A. Training Details

The key optimization and training hyperparameters for all experiments are summarized in Table 8.For all experiments, we fully fine-tune with freezing vision modules. Unless otherwise specified, we conduct training on a single node with 8 NVIDIA A100 GPUs.

For multimodal math reasoning and visual grounding, we prompt the model first to generate its reasoning within `<think>...</think>`, then produce a final answer in `<answer>...</answer>`, and finally output a verification score in `<score>...</score>`. Reward function consists of accuracy and formatting components. The accuracy component uses the widely adopted Python library math-verify to extract the model's answer and compare it with the ground truth. The format component ensures the correct structure and ordering of the `<think>` and `<answer>` tags, as well as the `<score>` tag.

For the math reasoning domain, we follow the open-r1-multimodal settings. We train on the curated multimodal math reasoning dataset, containing curated multimodal math problems with images and answers and evaluated on MathVista and MMMU. For the visual grounding tasks, we follow the VLM_R1 settings. We train on RefCOCO training set and evaluate on ReasonSeg.

For mobile agent tasks, we prompt the model dirctly to generate the tool call following the Qwen2.5-VL mobile use format in `<tool_call>...</tool_call>` and then output the verification score in `<score>...</score>`. Each sample contains a resized high-resolution screenshot, a natural-language goal together with a history of previous actions (`pre_act`), and a ground-truth tool call of the form `mobile_use(...)` specifying the action type (click, swipe, type, open, system_button, etc.) and its parameters (coordinates, text, button type, and so on).

The answer reward comprises both format and accuracy components. The format component ensures the correct structure and ordering of the `<tool_call>` and `<score>` tags. The accuracy component follows a step-wise verification process. First, we validate the action type against the ground truth, granting a $+1$ reward for a match. Subsequently, we evaluate the action parameters. For example, a click action earns an additional $+1$ reward if the Euclidean distance between the predicted coordinates and the ground truth label is less than $0.14 \times$ screen diagonal.

B. Ablation and Analysis

To better understand the effect of our preference verification reward in the mobile agent setting, we analyze training dynamics on the AndroidControl. As the policy improves, the binary verification reward quickly becomes dominated by correct trajectories, leading to a severe class imbalance where almost all reward-1 samples correspond to already-correct actions. As shown in Fig. 7, the subset of rollouts that receive binary reward $= 1$ rapidly collapses to near-perfect accuracy, providing little signal to separate moderate-quality actions from the very best ones. In contrast, the preference verification reward maintains a more informative mixture of correct and incorrect rollouts among its positive signals, preserving contrastive supervision even when overall task success is high.

Hyperparameter	Value
num_generations	8
per_device_train_batch_size	8
gradient_accumulation_steps	2
torch_dtype	bfloat16
data_seed	42
gradient_checkpointing	true
attn_implementation	flash_attention_2
learning_rate	1e-6
β	0.01

Table 8. Hyperparameter settings used in the training experiments.

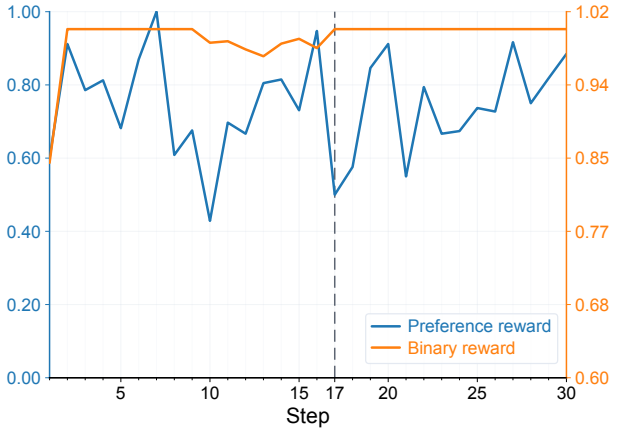

Figure 7. **Analysis of reward signal distribution during training.** The **Blue** line shows the proportion of correct answers among responses with preference verification reward = 1. The **Orange** line shows the proportion of correct answers among responses with binary verification reward = 1.

This difference in supervision is reflected in the learned verification scores. Figure 8 compares score distributions for models trained with binary versus preference verification reward on AndroidControl. The binary reward model tends to output highly discretized scores concentrated near the extremes, consistent with the imbalanced 0/1 supervision. In contrast, the preference reward model produces a smoother and more diverse score distribution, assigning fine-grained scores that better reflect relative action quality. This diversity is crucial for best-of-N selection, where ranking among multiple candidate trajectories matters more than predicting a single calibrated probability.

Tables 9 and 10 provide more fine-grained ablation results complementing the main figures. For MathVista and

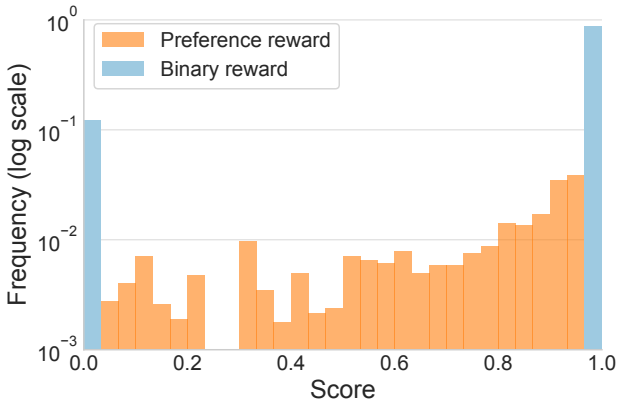

Figure 8. **Comparison of score distributions between models trained with Binary Reward and Preference Reward.** While the binary reward model tends to output discrete scores, the preference reward model produces a more diverse distribution.

AndroidControl, we report task performance using answer accuracy under pass@1 and best@8, while for ReasonSeg we use acc@(IoU > 0.5) under the same best-of-k protocol; all three domains share the same self-verification metrics AUC and AP. In the reward ablation, replacing the binary verification reward with our preference verification reward consistently strengthens self-verification, with especially large gains in AUC and AP on AndroidControl, indicating a much more reliable verifier under class imbalance. In the advantage ablation, decoupled advantage improves both pass@1 and best@8 performance across domains compared to entangled advantage, while also enhancing verification metrics, showing that separating generation and verification advantages benefits both task accuracy and ranking quality.

Table 9. **Ablation of Preference reward.** Replacing the binary answer reward with our *preference reward* consistently strengthens self-verification (↑AUC/AP) and improves best of N selection performance on *Math*, *Grounding*, and *GUI Agent*.

Domain	Method	Performance		Verification	
		pass@1	best@8	AUC	AP
MathVista	Binary verification reward	61.9	63.4	0.509	0.640
	Preference verification reward	**62.4**	**65.0**	**0.522**	**0.653**
ReasonSeg	Binary verification reward	**71.8**	72.6	0.636	0.789
	Preference verification reward	71.7	**73.5**	**0.672**	**0.804**
AndroidControl	Binary verification reward	70.6	70.6	0.609	0.765
	Preference verification reward	**70.9**	**72.7**	**0.727**	**0.841**

Table 10. **Ablation study on decoupled advantages.** Our advantage decoupled optimization consistently outperforms entangled advantage in both task performance and solution verification across mathematical reasoning, grounding, and GUI agent tasks.

Domain	Method	Performance		Verification	
		pass@1	best@8	AUC	AP
MathVista	Entangled Advantage	61.5	63.1	50.4	61.8
	Decoupled Advantage	**62.4**	**65.0**	**52.2**	**65.3**
ReasonSeg	Entangled Advantage	70.4	71.0	0.641	0.774
	Decoupled Advantage	**71.7**	**73.5**	**0.672**	**0.804**
AndroidControl	Entangled Advantage	71.0	70.4	0.542	0.715
	Decoupled Advantage	**70.9**	**72.7**	**0.727**	**0.841**

C. Extended Experimental Results

We conduct extensive experiments by enumerating all combinations of base, GRPO, and ADPO as both generators and verifiers across math reasoning, visual grounding, and mobile agent benchmarks (Tabs. 11 to 13). These results show that ADPO matches GRPO in pass@1 generation quality while providing substantially stronger verification for best-of-N selection, and that this unified training only incurs about 10% additional training time compared to GRPO with exactly the same data. In contrast to traditional pipelines that train a separate reward model with extra preference data, ADPO jointly learns generation and verification within a single policy, avoiding additional data collection and separate training runs and thus reducing both training and deployment cost.

Table 11 reports extended MathVista and MMMU results. With single-sample decoding (Sample 1), GRPO and ADPO generators reach similar pass@1 accuracy (62.2% vs 62.4% on MathVista; 48.7% vs 47.7% on MMMU), showing ADPO matches GRPO. Under best-of-12 decoding (Sample 12), pairing the ADPO generator and verifier increases MathVista accuracy from 62.5% to 65.3% and MMMU from 50.8% to 52.3%, indicating a stronger verifier under the same sampling budget.

Table 12 gives analogous ReasonSeg visual grounding results across short, long, and overall queries with gIoU, cIoU, and acc@(IoU > 0.5). On Sample 1, overall accuracy rises from 68.4% for the base generator to 71.1% for GRPO and 71.7% for ADPO. For best-of-12 (Sample 12), using ADPO for both generation and verification attains 73.2% overall, outperforming GRPO–GRPO (71.7%) and the base generator with majority voting (69.1%), confirming ADPO as the strongest box-ranking verifier.

For the majority-voting baseline on ReasonSeg, we aggregate continuous box predictions with an online clustering procedure. Specifically, predicted bounding boxes are processed sequentially and assigned to the existing cluster whose centroid has the highest IoU, if that IoU exceeds 0.5; otherwise, a new cluster is created. After each assignment, the cluster centroid is updated by averaging the coordinates of all boxes in that cluster. The final prediction is the centroid of the largest cluster.

Table 13 summarizes mobile agent results on AndroidControl and GUI Odyssey using step success rate (SR) and related metrics. On Sample 1, GRPO and ADPO generators achieve nearly identical SR (71.0% vs 70.9% on AndroidControl; 79.8% vs 79.7% on GUI Odyssey), indicating no pass@1 degradation. Under best-of-8 (Sample 8), the ADPO generator–verifier pair improves SR from 70.9% to 72.7% on AndroidControl and from 80.6% to 81.7% on GUI Odyssey, while the base generator with majority voting lags behind (58.3% and 46.6%), quantifying ADPO's stronger verification despite identical training data.

Table 11. **Evaluation results on multimodal math reasoning benchmarks.** Rows correspond to generators and columns correspond to verifiers. We use Qwen2-VL-7B as the base model, with GRPO and ADPO representing the finetuned models. Majority voting serves as the verifier baseline. Values are accuracy (%). GVQA: General VQA; MVQA: Math Target VQA; ARD: Art & Design; BUS: Business; HEM: Health & Medicine; HSS: Human & Social Science; SCI: Science; TEN: Technology & Engineering.

Generator	Verifier	MathVista (In-domain)			MMMU (OOD)						
		GVQA	MVQA	**ALL**	ARD	BUS	HEM	HSS	SCI	TEN	**ALL**
Sample 1											
Base	✗	68.9	48.5	57.9	67.5	39.1	49.3	69.0	33.9	36.7	47.1
GRPO	✗	69.8	55.7	62.2	65.0	45.9	48.2	68.2	35.9	39.8	48.7
ADPO	✗	68.7	57.0	62.4	63.1	46.2	50.2	71.1	33.3	35.3	47.7
Sample 4											
Base	Major	65.7	51.9	58.2	66.7	47.3	50.7	65.8	34.0	38.1	48.6
	Base	63.9	48.7	55.7	60.0	44.0	50.7	60.8	32.7	33.8	45.2
	GRPO	63.3	48.9	55.5	61.7	43.3	50.0	63.3	30.0	36.7	45.8
	ADPO	63.3	50.6	56.4	66.7	49.3	53.3	70.8	34.0	37.6	49.9
GRPO	Major	69.8	58.0	63.4	65.8	44.7	50.0	70.0	42.0	36.7	49.4
	Base	70.2	55.7	62.4	62.5	44.0	50.0	70.8	40.0	39.5	49.3
	GRPO	69.6	55.7	62.1	62.5	44.0	50.0	70.8	40.0	39.5	49.9
	ADPO	69.6	55.6	62.0	64.2	46.7	50.7	71.7	39.3	39.5	50.1
ADPO	Major	71.7	56.1	63.3	66.7	48.0	52.7	70.0	39.3	39.0	50.7
	Base	68.5	55.6	61.5	64.2	44.0	52.7	67.5	36.7	36.7	48.3
	GRPO	68.3	56.9	62.1	65.0	44.0	52.0	68.3	40.7	36.7	49.1
	ADPO	71.3	59.3	64.8	68.3	48.0	52.0	69.2	39.3	39.5	50.8
Sample 8											
Base	Major	68.0	53.3	60.1	68.3	50.0	53.3	68.3	32.7	36.7	49.4
	Base	63.0	51.9	57.0	65.0	40.7	48.0	65.0	32.0	32.4	45.0
	GRPO	62.8	50.9	56.4	65.8	40.0	49.3	65.8	32.7	37.1	46.6
	ADPO	63.5	50.6	56.5	67.5	48.7	54.0	71.7	36.0	41.0	51.2
GRPO	Major	70.4	56.5	62.9	66.7	48.7	51.3	74.2	42.7	36.7	51.1
	Base	67.6	55.0	60.7	62.5	47.3	51.3	72.5	38.7	37.6	49.7
	GRPO	67.6	55.0	60.8	62.5	46.7	50.0	70.0	39.3	38.6	49.3
	ADPO	67.6	54.4	60.5	63.3	46.7	53.3	68.3	38.7	39.0	49.8
ADPO	Major	71.1	58.0	64.0	65.0	49.3	56.7	71.7	38.7	40.5	51.8
	Base	70.0	55.7	62.3	63.3	52.0	53.3	66.7	42.7	41.0	51.6
	GRPO	69.8	55.9	62.3	63.3	52.7	54.7	65.8	42.0	39.0	51.2
	ADPO	72.2	58.9	65.0	65.8	54.0	54.7	66.7	40.7	41.0	52.1
Sample 12											
Base	Major	67.4	55.0	60.7	69.2	52.0	50.7	70.8	38.0	36.7	50.7
	Base	63.7	51.1	56.9	59.2	47.3	51.3	64.2	30.0	33.8	45.8
	GRPO	62.4	51.1	56.3	58.3	45.3	49.3	63.3	30.0	35.2	45.2
	ADPO	62.6	48.5	55.0	65.0	52.7	53.3	70.0	40.0	35.2	50.6
GRPO	Major	70.7	57.2	63.4	64.2	50.0	51.3	73.3	43.3	39.5	51.7
	Base	70.0	56.3	62.6	63.3	48.0	54.0	68.3	42.7	41.0	51.2
	GRPO	69.3	56.7	62.5	63.3	48.7	52.7	67.5	42.0	40.5	50.8
	ADPO	69.6	55.2	61.8	64.2	48.0	52.7	69.2	43.3	41.0	51.3
ADPO	Major	72.0	58.3	64.6	65.8	50.0	53.3	75.0	36.7	39.0	51.2
	Base	70.7	56.5	63.0	62.5	52.0	54.7	70.0	41.3	41.4	52.0
	GRPO	71.3	56.9	63.5	63.3	53.3	52.7	70.8	41.3	43.3	52.6
	ADPO	71.7	59.8	65.3	67.5	53.3	54.0	71.7	38.7	40.5	52.3

Table 12. **Evaluation results on image grounding benchmarks.** Rows correspond to generators and columns correspond to verifiers. We use Qwen2.5-VL-7B as the base model, with GRPO and ADPO representing the finetuned models. Majority voting serves as the verifier baseline. Models are trained on RefCOCO training set and tested on ReasonSeg (out-of-domain).

Generator	Verifier	Short query			Long query			Overall		
		gIoU	cIoU	ACC	gIoU	cIoU	ACC	gIoU	cIoU	ACC
Sample 1										
Base	✗	49.5	53.0	67.0	56.8	57.5	68.5	56.3	57.2	68.4
GRPO	✗	51.8	55.5	67.9	59.1	59.7	71.3	58.6	59.5	71.1
ADPO	✗	51.7	54.8	68.0	60.2	59.4	71.9	58.1	59.1	71.7
Sample 4										
Base	Major	47.8	52.0	66.0	57.3	57.9	69.3	56.7	57.5	69.1
	Base	47.2	51.4	66.0	56.9	57.5	68.6	56.3	57.1	68.4
	GRPO	49.6	53.2	67.0	57.4	57.9	68.7	56.9	57.7	68.6
	ADPO	50.4	53.5	68.0	57.3	57.9	69.1	56.9	57.7	69.1
GRPO	Major	54.5	57.0	68.0	58.8	59.5	72.1	58.5	59.4	71.8
	Base	55.2	57.5	68.0	59.7	60.3	72.9	59.4	60.2	72.6
	GRPO	53.4	56.4	68.9	59.1	59.7	72.0	58.8	59.5	71.8
	ADPO	55.1	57.8	68.0	60.5	61.1	73.4	60.2	60.9	73.1
ADPO	Major	52.7	55.3	67.0	60.1	60.5	72.0	59.6	60.2	71.7
	Base	51.2	54.1	66.0	59.3	59.9	71.7	58.8	59.6	71.4
	GRPO	51.0	54.3	67.0	60.0	60.7	72.7	59.4	60.3	72.4
	ADPO	52.2	55.1	67.0	61.0	61.5	73.3	60.5	61.1	72.9
Sample 8										
Base	Major	47.8	51.4	63.1	57.2	57.8	69.2	56.6	57.4	68.8
	Base	47.9	51.4	62.1	56.6	57.2	68.1	56.1	56.9	67.7
	GRPO	47.1	50.5	62.1	56.8	57.5	68.4	56.2	57.0	68.0
	ADPO	47.4	50.9	61.2	57.7	58.4	69.4	57.1	57.9	68.9
GRPO	Major	52.0	55.6	68.0	59.2	59.9	72.0	58.7	59.6	71.7
	Base	51.7	55.0	67.0	60.1	60.7	72.4	59.6	60.4	72.1
	GRPO	51.5	55.3	68.0	60.0	60.7	72.4	59.5	60.4	72.1
	ADPO	54.4	57.2	67.0	60.6	61.3	73.8	60.2	61.1	73.3
ADPO	Major	53.2	56.1	67.0	58.8	59.4	71.3	58.5	59.2	71.0
	Base	52.9	55.4	67.0	59.6	60.2	72.0	59.2	59.9	71.7
	GRPO	55.6	57.8	68.9	59.9	60.6	72.9	59.6	60.5	72.7
	ADPO	53.2	56.0	67.0	60.9	61.5	73.7	60.4	61.2	73.5
Sample 12										
Base	Major	50.2	53.7	66.0	57.2	57.8	69.3	56.8	57.6	69.1
	Base	49.5	53.3	68.0	56.9	57.6	68.9	56.5	57.4	68.8
	GRPO	50.0	53.1	66.0	57.1	57.9	69.1	56.7	57.6	68.9
	ADPO	49.8	52.6	65.1	57.6	58.2	69.2	57.1	57.8	68.9
GRPO	Major	55.6	58.1	69.9	58.8	59.5	72.2	58.6	59.4	72.0
	Base	48.8	52.4	65.1	59.6	60.2	72.5	58.9	59.7	72.1
	GRPO	53.2	55.8	68.0	59.1	59.9	71.9	58.8	59.6	71.7
	ADPO	54.1	56.7	68.9	60.9	61.6	74.0	60.5	61.3	73.7
ADPO	Major	53.3	55.3	66.0	59.3	60.0	71.8	58.9	59.8	71.5
	Base	55.0	57.4	68.9	60.2	60.9	72.1	59.9	60.7	71.9
	GRPO	53.8	56.5	68.9	60.3	61.0	72.4	59.9	60.7	72.1
	ADPO	53.9	56.2	67.0	61.3	62.0	73.6	60.9	61.6	73.2

Table 13. **Evaluation results on mobile agent benchmarks.** Rows correspond to generators and columns correspond to verifiers. We use Qwen2.5-VL-7B as the base model, with GRPO and ADPO representing the finetuned models.

Generator	Verifier	AndroidControl			GUI Odyssey		
		Type	Grounding	**SR**	Type	Grounding	**SR**
Sample 1							
Base	✗	82.2	73.6	61.3	81.1	61.4	52.8
GRPO	✗	86.0	76.9	71.0	93.1	83.9	79.8
ADPO	✗	85.8	76.2	70.9	94.2	82.5	79.7
Sample 4							
Base	Major	76.3	68.1	56.0	76.9	55.3	46.5
	Base	72.1	67.8	52.5	75.3	55.6	45.2
	GRPO	74.9	71.4	57.7	75.3	55.2	45.3
	ADPO	76.4	74.5	60.7	75.1	55.7	45.6
GRPO	Major	85.5	77.2	71.0	94.7	83.9	81.3
	Base	85.4	77.2	71.0	94.3	83.8	81.0
	GRPO	85.4	77.3	70.8	94.4	83.7	80.7
	ADPO	85.6	77.7	71.2	94.5	84.0	81.4
ADPO	Major	86.6	77.1	71.6	93.9	83.5	79.8
	Base	86.4	76.4	71.0	94.7	84.2	81.2
	GRPO	86.4	77.9	72.0	94.7	84.2	81.1
	ADPO	86.3	79.5	72.7	94.7	84.5	81.6
Sample 8							
Base	Major	78.7	68.8	58.3	76.7	55.4	46.6
	Base	73.9	68.4	54.3	75.1	55.2	44.9
	GRPO	77.3	73.4	61.0	74.4	54.3	44.5
	ADPO	79.7	76.5	64.7	73.9	54.6	44.6
GRPO	Major	85.6	76.9	70.8	94.6	84.4	81.5
	Base	85.6	77.1	71.0	93.7	84.4	80.7
	GRPO	85.4	77.4	70.9	93.7	84.4	80.6
	ADPO	85.6	77.7	71.4	93.9	84.8	81.2
ADPO	Major	86.5	76.4	71.3	94.8	84.0	80.9
	Base	86.1	76.2	70.8	95.1	84.6	81.6
	GRPO	85.8	77.7	71.4	94.9	84.4	81.4
	ADPO	86.4	78.7	72.7	94.8	84.7	81.7
Sample 12							
Base	Major	78.9	68.7	58.3	76.9	55.5	46.9
	Base	73.4	67.9	53.6	74.5	55.5	44.6
	GRPO	76.8	73.2	60.7	73.5	54.3	44.0
	ADPO	79.2	76.7	64.5	72.9	53.8	43.6
GRPO	Major	85.6	77.4	71.1	94.5	84.0	81.1
	Base	85.6	78.0	71.4	93.0	84.0	79.9
	GRPO	85.4	77.5	70.9	93.1	83.9	79.7
	ADPO	85.7	77.9	71.5	93.2	84.2	80.3
ADPO	Major	86.6	76.7	71.9	94.4	83.7	80.5
	Base	86.5	76.3	71.6	94.8	84.6	81.5
	GRPO	85.4	78.6	71.9	94.6	84.1	81.1
	ADPO	86.3	78.9	72.9	94.4	84.5	81.4