model shows substantial improvements, with gains up to 13.79% on average, while the 7B model exhibits diminishing returns, mirroring our observations in the visual domain and reinforcing the "difficulty-capacity matching" principle. Furthermore, the relative effectiveness of the self-supervised tasks is contingent upon the nature of the downstream objective. Tasks emphasizing structural reasoning (Link and Neighbor Prediction) yield better performance on relation-centric tasks such as link prediction. Conversely, tasks focused on feature reconstruction (Attribute Mask) demonstrate a comparative advantage on node classification benchmarks. These findings not only validate the generalizability of the SSL4RL framework beyond the visual modality but also highlight the critical importance of aligning the pretext task's inductive bias with the target application.

Table 6: Test performance (%) of 3B models on downstream graph tasks.

Category	**Model**	Cora	PubMed	WikiCS	Products	fb15k237	wn18rr	***Average***
Base model	Qwen2.5-VL-3B	21.80	64.26	30.50	7.93	26.30	29.80	30.09
SSL4RL-3B	Attribute	**55.80**	73.27	**57.62**	8.03	32.50	36.10	**43.88**
	Neighbor	39.00	**74.37**	50.84	**12.65**	36.10	36.60	41.59
	Link	31.30	71.97	55.93	4.91	**46.50**	**41.10**	41.95
Maximal Improvement		**↑ 34.00**	**↑ 10.11**	**↑ 27.12**	**↑ 4.42**	**↑ 20.20**	**↑ 11.30**	**↑ 13.79**

6 Conclusions

We have introduced SSL4RL, a framework that repurposes self-supervised tasks as verifiable reinforcement learning rewards for post-training vision–language models. Our study shows that SSL4RL not only improves performance on vision-centric benchmarks such as ImageNet-1K, but also enhances multimodal reasoning, achieving substantial gains on MMBench and SEED-Bench. These findings suggest a broader principle: verifiable and scalable supervision signals are already embedded in self-supervision, and with proper task selection they can drive alignment and reasoning in VLMs without reliance on external verifiers, judges, or costly human labels. Looking forward, SSL4RL opens a path toward safer and more capable multimodal foundation models by unifying the strengths of self-supervision and reinforcement learning.

Ethics Statement

We are not aware of any specific ethical concerns related to this work. All experiments are conducted on publicly available or synthetic datasets, without the use of sensitive or proprietary information.

Reproducibility Statement

We provide complete details of our methods, hyperparameters, datasets, and evaluation metrics in both the main paper and the appendix. To further support transparency and reproducibility, we will release our code upon acceptance.

References

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 arXiv:2502.13923*, 2025.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jason Kernion, Jackson Jones, Andy Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv:2212.08073*, 2022. URL https://arxiv.org/abs/2212.08073.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. URL https://arxiv.org/abs/2104.14294.

Changyu Chen, Xiting Wang, Ting-En Lin, Ang Lv, Yuchuan Wu, Xin Gao, Ji-Rong Wen, Rui Yan, and Yongbin Li. Masked thought: Simply masking partial reasoning steps can improve mathematical reasoning learning of language models. In *ACL*, 2024a. URL https://arxiv.org/abs/2403.02178.

Guanting Chen et al. Humans or llms as the judge? a study on judgement bias. In *EMNLP*, 2024b. URL https://aclanthology.org/2024.emnlp-main.474.pdf.

Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Mixed autoencoder for self-supervised visual representation learning. In *CVPR*, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. URL https://arxiv.org/abs/2002.05709.

Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, 132(1):208–223, 2024c.

Kanzhi Cheng, Wenpo Song, Jiaxin Fan, Zheng Ma, Qiushi Sun, Fangzhi Xu, Chenyang Yan, Nuo Chen, Jianbing Zhang, and Jiajun Chen. Caparena: Benchmarking and analyzing detailed image captioning in the llm era. In *ACL Findings*, 2025.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang,

Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. URL https://arxiv.org/abs/1810.04805.

Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015. URL https://arxiv.org/abs/1505.05192.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *ACM MM*, 2024.

Jiarui Feng, Hao Liu, Lecheng Kong, Mingfang Zhu, Yixin Chen, and Muhan Zhang. Taglas: An atlas of text-attributed graph datasets in the era of large graph and language models. *arXiv preprint arXiv:2406.14683*, 2024.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: Vlms overlook their visual representations. *arXiv preprint arXiv:2506.08008*, 2025.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *ECCV*, 2024.

Songwei Ge, Shlok Mishra, Chun-Liang Li, Haohan Wang, and David Jacobs. Robust contrastive learning using negative samples with diminished semantics. In *NeurIPS*, 2021.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018. URL https://arxiv.org/abs/1803.07728.

Jean-Bastien Grill et al. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020. URL https://arxiv.org/abs/2006.07733.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. URL https://arxiv.org/abs/1911.05722.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. URL https://arxiv.org/abs/2111.06377.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. Graphmae: Self-supervised masked graph autoencoders. In *KDD*, 2022. URL https://arxiv.org/abs/2205.10803.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2019.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *ICLR*, 2020. URL https://arxiv.org/abs/1905.12265.

Pu Jian, Junhong Wu, Wei Sun, Chen Wang, Shuo Ren, and Jiajun Zhang. Look again, think slowly: Enhancing visual reflection in vision-language models. In *EMNLP*, 2025.

Wei Jin, Tyler Derr, Haochen Liu, Yiqi Wang, Suhang Wang, Zitao Liu, and Jiliang Tang. Self-supervised learning on graphs: Deep insights and new direction. *arXiv preprint arXiv:2006.10141*, 2020.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In *NeurIPS*, 2020.

Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

Hung Le, Yue Wang, Akhilesh D. Gotmare, Silvio Savarese, and Steven C.H. Hoi. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. In *NeurIPS*, 2022. URL https://arxiv.org/abs/2207.01780.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, 2020. URL https://arxiv.org/abs/1910.13461.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh D. Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. URL https://arxiv.org/abs/2107.07651.

Xiaodong Liu et al. Rpt: Reinforced pre-training of large language models. *arXiv:2505.07185*, 2025a. URL https://arxiv.org/abs/2505.07185.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, 2024.

Yuhong Liu, Beichen Zhang, Yuhang Zang, Yuhang Cao, Long Xing, Xiaoyi Dong, Haodong Duan, Dahua Lin, and Jiaqi Wang. Spatial-ssrl: Enhancing spatial understanding via self-supervised reinforcement learning. *arXiv preprint arXiv:2510.27606*, 2025b.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016. URL https://arxiv.org/abs/1603.09246.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022. URL https://arxiv.org/abs/2203.02155.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. URL https://arxiv.org/abs/2103.00020.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023. URL https://arxiv.org/abs/2305.18290.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020. URL https://arxiv.org/abs/1910.10683.

Vasu Raina et al. Is llm-as-a-judge robust? investigating universal adversarial attacks on zero-shot llm assessment. In *EMNLP*, 2024. URL https://aclanthology.org/2024.emnlp-main.427.pdf.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhoujun Shao et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models via grpo. *arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *NeurIPS*, 2023. URL https://arxiv.org/abs/2303.11366.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat

Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*, 2024.

Xuezhi Wang et al. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023a. URL https://arxiv.org/abs/2203.11171.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *ACL*, 2023b. URL https://arxiv.org/abs/2212.10560.

Zifu Wang, Junyi Zhu, Bo Tang, Zhiyu Li, Feiyu Xiong, Jiaqian Yu, and Matthew B. Blaschko. Jigsaw-r1: A study of rule-based visual reinforcement learning with jigsaw puzzles, 2025.

Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *CVPR*, 2024.

xAI. Grok-1.5 vision preview. https://x.ai/news/grok-1.5v, April 2024. URL https://x.ai/news/grok-1.5v. Accessed: [Today's Date].

Hong Xuan, Abby Stylianou, Xiaotong Liu, and Robert Pless. Hard negative examples are hard, but useful. In *ECCV*, 2020.

Yonglong You, Tianlong Chen, Zhangyang Sui, and Yang Wang. Graph contrastive learning with augmentations. In *NeurIPS*, 2020. URL https://proceedings.neurips.cc/paper/2020/file/3fe230348e9a12c13120749e3f9fa4cd-Paper.pdf.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *NeurIPS*, 2022. URL https://arxiv.org/abs/2203.14465.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024.

THE USE OF LARGE LANGUAGE MODELS (LLMS)

In this work, LLMs are primarily employed for polishing the language of the manuscript to ensure grammatical correctness and coherence. Importantly, all conceptual development, theoretical analysis, experimental design, and result interpretation are conducted independently by the authors. The use of LLMs is strictly limited to auxiliary tasks, ensuring that the scientific contributions of this paper remain entirely unaffected by such tools.

A DETAILED RESULTS OF MMBENCH

In this section, we present the detailed leaf task results of MMBench in Tables 7 and 8.

Table 7: Test performance (%) of 3B models on MMBench downstream tasks. IR: Identity Reasoning, PPR: Physical Property Reasoning, FR: Function Reasoning, OL: Object Localization, SIU: Structuralized Imagetext Understanding, AtR: Attribute Recognition, FP: Future Prediction, SR: Spatial Relationship, IS: Image Scene, IQ: Image Quality, Ace: Action Recognition, AC: Attribute Comparison, IT: Image Topic, NR: Nature Relation, PR: Physical Relation, SR: Social Relation, CR: Celebrity Recognition, IS: Image Style, OCR: OCR, IE: Image Emotion.

Category	Model	IR	PPR	FR	OL	SIU	AtR	FP	SR	IS	IQ	Ace	AC	IT	NR	PR	SR	CR	IS	OCR	IE	*Average*
Base	Qwen2.5-VL-3B	86.36	60.27	83.22	63.17	71.99	81.44	51.54	45.20	91.15	46.67	84.65	63.12	82.86	64.25	47.87	38.37	89.39	82.55	94.23	64.50	72.99
SSL4RL-3B	Rotation	**97.16**	**66.67**	**87.83**	63.49	75.53	85.23	56.15	**56.50**	95.82	55.33	89.30	**68.79**	87.14	**70.39**	**64.89**	88.37	**95.45**	87.26	**94.87**	75.50	**80.38**
	Jigsaw	91.48	64.38	85.20	61.59	73.40	89.02	52.31	53.11	93.86	56.00	86.51	63.83	**87.86**	64.25	59.57	81.98	94.95	85.38	91.67	66.50	77.82
	Contrastive	90.34	52.05	73.03	60.32	69.15	78.03	53.08	32.77	88.45	42.00	81.40	60.99	82.14	36.87	57.45	81.40	90.40	61.79	85.26	52.50	69.27
	Position	95.45	64.38	86.84	66.03	**78.37**	86.36	**56.92**	42.37	**96.31**	**58.00**	89.77	67.38	86.43	68.16	60.64	85.47	94.95	**91.51**	94.23	**78.50**	80.08
	Combination	95.45	61.64	82.24	**68.25**	77.66	**89.77**	54.62	45.76	94.10	55.33	**90.23**	63.83	**87.86**	61.45	54.26	**90.12**	94.19	89.15	**94.87**	67.50	78.77

Table 8: Test performance (%) of 7B models on MMBench downstream tasks. IR: Identity Reasoning, PPR: Physical Property Reasoning, FR: Function Reasoning, OL: Object Localization, SIU: Structuralized Imagetext Understanding, AtR: Attribute Recognition, FP: Future Prediction, SR: Spatial Relationship, IS: Image Scene, IQ: Image Quality, Ace: Action Recognition, AC: Attribute Comparison, IT: Image Topic, NR: Nature Relation, PR: Physical Relation, SR: Social Relation, CR: Celebrity Recognition, IS: Image Style, OCR: OCR, IE: Image Emotion.

Category	Model	IR	PPR	FR	OL	SIU	AtR	FP	SR	IS	IQ	Ace	AC	IT	NR	PR	SR	CR	IS	OCR	IE	*Average*
Base	Qwen2.5-VL-7B	**98.30**	66.67	92.11	74.60	82.98	88.64	70.00	76.27	97.05	58.00	92.09	85.11	91.43	88.83	69.15	92.44	97.22	94.81	96.15	82.00	86.37
SSL4RL-7B	Rotation	97.73	71.23	**92.43**	71.11	89.72	90.15	67.69	**78.53**	**97.79**	61.33	92.56	**94.33**	90.71	85.47	**71.28**	**93.60**	96.97	**96.23**	**97.44**	83.50	87.50
	Jigsaw	97.73	**74.89**	91.45	**75.24**	86.52	93.56	**73.85**	74.01	97.54	60.67	91.63	87.94	90.00	**89.39**	70.21	91.86	**97.47**	94.81	**97.44**	**85.50**	**87.73**
	Contrastive	**98.30**	63.01	**92.43**	74.29	85.46	85.61	70.00	77.97	97.30	**63.33**	**93.49**	84.40	88.57	85.47	67.02	91.28	**97.47**	95.75	93.59	84.50	86.25
	Position	**98.30**	63.01	**92.43**	74.29	85.46	85.61	70.00	77.97	97.30	**63.33**	**93.49**	84.40	88.57	85.47	67.02	91.28	**97.47**	95.75	93.59	84.50	86.25
	Combination	97.73	73.97	92.11	71.43	**90.43**	**95.08**	70.77	74.58	**97.79**	61.33	**93.49**	81.56	**92.14**	**89.39**	68.09	93.02	**97.47**	93.87	96.79	84.50	85.78

B RESULTS OF SSL4RL 7B-MODEL ON SEED-BENCH

In Table 9, we present the SEED-Bench results for the SSL4RL 7B-models.

Table 9: Test performance (%) of 7B models on SEED-Bench downstream tasks. TU: Text Understanding, VR: Visual Reasoning, SU: Scene Understanding, IId: Instance Identity, IIn: Instance Interaction, IA: Instance Attributes, IL: Instance Location, SR: Spatial Relation, IC: Instances Counting.

Category	**Model**	TU	VR	SU	IId	IIn	IA	IL	SR	IC	***Average***
Base	Qwen2.5-VL-7B	72.62	77.95	77.99	77.44	75.26	76.19	71.98	62.56	69.55	74.70
SSL4RL-7B	Rotation	**76.19**	78.25	78.59	77.94	73.20	76.90	**73.11**	64.23	69.76	75.33
	Jigsaw	70.24	79.15	78.44	77.66	**76.29**	76.58	73.01	63.77	69.27	75.05
	Contrastive	71.43	78.85	78.28	**78.32**	**76.29**	76.47	72.70	**64.69**	**70.33**	75.27
	Position	70.24	**80.66**	**78.82**	77.44	71.13	**77.97**	72.19	64.38	69.39	**75.56**
Maximal Improvement		**↑ 3.57**	**↑ 2.71**	**↑ 0.83**	**↑ 0.88**	**↑ 1.03**	**↑ 1.78**	**↑ 1.13**	**↑ 2.13**	**↑ 0.78**	**↑ 0.86**

C Detailed Results of the Ablation Study about Difficulty

From Table 10 to Table 13, we provide detailed experimental results for the ablation study on task difficulty of MMBench, SEED-Bench, and ImageNet1k.

Table 10: Test performance (%) of 3B models trained with different task difficulties on MM-Bench downstream tasks. IR: Identity Reasoning, PPR: Physical Property Reasoning, FR: Function Reasoning, OL: Object Localization, SIU: Structuralized Imagetext Understanding, AtR: Attribute Recognition, FP: Future Prediction, SR: Spatial Relationship, IS: Image Scene, IQ: Image Quality, Ace: Action Recognition, AC: Attribute Comparison, IT: Image Topic, NR: Nature Relation, PR: Physical Relation, SR: Social Relation, CR: Celebrity Recognition, IS: Image Style, OCR: OCR, IE: Image Emotion.

Model	Difficulty	IR	PPR	FR	OL	SIU	AtR	FP	SR	IS	IQ	Ace	AC	IT	NR	PR	SR	CR	IS	OCR	IE	*Average*
Qwen2.5-VL-3B	–	86.36	60.27	83.22	63.17	71.99	81.44	51.54	45.20	91.15	46.67	84.65	63.12	82.86	64.25	47.87	38.37	89.39	82.55	94.23	64.50	72.99
Rotation	90-degree	**97.16**	**66.67**	**87.83**	63.49	75.53	85.23	56.15	**56.50**	95.82	55.33	89.30	68.79	87.14	**70.39**	**64.89**	88.37	95.45	87.26	**94.87**	75.50	80.38
	45-degree	93.75	64.84	**87.83**	**66.67**	74.11	**90.15**	58.46	46.33	95.33	53.33	88.84	65.25	87.14	65.92	54.26	**90.12**	93.94	86.79	92.31	70.50	79.16
Jigsaw	2x2	91.48	64.38	85.20	61.59	73.40	89.02	52.31	53.11	93.86	56.00	86.51	63.83	87.86	64.25	59.57	81.98	94.95	85.38	91.67	66.50	77.82
	3x3	93.75	60.27	82.57	60.95	65.60	84.47	49.23	51.41	91.89	56.67	83.26	65.25	82.86	66.48	50.00	78.49	85.35	85.85	88.46	72.00	75.12
Contrastive	Weak	90.34	52.05	73.03	60.32	69.15	78.03	53.08	32.77	88.45	42.00	81.40	60.99	82.14	36.87	57.45	81.40	90.40	61.79	85.26	52.50	69.27
	Strong	94.89	64.84	82.24	63.49	77.30	86.36	51.54	44.07	95.82	47.33	89.30	70.21	85.71	59.78	52.13	86.63	94.44	87.74	92.31	70.50	77.89
Position	2x2	95.45	64.38	86.84	66.03	**78.37**	86.36	56.92	42.37	**96.31**	58.00	**89.77**	67.38	86.43	68.16	60.64	85.47	94.95	91.51	94.23	**78.50**	80.08
	3x3	**97.16**	63.47	87.50	66.35	75.53	88.64	**60.00**	52.54	95.33	**58.67**	86.51	**76.60**	**88.57**	**70.39**	58.51	88.95	**95.96**	**92.45**	93.59	77.50	**81.03**

Table 11: Test performance (%) of 3B models trained with different task difficulties on SEED-Bench downstream tasks. IC: Instances Counting, IA: Instance Attributes, SU: Scene Understanding, IId: Instance Identity, IIn: Instance Interaction, VR: Visual Reasoning, IL: Instance Location, SR: Spatial Relation, TU: Text Understanding.

Model	Difficulty	IC	IA	SU	IId	IIn	VR	IL	SR	TU	*Average*
Qwen2.5-VL-3B	–	60.52	62.87	60.35	63.24	64.95	53.78	58.79	51.60	41.67	60.83
Rotation	90-degree	64.12	71.03	73.65	72.80	67.01	73.41	61.76	54.03	45.24	69.10
	45-degree	60.60	69.26	72.70	72.53	68.04	72.81	63.60	55.25	38.10	67.81
Jigsaw	2x2	62.93	70.19	70.30	71.65	63.92	69.79	62.68	53.12	48.81	67.67
	3x3	61.22	67.58	69.28	68.87	**69.07**	64.35	61.35	51.14	48.81	65.66
Contrastive	Weak	54.03	61.22	67.10	68.38	64.95	67.07	63.70	51.75	28.57	61.90
	Strong	57.54	64.75	70.68	70.56	67.01	70.39	63.39	55.86	29.76	65.00
Position	2x2	**64.20**	72.51	73.56	72.75	62.89	70.69	**64.62**	55.25	**52.38**	69.77
	3x3	61.87	**72.53**	**73.84**	**73.89**	**69.07**	**74.02**	**64.62**	**58.30**	44.05	**69.80**

Table 12: Test performance (%) of 7B models trained with different task difficulties on MM-Bench downstream tasks. IR: Identity Reasoning, PPR: Physical Property Reasoning, FR: Function Reasoning, OL: Object Localization, SIU: Structuralized Imagetext Understanding, AtR: Attribute Recognition, FP: Future Prediction, SpR: Spatial Relationship, ISc: Image Scene, IQ: Image Quality, Ace: Action Recognition, AC: Attribute Comparison, IT: Image Topic, NR: Nature Relation, PR: Physical Relation, SoR: Social Relation, CR: Celebrity Recognition, ISt: Image Style, OCR: OCR, IE: Image Emotion.

Model	Difficulty	IR	PPR	FR	OL	SIU	AtR	FP	SpR	ISc	IQ	Ace	AC	IT	NR	PR	SoR	CR	ISt	OCR	IE	*Average*
Base	Qwen2.5-VL-7B	**98.30**	66.67	92.11	74.60	82.98	88.64	70.00	76.27	97.05	58.00	92.09	85.11	**91.43**	88.83	69.15	92.44	97.22	94.81	96.15	82.00	**86.37**
Jigsaw	3x3	**98.30**	69.41	92.11	76.83	86.88	90.91	70.77	**79.10**	97.54	62.00	92.56	87.94	90.00	89.94	**70.21**	**93.02**	97.47	95.28	**98.72**	84.50	86.17
	4x4	97.73	66.21	92.43	75.24	86.17	92.80	**73.08**	76.27	97.79	62.67	92.09	87.94	90.71	87.71	69.15	91.86	97.47	94.81	97.44	**85.00**	85.73
	5x5	97.73	67.58	92.76	74.29	85.46	90.15	**73.08**	**79.10**	97.05	61.33	93.02	87.94	90.71	89.94	**70.21**	91.28	97.98	**95.75**	95.51	84.00	85.74
Position	3x3	95.45	68.49	93.09	**77.46**	88.30	94.70	**73.08**	77.40	97.79	62.67	94.42	89.36	90.71	87.15	67.02	89.53	97.47	94.34	97.44	81.50	85.87
	4x4	97.73	64.84	**94.74**	74.92	88.65	95.45	**73.08**	72.32	97.79	58.00	94.42	82.98	89.29	88.83	69.15	90.12	**98.23**	94.34	98.08	82.50	85.27
	5x5	97.73	**72.15**	93.75	75.24	**90.07**	**95.83**	70.00	67.80	**98.53**	**63.33**	**95.81**	**90.07**	90.71	**91.62**	67.02	91.86	**98.23**	95.28	95.51	81.00	86.08

Table 13: Test performance (%) of models trained with different task difficulties on ImageNet-1K.

Model	Difficulty	Completion	Choice10	Choice200
Qwen2.5-VL-3B	–	24.93	85.22	57.10
Position	2x2	28.76	**92.35**	**67.14**
	3x3	27.37	88.99	59.93
Contrastive	Weak	26.84	89.51	61.78
	Strong	26.93	89.44	61.31
Rotation	90-degree	29.19	87.26	58.48
	45-degree	**29.91**	89.94	60.52

Table 14: Test performance (%) of 7B models trained with different task difficulties on SEED-Bench downstream tasks. IC: Instances Counting, IA: Instance Attributes, SU: Scene Understanding, IId: Instance Identity, IIn: Instance Interaction, VR: Visual Reasoning, IL: Instance Location, SR: Spatial Relation, TU: Text Understanding.

Category	Model	TU	VR	SU	IId	IIn	IA	IL	SR	IC	*Average*
Base	Qwen2.5-VL-7B	**72.62**	**77.95**	77.99	77.44	75.26	76.19	71.98	62.56	69.55	**74.70**
Jigsaw	3x3	69.88	77.01	**78.78**	77.88	**76.29**	**79.46**	71.57	**64.69**	73.81	74.37
	4x4	69.72	76.98	77.83	78.15	71.13	78.85	72.29	62.10	72.62	73.30
	5x5	69.84	76.43	78.34	78.15	71.13	78.85	**73.21**	63.17	70.24	73.26
Position	3x3	68.82	77.93	78.53	**78.75**	74.23	78.55	73.01	64.08	76.19	74.45
	4x4	68.98	77.41	77.58	77.33	**76.29**	78.55	71.47	61.64	76.19	73.94
	5x5	69.27	77.39	78.59	77.72	73.20	79.15	**73.21**	63.47	**77.38**	74.38

D Extension to Other Domains: An Example on Graph

Having established SSL4RL for vision-language reasoning, we now explore its broader applicability. The paradigm's core principle—generating rewards from data transformations—naturally extends to any domain with rich structural information. Beyond images, graph-structured data presents a compelling candidate, given its explicit relational semantics that are amenable to various pretext tasks. In this section, we empirically validate this potential by adapting SSL4RL to the graph domain.

Tasks and Reward Definitions. We introduce three graph-based SSL tasks, defined as follows: (1) ***Attribute Mask*** (Jin et al., 2020; Hu et al., 2019): A subset of node descriptions is randomly masked. The reward is quantified by the model's accuracy in reconstructing the original masked features. (2) ***Neighbor Prediction*** (Kipf & Welling, 2016): For a target node within a partially observed graph, the model is rewarded for correctly identifying its adjacent nodes. (3) ***Link Prediction*** (Hu et al., 2020; Hou et al., 2022): Given a pair of nodes and a partial graph structure, the model receives a reward for accurately classifying the presence or absence of an edge connecting them.

Benchmarks. We evaluate our method on benchmark datasets curated from TAGLAS (Feng et al., 2024), a comprehensive collection of text-attributed graphs. The evaluation encompasses two key tasks: (1) *Node-level classification* on the Cora and PubMed co-citation graphs and the WikiCS page relation graph; (2) *Link-level prediction* on the Products co-purchase graph and the fb15k237 and wn18rr knowledge graphs.

Results and Observations. The results in Table 15 and Table 16 demonstrate the successful application of SSL4RL to graph-structured data. The 3B model shows substantial improvements, with gains up to 13.79% on average, while the 7B model exhibits diminishing returns, mirroring our observations in the visual domain and reinforcing the "difficulty-capacity matching" principle. Furthermore, the relative effectiveness of the self-supervised tasks is contingent upon the nature of the downstream objective. Tasks emphasizing structural reasoning (Link and Neighbor Prediction) yield better performance on relation-centric tasks such as link prediction. Conversely, tasks focused on feature reconstruction (Attribute Mask) demonstrate a comparative advantage on node classification benchmarks. These findings not only validate the generalizability of the SSL4RL framework beyond the visual modality but also highlight the critical importance of aligning the pretext task's inductive bias with the target application.

Table 15: Test performance (%) of 3B models on downstream graph tasks.

Category	Model	Cora	PubMed	WikiCS	Products	fb15k237	wn18rr	*Average*
Base model	Qwen2.5-3B	21.80	64.26	30.50	7.93	26.30	29.80	30.09
SSL4RL-3B	Attribute	**55.80**	73.27	**57.62**	8.03	32.50	36.10	**43.88**
	Neighbor	39.00	**74.37**	50.84	**12.65**	36.10	36.60	41.59
	Link	31.30	71.97	55.93	4.91	**46.50**	**41.10**	41.95
Maximal Improvement		**↑ 34.00**	**↑ 10.11**	**↑ 27.12**	**↑ 4.42**	**↑ 20.20**	**↑ 11.30**	**↑ 13.79**

Table 16: Test performance (%) of 7B models on downstream graph tasks.

Category	Model	Cora	PubMed	WikiCS	Products	fb15k237	wn18rr	*Average*
Base model	Qwen2.5-7B	64.80	69.86	49.15	50.50	30.10	32.50	57.73
SSL4RL-7B	Attribute	63.80	59.55	54.23	**51.30**	**37.00**	32.70	49.76
	Neighbor	63.10	70.87	52.54	43.47	33.00	34.00	49.49
	Link	**67.70**	**72.27**	**55.93**	48.89	21.30	**39.50**	**50.93**
Maximal Improvement		**↑ 2.90**	**↑ 2.41**	**↑ 6.78**	**↑ 0.80**	**↑ 6.90**	**↑ 7.00**	**↑ 1.45**

E THE IMPACT OF TRAINING DATA VOLUME

A key question for data-driven methods is the ability to leverage increasing amounts of data. To characterize the relationship between performance and data scaling for SSL4RL, we progressively expand our training dataset and evaluate the resulting models on the MMBench benchmark. We design three training regimes with increasing data volume:

- **Base Set**: ∼4,000 samples from MMBench.
- **Extended Set**: ∼18,000 samples from a mixture of MMBench and SEED-Bench.
- **Full Set**: ∼118,000 samples from MMBench, SEED-Bench, and ImageNet.

We select the Position task as the representative SSL4RL method for evaluation. The results, detailed in Table 17 across almost all subtasks, demonstrate a clear positive scaling relationship. When scaling from the Base Set to the Extended Set, the average performance on MMBench improved from 80.08% to 81.38%, a gain of +1.30%. A further expansion to the Full Set yielded an additional +1.04% improvement, reaching 82.42%. This consistent, monotonic improvement suggests that **the effectiveness of SSL4RL continues to benefit from larger, more diverse datasets**. The performance gains are not uniform across all sub-categories, which provides deeper insight. For instance, the most significant improvements are observed in Cross-Instance Fine-grained Perception, which saw a substantial jump of 6.36 percentage points from the smallest to the largest dataset. This indicates that tasks requiring nuanced comparisons across different images benefit from exposure to a broader visual world. Conversely, capabilities like Relation and Attribute Reasoning showed more modest gains, potentially plateauing earlier or requiring more targeted data.

In summary, our analysis confirms that SSL4RL effectively translates increased data volume into enhanced performance. The experiments indicate the potential for further gains through even more extensive pre-training, establishing SSL4RL as a promising and scalable strategy for vision-language model alignment.

Table 17: Performance (%) on MMBench benchmark with increasing training volumes. Logical: Logical Reasoning, Relation: Relation Reasoning, Attribute: Attribute Reasoning, Coarse: Coarse Perception, Cross Inst.: Cross-Instance Fine-grained Perception, Single-Inst.: Single-Instance Fine-grained Perception.

Model	Training Volumes	Logical	Relation	Attribute	Coarse	Cross-Inst.	Single-Inst.	*Average*
Qwen2.5-VL-3B	–	61.77	41.54	76.62	73.55	64.32	82.06	72.99
SSL4RL-Position	4000	67.65	77.19	82.22	82.15	66.51	85.39	80.08
	18,000	68.71	79.26	**85.03**	81.65	70.16	86.47	81.38
	118,000	**68.77**	**79.29**	84.26	**83.66**	**72.87**	**88.31**	**82.42**

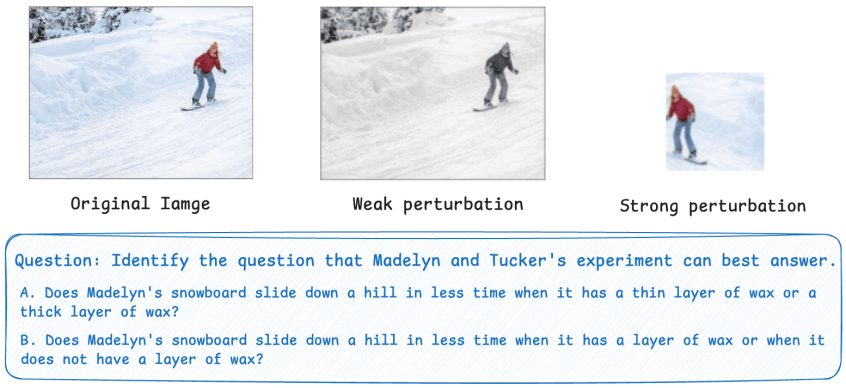


Figure 6: Perturbation examples on MMBench.

F Robustness Analysis

To evaluate our model's robustness, we conduct a comprehensive evaluation under various image perturbations. We design two levels of perturbation—weak and strong—to assess the models' resilience to visual corruptions. The weak perturbation applies a series of common image transformations, each with a probability of 0.5, including color jittering, random conversion to grayscale, Gaussian blur, and horizontal flipping. The strong perturbation builds upon the weak one by incorporating an additional multi-crop strategy. This strategy generates two 224×224 pixel crops per image via randomly resized cropping (scale range [0.08, 1.0]), presenting a more significant challenge to the model's perception. Examples of these perturbed images are provided in Figure 6.

We evaluate our SSL4RL models and the base model on the perturbed MMBench benchmarks. As shown in Table 18 and Table 19, **our method consistently demonstrates superior robustness across both perturbation levels**. For instance, under weak perturbation, our SSL4RL-Rotation model achieves an average score of 72.11%, outperforming the base model at 66.80%. This performance gap is maintained under strong perturbation, where our SSL4RL-Position model scores 64.37% compared to the base model's 59.62%. These results confirm that our SSL4RL strategy effectively enhances the model's robustness to photometric variations.

Table 18: Performance (%) on weakly perturbed MMBench. Logical: Logical Reasoning, Relation: Relation Reasoning, Attribute: Attribute Reasoning, Coarse: Coarse Perception, Cross Inst.: Cross-Instance Fine-grained Perception, Single-Inst.: Single-Instance Fine-grained Perception.

Category	**Model**	Logical	Relation	Attribute	Coarse	Cross-Inst.	Single-Inst.	***Average***
Base model	Qwen2.5-VL-3B	49.58	45.06	76.94	67.80	59.33	68.46	66.80
SSL4RL-3B	Rotation	**61.00**	**72.87**	78.66	70.18	**62.57**	72.89	**72.11**
	Jigsaw	55.47	70.80	**79.38**	**71.23**	56.28	71.84	70.84
	Contrastive	54.46	62.05	67.33	55.80	47.64	64.02	59.94
	Position	57.97	72.77	79.01	70.40	57.55	**73.39**	71.70

Table 19: Performance (%) on strongly perturbed MMBench. Logical: Logical Reasoning, Relation: Relation Reasoning, Attribute: Attribute Reasoning, Coarse: Coarse Perception, Cross Inst.: Cross-Instance Fine-grained Perception, Single-Inst.: Single-Instance Fine-grained Perception.

Category	**Model**	Logical	Relation	Attribute	Coarse	Cross-Inst.	Single-Inst.	***Average***
Base model	Qwen2.5-VL-3B	45.08	44.99	71.65	59.88	53.89	58.62	59.62
SSL4RL-3B	Rotation	52.69	60.47	70.96	64.33	**55.04**	62.51	63.38
	Jigsaw	46.45	59.62	72.86	63.90	52.40	62.47	62.00
	Contrastive	45.56	46.36	60.86	49.50	42.14	53.20	51.25
	Position	**53.43**	**61.37**	**72.95**	**65.85**	53.68	**65.12**	**64.37**

G Comparison with Verifier-Driven RL using Downstream Golden Rewards

To contextualize the performance of our SSL4RL method, we include a strong, task-specific baseline *VLM-R1* (Shen et al., 2025). It extends R1-style reinforcement learning to VLMs with rule-based reward formulation. Specifically, we randomly split downstream datasets into 60% for training and 40% for held-out testing. Then we fine-tune Qwen2.5-VL-3B using the GRPO algorithm, where the reward is computed by comparing it directly to the golden answer. We denote the tuned model as **Golden-3B**. Notably, the model is trained with **oracle reward signals**, and the setup simulates an ideal scenario where the reward model is perfectly aligned with the evaluation metric. The results are presented in Table 20 and Table 21, where we draw two key conclusions:

- **The Oracle Baseline has a high but finite ceiling.** Even with direct supervision from golden answers, the Golden-3B model achieves 84.93% on MMBench and 73.21% on SEED-Bench. This indicates inherent challenges in these benchmarks that are not fully solved even with ideal, verifiable rewards.
- **Our SSL4RL method is highly competitive.** The performance gap between our best SSL4RL variants and the Golden-3B oracle is relatively small (*e.g.*, 81.35% vs. 84.93% on MMBench, and 69.80% vs. 73.21% on SEED-Bench), compared to our improvements over the base model. This demonstrates that our self-supervised objectives, which require *no labeled downstream data*, can effectively close the gap to the performance driven by idealized, task-specific reward signals.

This comparison paints an encouraging picture for SSL4RL. Our method, which deliberately avoids using any downstream labels, can close much of the gap to a model trained with direct oracle signals. It suggests that the reinforcement learning process can be steered effectively by self-supervised objectives. While perfect verifiable rewards remain a powerful tool, our work shows that highly competitive performance can be achieved through a more scalable and generalizable pathway.

Table 20: Performance (%) of on the MMBench test set. Logical: Logical Reasoning, Relation: Relation Reasoning, Attribute: Attribute Reasoning, Coarse: Coarse Perception, Cross Inst.: Cross-Instance Fine-grained Perception, Single-Inst.: Single-Instance Fine-grained Perception.

Category	**Model**	Logical	Relation	Attribute	Coarse	Cross-Inst.	Single-Inst.	***Average***
Base model	Qwen2.5-VL-3B	53.03	46.19	77.10	74.56	67.94	81.41	72.63
RL with golden rewards	Golden-3B	71.82	86.70	88.34	83.74	79.57	89.39	84.93
SSL4RL-3B	Rotation	66.06	**82.97**	**87.38**	80.07	**75.07**	85.93	**81.35**
	Jigsaw	62.53	79.06	85.84	77.39	66.43	84.14	78.46
	Contrastive	60.50	71.50	73.75	66.31	56.17	77.72	68.82
	Position	**67.58**	78.37	85.16	**81.22**	70.70	**87.87**	81.12

Table 21: Performance (%) on the SEED-Bench test set. TU: Text Understanding, VR: Visual Reasoning, SU: Scene Understanding, IId: Instance Identity, IIn: Instance Interaction, IA: Instance Attributes, IL: Instance Location, SR: Spatial Relation, IC: Instances Counting.

Category	**Model**	TU	VR	SU	IId	IIn	IA	IL	SR	IC	***Average***
Base model	Qwen2.5-VL-3B	59.04	60.44	67.17	63.83	61.80	51.28	57.20	59.09	59.54	62.00
RL with golden rewards	Golden-3B	67.68	68.41	74.76	77.03	74.69	82.05	63.84	77.27	77.10	73.21
SSL4RL-3B	Rotation	62.60	62.64	72.28	73.69	71.28	58.46	55.35	**68.18**	**77.10**	69.01
	Jigsaw	60.16	60.99	71.31	74.01	69.53	58.46	**59.78**	63.64	74.81	67.97
	Contrastive	55.08	**64.01**	69.79	68.76	61.59	30.77	56.09	56.82	74.05	63.02
	Position	**63.21**	62.36	**72.83**	**74.17**	**72.40**	**58.97**	57.93	65.91	73.28	**69.80**

H Ablation Study on Base Models

In this section, we evaluate whether the benefits of SSL4RL extend to architectures beyond the Qwen2.5-VL series. For this purpose, we select Gemma3-4B (Team et al., 2025), a recent and

powerful model from Google with a distinct architecture. To ensure a fair comparison under computational constraints, we adapted our training setup by halving the batch size to 256 while meticulously preserving all other hyperparameters and training procedures. Results presented in Table 22 provide strong evidence for the general applicability of our method. On the Gemma3 base model, our SSL4RL models yield a consistent and notable average performance improvement of 2.88% on MMBench. **The gain is particularly pronounced in Cross-instance Perception (5.76%), Logical Reasoning (4.35%), and Attribute Reasoning (4.13%).** Moreover, the relative efficacy of the individual SSL4RL tasks is remarkably consistent with our prior findings. As with Qwen2.5-VL, tasks like Rotation and Position confer the most significant benefits, while the simpler Contrastive task shows more modest gains. The consistent results validate that SSL4RL is a general-purpose principle for VLMs, not an artifact of a particular model family.

Table 22: Test performance (%) on MMBench downstream tasks. The base model is Gemma3-4B. Logical: Logical Reasoning, Relation: Relation Reasoning, Attribute: Attribute Reasoning, Coarse: Coarse Perception, Cross Inst.: Cross-Instance Fine-grained Perception, Single-Inst.: Single-Instance Fine-grained Perception.

Category	**Model**	Logical	Relation	Attribute	Coarse	Cross-Inst.	Single-Inst.	***Average***
Base	Gemma3-4B	63.12	80.56	83.24	80.43	68.65	78.85	78.30
SSL4RL	Rotation	**67.47**	81.16	86.12	**83.27**	72.35	**82.09**	**81.38**
	Jigsaw	63.77	**84.39**	86.69	82.99	73.72	80.67	81.10
	Contrastive	63.89	81.94	**87.37**	82.30	71.79	81.70	80.75
	Position	64.81	82.26	86.54	83.08	**74.41**	80.89	81.15

I EVALUATION ON VISION-LANGUAGE BENCHMARKS

In this section, we evaluate the SSL4RL strategy on broader vision-language benchmarks, including Vision Questioning Answer (VQA) tasks and the open-ended image-caption task.

I.1 VISUAL QUESTION ANSWERING (VQA) TASKS

Besides MMBench and SEED-Bench, we further consider four representative VQA benchmarks, V*(Wu & Xie, 2024), RealWorldQA (xAI, 2024), BLINK (Fu et al., 2024), MME-RealWorld-Lite (Zhang et al., 2024), featuring challenging current VLMs' ability of visual perception, spatial understanding, detail capturing, real-world applications, and so on. The detailed descriptions of benchmarks are as follows:

- **MMBench** (Liu et al., 2024): A diverse benchmark with over 3,000 multiple-choice questions spanning 20 distinct ability dimensions. All results on MMBench are reported for the DEV-EN split, showing the average performance per category (detailed per-dimension results are in Appendix A).
- **SEED-Bench** (Li et al., 2023): A comprehensive benchmark with human annotations for image and video modalities. For evaluation, we select the 9 core dimensions pertaining to image understanding, which comprise 14,232 examples in total.
- **V*** (Wu & Xie, 2024): A benchmark based on 191 high-resolution images with an average image resolution of 2246×1582. It is specifically designed to quantitatively evaluate VLMs' ability in challenging scenarios where the image contains abundant and complex information, and the visual information needed might not be easily found.
- **RealWorldQA** (xAI, 2024): A benchmark designed to evaluate the real-world visual understanding capabilities of VLMs. It assesses how well these models comprehend physical environments. The benchmark consists of 700+ images drawn from real-world scenarios, including those captured from vehicles.
- **BLINK** (Fu et al., 2024): A benchmark for VLMs that focuses on core visual perception abilities, including relative depth estimation, visual correspondence, forensics detection, and multi-view reasoning, etc. BLINK contains visual commonsense problems that humans can answer within seconds, rarely requiring domain knowledge.

- **MME-RealWorld-Lite** (Zhang et al., 2024): A benchmark contains 13K high-quality images annotated humans, resulting in 29K question-answer pairs that cover 43 subtasks across 5 real-world scenarios, featuring the highest resolution and a targeted focus on real-world applications. For inference acceleration, we utilize its public lite version MME-RealWorld-Lite with 50 samples per task.

We consider the four SSL4RL strategies: Rotation, Jigsaw, Contrastive, and Position, under the same training and evaluation settings in Section 4. We report the general performances in Table 24 and Table 28. The results consistently demonstrate the effectiveness of SSL4RL. For the 3B models (Table 24), **our method provides a substantial average performance gain of 7.27% over the base model**. The improvements are particularly pronounced on V* (+8.90%) and RealWorldQA (+9.55%), underscoring that SSL4RL significantly enhances the model's ability to process fine-grained details in complex, real-world images. These consistent results across a diverse set of benchmarks verify the effectiveness of the proposed SSL4RL framework.

Table 23: Test performance (%) on SEED-Bench downstream tasks. TU: Text Understanding, VR: Visual Reasoning, SU: Scene Understanding, IId: Instance Identity, IIn: Instance Interaction, IA: Instance Attributes, IL: Instance Location, SR: Spatial Relation, IC: Instances Counting.

Category	Model	TU	VR	SU	IId	IIn	IA	IL	SR	IC	*Average*
Base	Qwen2.5-VL-3B	41.67	53.78	60.35	63.24	64.95	62.87	58.79	51.60	60.52	60.83
SSL4RL	Rotation	45.24	**73.41**	**73.65**	**72.80**	**67.01**	71.03	61.76	54.03	64.12	69.10
	Jigsaw	48.81	69.79	70.30	71.65	63.92	70.19	62.68	53.12	62.93	67.67
	Contrastive	28.57	67.07	67.10	68.38	64.95	61.22	63.70	51.75	54.03	61.90
	Position	**52.38**	70.69	73.56	72.75	62.89	**72.51**	**64.62**	**55.25**	**64.20**	**69.77**
Maximal Improvement		**↑ 10.71**	**↑ 19.63**	**↑ 13.30**	**↑ 9.56**	**↑ 2.06**	**↑ 9.64**	**↑ 5.83**	**↑ 3.65**	**↑ 3.68**	**↑ 8.94**

Table 24: Test performance (%) on extended VQA benchmarks.

Category	Model	V*	RealWorldQA	BLINK	MME-RealWorld-Lite	*Average*
Base	Qwen2.5-VL-3B	59.16	52.67	42.13	32.41	45.85
SSL4RL-3B	Rotation	62.30	62.09	**48.18**	34.18	51.68
	Jigsaw	63.35	**62.22**	45.18	35.12	51.46
	Contrastive	60.20	58.03	45.13	30.17	48.38
	Position	**68.06**	59.86	46.39	**38.19**	**53.12**
Maximal Improvement		**↑ 8.90**	**↑ 9.55**	**↑ 6.05**	**↑ 5.78**	**↑ 7.27**

I.2 Open-ended Vision-Language Task: Image Captioning

Relying solely on the data itself, SSL4RL requires no human labels, external verifiers, or heuristic judges, yet produces dense and scalable reinforcement signals. Intuitively, it has potential for open-ended tasks where ground truth is ill-defined and human annotations are expensive. To evaluate SSL4RL's potential on open-ended tasks, we leverage a recent image captioning platform, CapArena (Cheng et al., 2025), which contains over 6,000 human-annotated pairwise preference battles. In CapArena, captions from a test model are compared against those from strong baseline models (GPT-4o, CogVLM-19B, or MiniCPM-8B) using a LLM (GPT-4o) as a judge, with human references provided for context. The winner is assigned +1, the loser with -1, and 0 for a draw in each pairwise comparison. This challenging task mirrors the real-world challenge of improving a model without a single clear and correct answer.

We evaluate our four SSL4RL strategies on CapArena against the base Qwen2.5-VL-3B model. For easier comparison, we present the final scores after min-max normalization in Table 25. **SSL4RL consistently improves over the base model, with the largest performance gain of 8.14 points (56.45 vs. 48.31)**. This clearly shows that the self-supervised rewards provide a meaningful learning signal even in the absence of a verifiable ground truth.

Qualitative Insights. We conduct a qualitative comparison in Table 26, shed light on *how* SSL4RL enhances open-ended generation:

- **Enhanced Detail Capture.** SSL4RL models correctly capture more details in the image. For instance, our SSL4RL model correctly identifies a "scoreboard" displaying "1 0 0" and a "disabled persons' sign," which the base model either misinterprets or omits.
- **Improved Spatial Reasoning.** SSL4RL models tend to use more precise spatial descriptors like "*behind the fence*," "*to the left of*," and "*the center of the court*", aligning well with the spatial reasoning ability required by SSL pretext tasks.

In summary, this experiment demonstrates that SSL4RL is not limited to tasks with easily verifiable rewards. By providing a dense, automated learning signal derived from the data's intrinsic structure, our method successfully improves model performance on the complex, open-ended task of image captioning.

Table 25: Performance of 3B-models on CapArena Platform. GPT-Score: the score compared with GPT-4o. Cog-Score: the score compared with CogVLM-19B. CPM-Score: the score compared with MiniCPM-8B. Average: the average score. Avg-Length: the average response length.

Category	**Model**	GPT-Score	Cog-Score	CPM-Score	***Average***	Avg-Length
Base	Qwen2.5-VL-3B	0.00	6.48	92.96	48.31	89.71
SSL4RL-3B	Rotation	19.15	49.30	96.48	55.28	92.91
	Jigsaw	8.87	54.93	98.59	55.64	87.48
	Contrastive	4.96	57.95	100.00	56.45	93.14
	Position	0.00	59.15	92.25	51.04	90.48

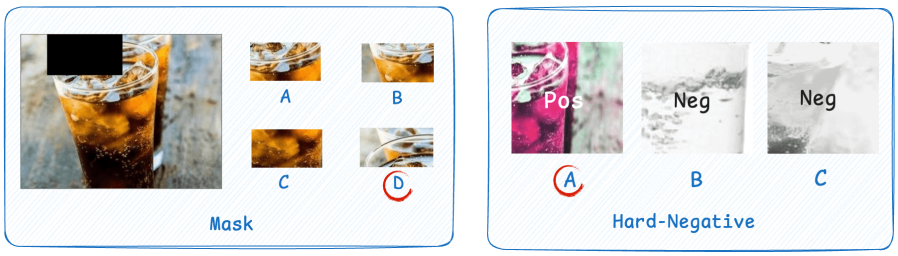


Figure 7: Illustrations of Harder SSL4RL Task: Mask and Hard-Negative

J TRAILS ON HARDER SSL4RL TASKS

Beyond the four SSL4RL tasks, we further explore two more challenging tasks on 7B models:

- **Mask**. Guided by the insights from Masked AutoEncoders (MAEs) (He et al., 2022; Chen et al., 2023; 2024c), we adopt the self-supervised strategy of mask-and-reconstruct. We randomly sample four patches from an image and mask one of them in the original view. The model must then identify which of the four patches is missing, a concurrent idea explored in Liu et al. (2025b).
- **Hard-Negative Contrastive.** Moving beyond the simple binary task Contrastive in Section 4, we increase the discrimination difficulty by introducing hard negatives (Xuan et al., 2020; Kalantidis et al., 2020; Ge et al., 2021). For a given image, we use a pretrained DINOv2 (Oquab et al., 2023) encoder to find its most semantically similar images from the dataset. The model's task is to identify the positive augmented view from two hard negative views, forcing it to learn finer-grained visual distinctions.

As shown in Table 27, on MMBench, our SSL4RL-7B models surpass the base model by an average of 1.41% (87.78% vs. 86.37%), and the Hard-Negative task drives a 5.23% improvement on the Logical Reasoning sub-task. **Crucially, extending our evaluation to four additional benchmarks reveals even larger improvements for the SSL4RL-7B models.** As shown in Table 28, our SSL4RL models outperform the base model by a significant margin of 5.05% on average (62.61%

Table 26: Examples: Image-Captioning Comparison on CapArena.

Image	Caption
	**Qwen2.5-VL-3B:** A view of a tennis court through a chain-link fence. The court features a red and blue surface with white boundary lines. The fence has a sign indicating "100" and "WON" on the top right side. The background includes a fence and greenery, suggesting an outdoor setting. **SSL4RL-Rotation:** The image depicts a tennis court viewed through a black chain-link fence. The court features a red and blue surface with white lines marking the boundaries and service boxes. A net is stretched across the center of the court, supported by black posts. Behind the fence, a scoreboard displays the score "1 0 0" indicating that one player has won all three sets. The background includes a chain-link fence enclosing the tennis court and some greenery beyond it. The sky appears clear, suggesting a sunny day.
	**Qwen2.5-VL-3B:** A white Porsche sports car is parked on the side of a road in a desert-like environment. The car is positioned near a "Reserved Parking" sign, which is mounted on a pole and partially obscured by a large fallen tree trunk. The surrounding area is rocky and barren, with sparse vegetation in the background. The sky is clear and blue, indicating a sunny day. The car's sleek design and modern features are prominently displayed, emphasizing its status as a high-performance vehicle. **SSL4RL-Rotation:** A white Porsche sports car is parked in a reserved parking spot on a rocky, barren landscape with sparse vegetation and a clear blue sky above. The car's sleek design and modern features are prominently displayed, contrasting with the rugged surroundings. A "Reserved Parking" sign is visible to the left of the car, indicating designated parking for individuals with disabilities. The overall scene suggests a location that could be part of a natural reserve or an eco-friendly area.

vs. 57.56%), particularly on complex VQA benchmarks like BLINK (+6.73%) and V* (+5.76%). These results are consistent with the positive findings on 7B models in concurrent works Liu et al. (2025b); Wang et al. (2025), verifying that the SSL4RL strategy is effective and has potential when scaled to larger models.

Table 27: Test performance (%) of 7B models on MMBench. Logical: Logical Reasoning, Relation: Relation Reasoning, Attribute: Attribute Reasoning, Coarse: Coarse Perception, Cross Inst.: Cross-Instance Fine-grained Perception, Single-Inst.: Single-Instance Fine-grained Perception.

Category	Model	Logical	Relation	Attribute	Coarse	Cross-Inst.	Single-Inst.	*Average*
Base	Qwen2.5-VL-7B	76.49	84.68	85.69	84.66	84.49	89.15	86.37
SSL4RL-7B	Mask	80.16	83.52	**86.94**	84.50	84.89	89.81	86.85
	Hard-Contrastive	**81.72**	**85.71**	86.21	**85.43**	**86.17**	**90.98**	**87.78**

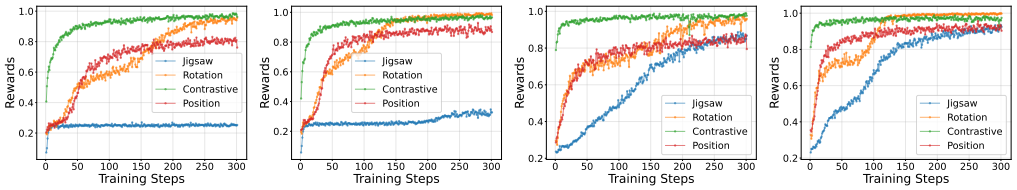
Table 28: Test performance (%) of 7B-models on extended VQA benchmarks.

Category	Model	V*	RealWorldQA	BLINK	MME-RealWorld-Lite	*Average*
Base	Qwen2.5-VL-7B	73.29	65.88	45.50	45.59	57.56
SSL4RL-7B	Mask	78.01	68.88	49.28	**49.03**	61.30
	Hard-Contrastive	**79.05**	**70.58**	**52.23**	48.61	**62.61**
Maximal Improvement		↑ **5.76**	↑ **4.70**	↑ **6.73**	↑ **3.44**	↑ **5.05**

K RL TRAINING CURVES

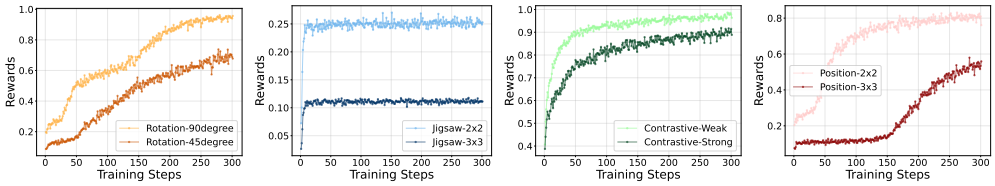
K.1 RL REWARD CURVES

Figure 8 illustrates the evolving reward signals of different SSL tasks during reinforcement learning training. Despite being intuitive for humans, these SSL tasks pose significant challenges for the base VLM. For instance, Qwen2.5-VL-3B initially achieves only about 20% accuracy on the Rotation task and nearly zero on the Jigsaw task, highlighting its limited low-level perceptual capability regarding spatial and structural image properties. Position and Rotation tasks exhibit gradual learning curves, plateauing at a reward of about 0.8 after approximately 200 epochs. For the harder Jigsaw task, the reward surges and plateaus within 20 epochs for 3B models, while 7B models present more smooth reward curves, indicating a stronger learning capability of large-scale models. For the Contrastive task, models rapidly surge to 1.0 within 50 epochs due to the task's easy complexity. Figure 9 presents the reward curves of tasks with different difficulties. For the same task, increasing the difficulty slows down the learning process and lowers the final converging rewards, showing the sensitivity of the models to tasks' difficulty.



(a) MMBench (3B) (b) SEED-Bench (3B) (c) MMBench (7B) (d) SEED-Bench (7B)

Figure 8: Reward curves of SSR4RL models during reinforcement learning.



(a) Rotation (b) Jigsaw (c) Contrastive (d) Position

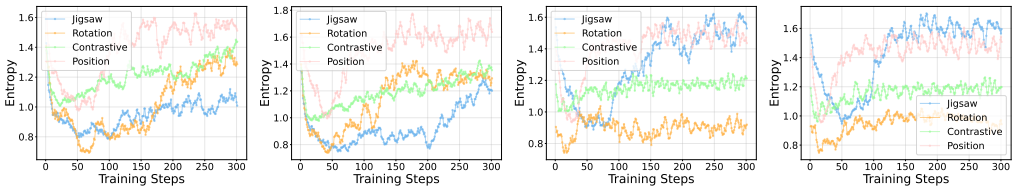
Figure 9: Rewards of SSL4RL 3B-models on MMBench, comparing different difficulties.

K.2 RL ENTROPY CURVES

Figure 10 presents the entropy trajectory of SSL4RL models during reinforcement training. In general, the entropy shows a decrease-then-increase trend. This is an emergent signature of a successful optimization process balancing two competing objectives: ***specialization*** and ***generalization***. At the first stage, the learning rewards incentivize a lower-entropy, specialized policy for high performance on the target task, resulting in a sharpening of the probability distribution. For a given prompt, the model becomes more confident in a subset of the vocabulary, directly leading to a reduction in entropy. At the second stage, to reduce the costly KL penalty and prevent mode collapse, the policy model slightly broadens probability distributions. This incentivizes a higher-entropy policy to maintain linguistic diversity and prevent catastrophic forgetting of the pre-training distribution. The two-stage entropy curves reflect a sustainable compromise between maximizing reward and preserving the foundational knowledge and generative diversity, supporting our experimental findings that SSL4RL models can generalize well to downstream vision-language tasks. S

K.3 RESPONSE LENGTH CURVES

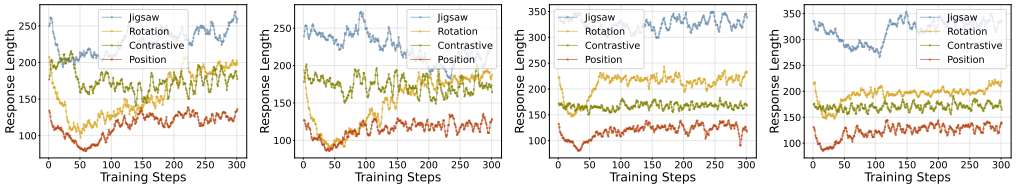
As illustrated in Figure 11, reinforcement learning induced no substantial change in model response length, with variations remaining within a 50-token margin. This is different from DeepSeek-R1



(a) MMBench (3B) (b) SEED-Bench (3B) (c) MMBench (7B) (d) SEED-Bench (7B)

Figure 10: Entropy of SSR4RL models during reinforcement learning.

(DeepSeek-AI et al., 2025), which reports longer responses featuring complex chains of thought after RL on complex mathematical and coding problems. The disparity is attributable to the nature of our SSL4RL tasks, which are perception-centric and necessitate less extensive reasoning. The second observation is that the harder Jigsaw task requires longer responses than other tasks, a correlation between task difficulty and response length that aligns with findings from Wang et al. (2025).



(a) MMBench (3B) (b) SEED-Bench (3B) (c) MMBench (7B) (d) SEED-Bench (7B)

Figure 11: Response Lengths of SSR4RL models during reinforcement learning.

L Case Analysis and Visualization

In Figure 12, we show that SSL4RL mitigates over-reliance on linguistic priors, fostering greater dependence on visual evidence. For example, when asked about a chandelier's color, the base model defaults to a common-sense response (*e.g.*, a typical decorative color), while our model first localizes the object and then answers based on the actual appearance.

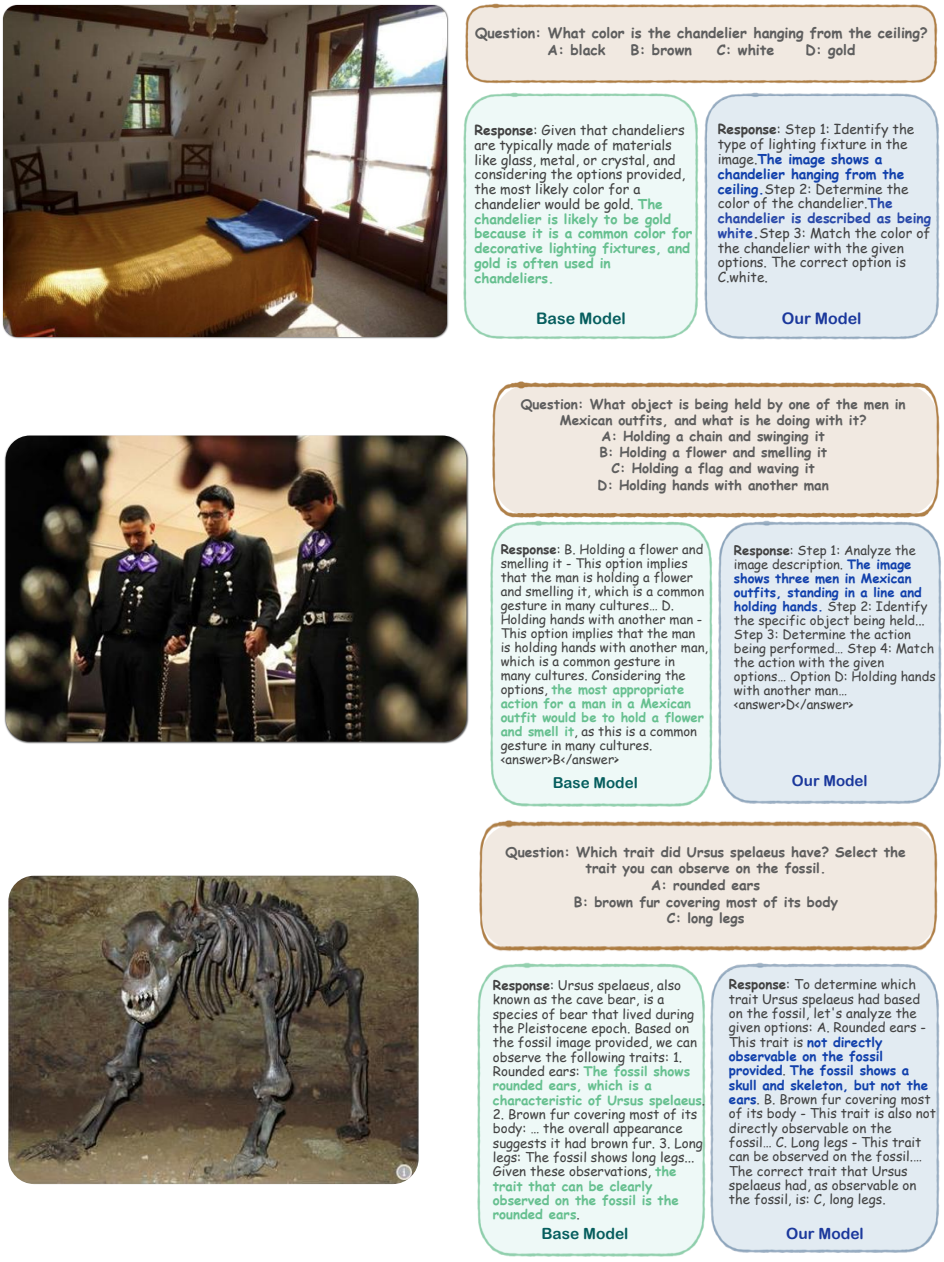


Figure 12: Instances of VLMs' loss on image information. After receiving textual instructions, VLMs may be more inclined to rely on the encoded textual knowledge for reasoning rather than carefully observing the content of the image.

Besides, we visualize the attention maps of the baseline model, *i.e.*, Qwen2.5-VL-3B and our models on several examples from the SEED-Bench dataset(Li et al., 2023). We pick a dominant token from the questions of each example, calculate the attention map of the first generated token to that input token, and average the attention matrices of all heads and all layers of the language model. The

results in Figure 13 illustrate that our models consistently display more focused attention towards the regions in the images corresponding to the selected token, which indirectly proves the superior performance of our models.

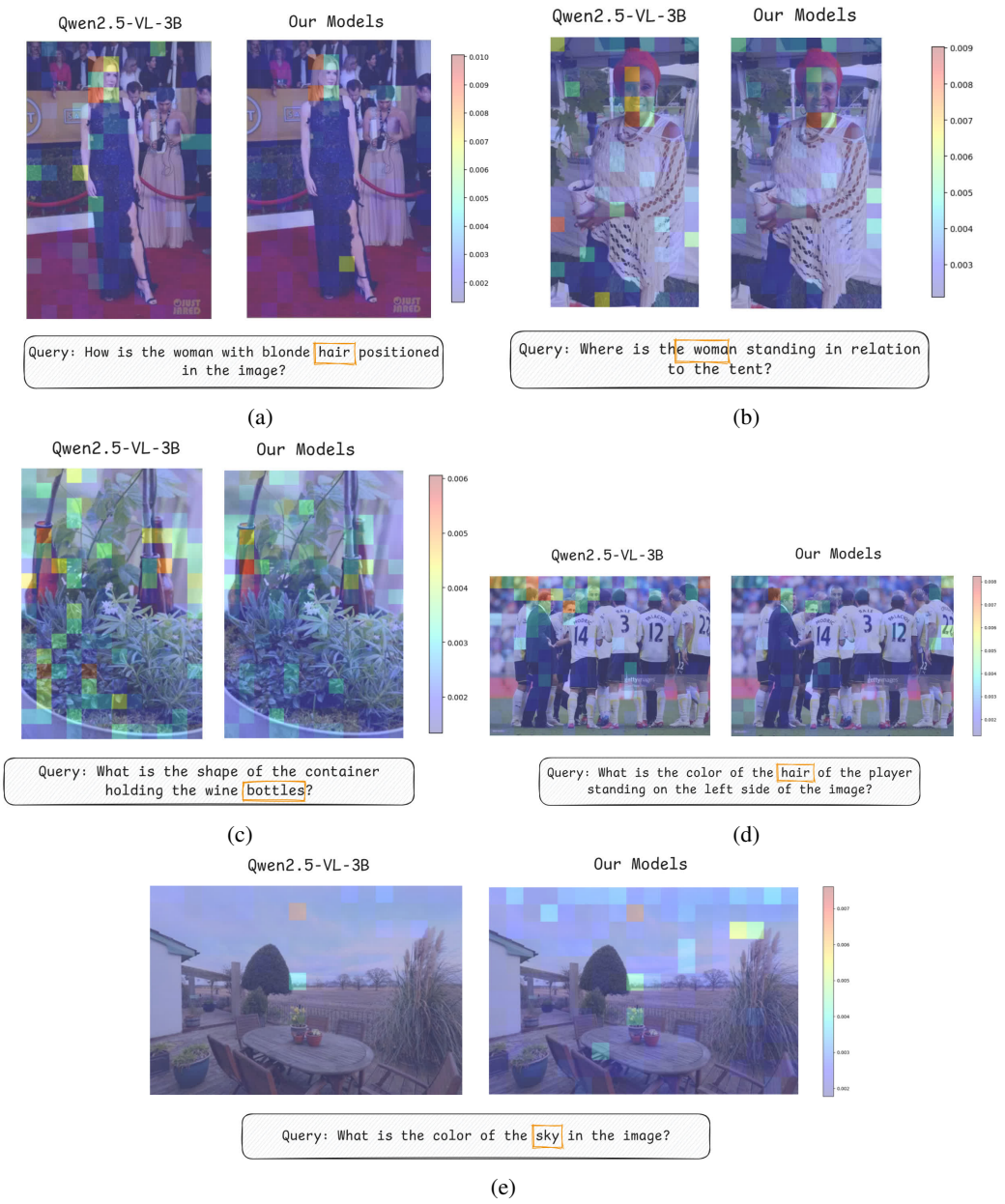


Figure 13: Comparisons of Attention Maps.

M SSL TASK EXAMPLES

In this section, we show a specific instance of Rotation, Jigsaw, Contrastive, and Position tasks to illustrate the SSL task design.

Rotation Example

Query: These are two images. The second image is a rotated version of the first image. Please determine how many degrees the second image has been rotated **counter-clockwise** relative to the first image.
You must reason step-by-step and then provide the final answer. The output **must strictly follow** this format: <think>your reasoning here </think><answer>number_of_degrees</answer>.

Answer: 270

Jigsaw Example

Query: <image><image><image><image>
<image><image><image><image><image>
The provided images represent 9 parts of an original image, divided into a 3x3 grid.
Your task is to determine the correct order of these parts to reconstruct the original image. Starting from the top-left corner, proceed row by row, from left to right and top to bottom, to arrange the parts.
The output should be a string of numbers, separated by a comma, where each number corresponds to the original position of the patches in the restored image. For instance, "3,1,9,2,8,5,4,6,7" would indicate the positions of the patches in the correct order.
Before providing the final result, you must reason through the puzzle step by step. Consider the relative placement of each part and how they fit together.
Your answer should strictly follow this format:
<think>your step-by-step reasoning here</think><answer>order</answer>

Answer: 2,7,6,1,3,5,9,8,4

Contrastive Example

Query: <image><image>
The provided images are augmentations of the same original image or two different images. The augmentations may include random cropping, color adjustments, grayscale conversion, blurring, and flipping. Please think step-by-step and determine if these two images are possibly derived from the same original image. If the provided images are from the same original image, respond with "positive"; if they correspond to different original images, respond with "negative".
Your answer should strictly follow this format:
<think>your step-by-step reasoning here</think><answer>positive/negative</answer>

Answer: positive

Position Example

Query: <image><image>
The second image in an augmented version of a crop in the first image. The augmentations may include grayscale, color jitter, solarization, etc. Please determine which part of the first image the second image is from. The second image is partitioned into 3x3 parts, and the first image can be only from one of the parts, but cannot be across two parts. The answer should be in the format of x/y, where x is the row number (from top to bottom) and y is the column number (from left to right). For example, 1/1 indicates the top-left part, and 1/3 indicates the top-right part. Both x and y may take values from 1 to 3.
Your answer should strictly follow this format:
<think>your step-by-step reasoning here</think><answer>x/y</answer>

Answer: 3/3

N Downstream Benchmark Example

In this section, we show a specific instance of Rotation, Jigsaw, Contrastive, and Position tasks to illustrate the SSL task design.

Imagenet-Completion Example

Query: <image>This is an image containing an object. Please identify the species of the object based on the image. The output answer format should be as follows: <think>... </think><answer>species name</answer>. Please strictly follow the format.
Answer: tench, Tinca, tinca

Imagenet-Choice10 Example

Query: <image>This is an image containing an object. Please identify the species of the object based on the image. The output answer format should be as follows: <think>... </think><answer>species name</answer>
Please strictly follow the format.

Please select the correct species name from the following options: Ursus americanus, shoe shop, brush wolf, essence, malemute, scoreboard, tench, ruddy turnstone, Salamandra salamandra, koala.

Answer: tench

MMBench Example

Query: <image>Identify the question that Madelyn and Tucker's experiment can best answer.

A. Does Madelyn's snowboard slide down a hill in less time when it has a thin layer of wax or a thick layer of wax?
B. Does Madelyn's snowboard slide down a hill in less time when it has a layer of wax or when it does not have a layer of wax?
C. NaN.
D. NaN.

Answer: B

SEED-Bench Example

Query: <image>How many towels are in the image?

A. One.
B. Two.
C. Three.
D. Four.

Answer: A