VOLD: Reasoning Transfer from LLMs to Vision-Language Models via On-Policy Distillation

Supplementary Material

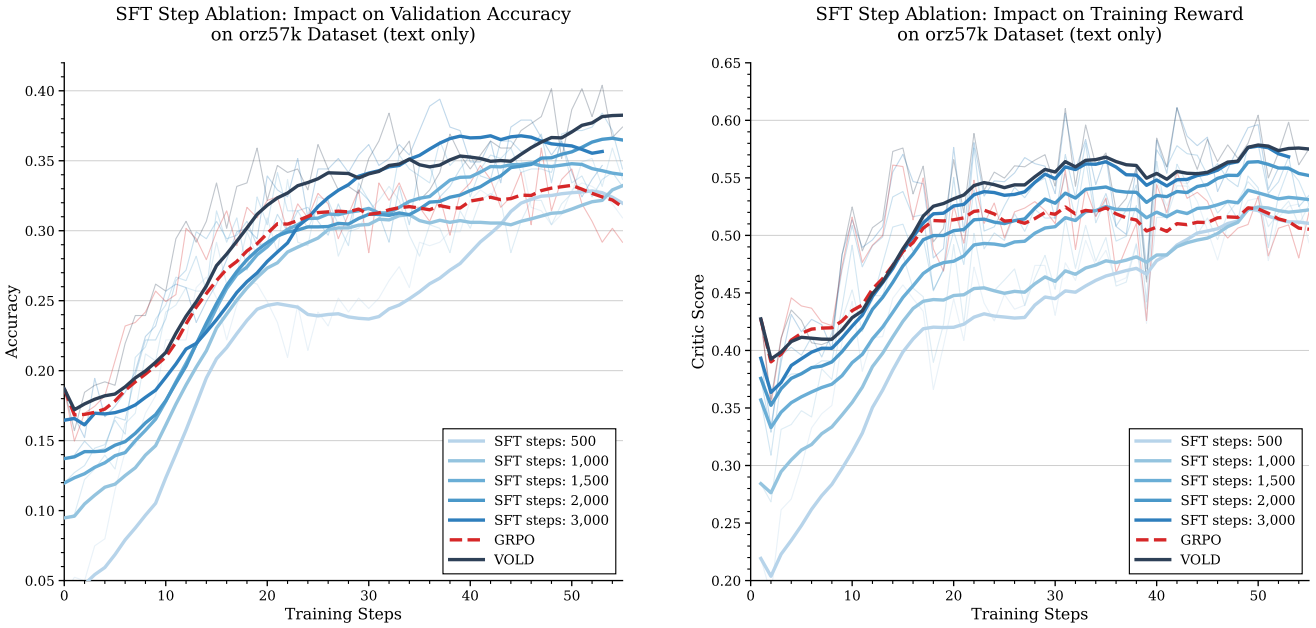

Figure 4. **Sufficient Policy Alignment is Crucial for On-Policy Distillation**. This figure illustrates that the benefit of our unified objective depends on the quality of the initial alignment from Stage 1. Models with short SFT phases (light blue) are poorly aligned with the teacher and fail to benefit from its guidance. As the alignment improves with more SFT steps (darker blue), the student can better leverage the on-policy distillation signal, unlocking significant performance gains over the GRPO-only baseline (red).

6. Implementation details

Detailed training hyperparameters for both stages of the VOLD framework are provided in Table 4. SFT experiments were conducted using 32 A100 GPUs and RL on 4 A100 GPUs, with gradient accumulation to achieve the specified effective batch sizes.

7. More Ablations

7.1. Impact of Cold Start

To understand when on-policy distillation becomes beneficial and how SFT duration affects subsequent RL training effectiveness, we evaluate different cold start checkpoints from various stages of teacher-trace SFT on MoT-Teacher-8B. We apply identical RL training with on-policy distillation to checkpoints saved at SFT steps 500, 1000, 1500, 2000, 2500, 3000, 3500, and 4000. Figure 4 shows the resulting training dynamics for both validation accuracy on Geo3K (left) and training reward (right), where the color gradient represents different starting checkpoints: light red corresponds to 500 SFT steps, progressing through darker reds to black representing 4000 SFT steps. The yellow curve shows the GRPO-only baseline starting from the 4000-step checkpoint. Early SFT checkpoints (light red, e.g., step 500) initially perform worse than the base model and show no benefit from distillation, indicating that minimal teacher-trace exposure is insufficient for distributional alignment. Performance progressively improves as SFT training continues, with the student's output distribution gradually converging toward the teacher's, effectively establishing a "breadcrumb trail" that guides subsequent RL exploration. Later checkpoints (darker colors approaching black) demonstrate substantial improvements over the GRPO-only baseline, as the student can better follow the teacher's reasoning patterns through on-policy distillation. The dynamics plateau after approximately 3000 SFT steps, suggesting that the student has sufficiently internalized the teacher's reasoning style and established a robust breadcrumb trail for RL guidance. These results demonstrate that adequate cold start training is crucial for creating the distri-

Table 4. Training hyperparameters for Stage 1 (SFT) and Stage 2 (RL) in the VOLD framework.

Hyperparameter	**Stage 1 (SFT)**	**Stage 2 (RL)**
Learning Rate	5×10^{-5}	6×10^{-6}
Batch Size	256	256
Training Steps	4000	60
Rollouts per Prompt	-	5
KL Coefficient (β)	-	1×10^{-3}
Clipping Threshold (ϵ_{upper})	-	0.3
Clipping Threshold (ϵ_{lower})	-	0.2
Max Sequence Length	8192	8192
Optimizer	AdamW	AdamW
Weight Decay	1×10^{-2}	1×10^{-2}
Warmup Steps	150	5
LR Schedule	Cosine	Cosine
Vision Encoder	Frozen	Frozen

butional bridge necessary for successful knowledge transfer during the unified RL+On-policy Distillation phase.

7.2. Teacher Size Ablation

To investigate the impact of teacher capacity on student performance, we compare VOLD training using different teacher model sizes during the RL+KD phase. We evaluate three teacher configurations: Qwen3-4B, Qwen3-8B, and Qwen3-14B, all sharing the same tokenizer with the Qwen2.5-VL-3B student to enable meaningful KL divergence computation. Table 5 presents results across representative benchmarks for each teacher size. The results demonstrate that larger teacher models generally provide better guidance, with the 8B teacher outperforming the 4B teacher on most benchmarks, particularly on MMStar (55.2% vs 53.66%) and LogicVista (44.97% vs 42.95%). This improvement can be attributed to the superior reasoning capabilities of larger teachers, which provide more valuable supervision during on-policy distillation. However, the performance gains begin to saturate beyond the 8B scale, with the 14B teacher showing similar performance compared to the 8B variant on certain tasks. This saturation suggests that the 3B student model has inherent capacity limitations that prevent it from fully leveraging the guidance from very large teachers. The student's reasoning capabilities appear to plateau once the teacher reaches sufficient quality, indicating there is only so much knowledge the student can effectively absorb through on-policy distillation given its architectural constraints.

7.3. Effect of Reward-Guided KL Masking

To evaluate the impact of our reward-guided KL masking mechanism, we compare three configurations during RL training: vanilla GRPO, VOLD with KL masking (our full method), and VOLD without KL masking. As shown in Figure 5, the reward-guided masking provides a clear benefit over both alternatives. While VOLD without masking (purple line) outperforms vanilla GRPO (red line), achieving approximately 0.56 vs 0.51 final reward, our complete VOLD framework with KL masking (blue line) achieves the highest performance at 0.58. The consistent gap throughout training demonstrates that selectively applying distillation only to incorrect responses allows the model to retain successful reasoning strategies while still

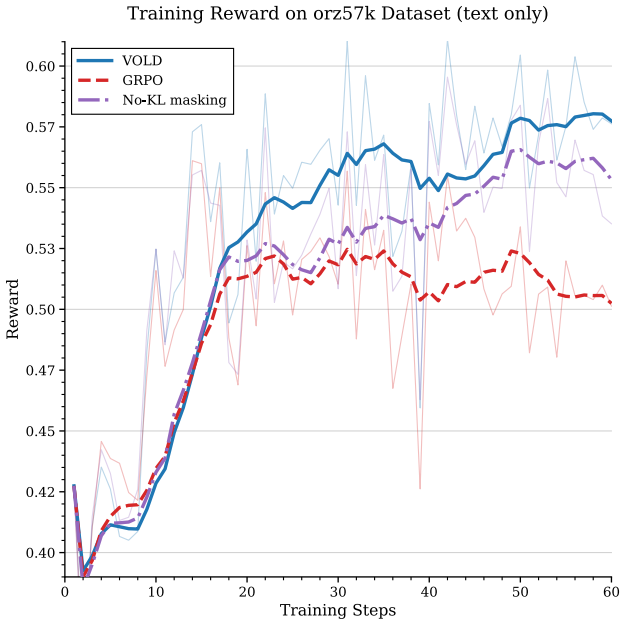

Figure 5. Training reward comparison: VOLD with KL masking (blue), without masking (purple), and vanilla GRPO (red). KL masking provides consistent performance gains throughout training.

Table 5. **Impact of Teacher Model Size**. We evaluate the final performance of VOLD when using teacher models of varying scales (4B, 8B, and 14B parameters). While increasing teacher size from 4B to 8B yields performance gains across most benchmarks, we observe diminishing returns with the 14B teacher, which provides no consistent improvement over the 8B model.

Dataset	Teacher 4B	Teacher 8B	Teacher 14B
MMMU-Pro	32.6	32.0	32.2
MMStar	53.7	55.2	55.1
Mathvision	26.1	28.0	27.8
MathVista	61.5	61.9	62.0
MathVerse	39.2	37.9	38.3
DynaMath$_{(Average)}$	49.0	50.7	50.9
WeMath	30.3	31.8	31.6
LogicVista	43.0	45.0	44.9

benefiting from teacher guidance on failed attempts. This validates our hypothesis that masking prevents interference between RL exploration of novel correct paths and teacher distillation objectives.

7.4. General Multimodal Capabilities

To evaluate whether training on text-only data degrades general multimodal capabilities, we extend our evaluation to additional benchmarks: MME [8] (split into Perception and Reasoning subscores), MMStar (split into Perception and Reasoning subcategories), and HallusionBench [10]. Table 6 presents the results for the Qwen2.5-VL baseline, VLAA-Thinker (trained on images), and VOLD (text-only training).

VOLD achieves the highest overall MME score (2268 vs 2157 baseline) and the strongest MMMU-Pro improvement (27.1→32.0). For MMStar, the overall drop is minimal (55.9→55.2, just 0.7%), but examining subcategories reveals a clear pattern: perception decreases (58.4→54.4) while reasoning improves (53.5→55.9). This trade-off is consistent across benchmarks (e.g., MME Reasoning: +24%). Notably, VLAA-Thinker, which is trained on images, shows the same pattern, indicating that this is not due to text-only training but an inherent effect of reasoning-focused methods. This perception-reasoning trade-off is well-documented in recent literature [2, 32]. Regarding hallucination, HallusionBench improves (49.8 vs 46.3), confirming that distillation from an LLM does not worsen hallucination.

7.5. VOLD as Foundation for Image-Based RL

We argue that when image data is available, VOLD provides a stronger starting point than training from scratch. To confirm this, we apply RL on VLAA-Thinker's image-text dataset starting from our VOLD checkpoint. Table 7 shows that VOLD + RL on images outperforms both VLAA-Thinker (which trains on images from scratch) and text-only VOLD, demonstrating that VOLD serves as a strong foundation that can be further enhanced with image data.

Table 6. **General Multimodal Capability Evaluation.** Perception slightly decreases while reasoning improves. This trade-off is shared by image-trained methods (VLAA-Th.), confirming it is inherent to reasoning-focused training, not text-only training.

Model	MME	MME Per.	MME Reas.	MMMU-Pro	MMStar (P / R)	HallBench
Qwen2.5-VL	2157	**1564**	594	27.1	**58.4** / 53.5	46.3
VLAA-Th.	2185	1552	634	24.6	53.2 / **58.6**	44.1
VOLD	**2268**	1530	**738**	**32.0**	54.4 / 55.9	**49.8**

Table 7. **VOLD as Foundation for Image-Based RL.** Starting from the VOLD checkpoint and applying RL on image-text data yields the best results, outperforming both text-only VOLD and image-only training from scratch.

Model	MathVista	MathVerse
Qwen2.5-VL (baseline)	61.2	31.2
VLAA-Thinker (RL on images)	61.0	36.4
VOLD (text-only)	61.9	37.9
VOLD + RL on images	**63.4**	**38.7**

7.6. Generalization to Other RL Algorithms

Since VOLD is orthogonal to the choice of RL algorithm, it can be integrated with any RL method. To verify this, we replace GRPO with GSPO [46] (Group Sequence Policy Optimization), which defines importance ratios at the sequence level rather than the token level. Figure 6 shows that on-policy distillation consistently improves training dynamics for both GRPO and GSPO, confirming that VOLD generalizes beyond a specific RL algorithm.

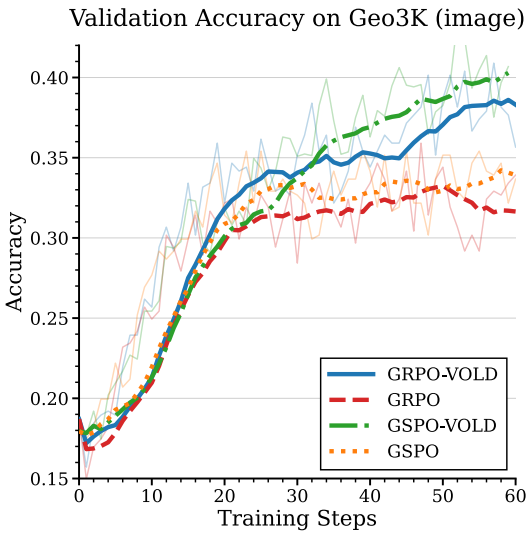

Figure 6. **Generalization to Other RL Algorithms.** Validation accuracy on Geo3K. On-policy distillation (OPD) improves both GRPO and GSPO, confirming VOLD is orthogonal to the choice of RL method.