DUAL TUNING: JOINT SCREENING OF DATA UTILITY AND SUPERVISION MODES FOR MULTIMODAL FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal fine-tuning requires decisions about both the training data and the form of supervision, such as chain of thought (CoT) or direct-answer (DA) supervision. These decisions are jointly shaped by the capability of the base model, the quality of CoT traces, and the characteristics of the target task, rather than by any fixed recipe. Consequently, isolated ablations are insufficient for joint assessment. Here we introduce Dual Tuning, a screening protocol conditioned on the training configuration that addresses a complementary question: for a given base model and candidate dataset, which task groups benefit from training, and which supervision mode is better supported by the observed gains? It uses one joint supervised fine-tuning run per configuration to obtain gains over the base model and differences between CoT and DA modes for each task group. Experiments on spatial, mathematical, and multidisciplinary tasks reveal heterogeneous patterns across subgroups and models. Qwen2.5-VL-7B favors DA under the evaluated spatial configuration and CoT on MathVista. However, a CoT advantage over DA does not always coincide with a genuine improvement over the base model. The results on the Massive Multi-discipline Multimodal Understanding and Reasoning (MMMU) benchmark identify 9 subjects favoring CoT, 4 favoring direct answers, and 10 with negative gains in both modes. Finally, subset retraining experiments provide directional support for using these signals in data curation. Dual Tuning offers a common empirical basis for screening candidate data and assigning supervision modes before subsequent training.

1 INTRODUCTION

Multimodal supervised fine-tuning (SFT) requires deciding both which training data to retain and which form of supervision to use. The same visual question can be paired with chain of thought (CoT) or direct answer (DA) supervision. CoT prompting and its multimodal extensions demonstrate the value of intermediate reasoning for language and vision language tasks (Wei et al., 2022; Zhang et al., 2023). For data curation, this motivates two coupled questions. Does training with the candidate data improve the base model, and which supervision mode is better supported for each task group?

These decisions depend on the base model, the construction of CoT traces, and their alignment with task characteristics rather than fixed task labels. A candidate dataset can contain task groups with different supervision preferences, even when the aggregate score favors one mode. The relevant unit of assessment is therefore the current training configuration, including the base model, paired training data, training recipe, and task labels linking data subsets to evaluation groups.

Evaluating a data subset in isolation does not directly establish how its associated task group behaves after training on a mixed dataset. Comparing supervision modes alone likewise does not establish whether either improves over the base model. Repeating separate experiments across data groups and supervision choices also expands the set of training comparisons. We introduce Dual Tuning, a screening protocol conditioned on the training configuration that jointly assesses data utility and supervision modes across multiple task groups.

Dual Tuning uses one joint SFT run per configuration to train on paired CoT and DA examples under their respective prompts. It then evaluates the base model and the jointly trained model in both

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

modes for each labeled task group. The resulting four scores measure gains over the stronger base model mode and differences between the trained modes. Combining these signals distinguishes a supervision preference accompanied by positive training gains from a relative advantage without improvement over the base model, providing a common basis for screening candidate data and assigning supervision modes before subsequent training.

Experiments on spatial, mathematical, and multidisciplinary tasks illustrate the value of this distinction. For Qwen2.5-VL-7B, CoT yields positive aggregate gains over the base model on the evaluated spatial benchmarks, while DA achieves larger gains, supporting DA supervision under this configuration (Table 3). In MathVista's Logical Reasoning category, CoT outperforms DA, yet neither mode improves over the base model (Table 4). On the Massive Multi-discipline Multimodal Understanding and Reasoning (MMMU) benchmark, the same model yields 9 subjects favoring CoT, 4 favoring direct answers, and 10 with negative gains in both modes (Table 5). These results show why curation requires both gains over the base model and comparisons between modes at task group granularity. Comparisons with models trained in a single mode assess the joint model as a screening instrument, while subset retraining provides directional support for using its signals in subsequent data selection.

Our contributions are threefold.

1. **A joint screening protocol.** Dual Tuning uses one joint SFT run per candidate configuration to obtain gains and mode comparisons across multiple labeled task groups.
2. **A joint criterion for data and supervision choices.** Combining gains over the base model with differences between CoT and DA modes separates positive training benefits from relative mode advantages.
3. **An empirical study of configuration dependent curation.** Experiments across tasks, base models, and CoT datasets characterize the screening outcomes, with subset retraining providing directional support for their use in data selection.

2 Related Work

Chain of thought (CoT) prompting uses intermediate steps to improve language-model reasoning (Wei et al., 2022), while multimodal extensions incorporate visual evidence into rationale generation and answer inference (Zhang et al., 2023). Earlier releases offered separate instruction-following and reasoning-specialized models, exemplified by Qwen2.5-Instruct (Qwen et al., 2024) and QwQ-32B (Qwen Team, 2025), respectively. Qwen3 subsequently integrates thinking and non-thinking modes in one model with inference-time thinking-budget control (Yang et al., 2025a). Data-centric approaches instead optimize training-data composition: gradient-based influence estimation selects instruction examples for target capabilities (Xia et al., 2024), whereas proxy-model-based reweighting adjusts domain proportions for language-model pretraining (Xie et al., 2023). Dual Tuning complements these approaches by jointly screening data utility and the choice between CoT and direct-answer supervision for a given multimodal training configuration.

3 Method

Given a target multimodal task $\mathcal{T}$, a base model $\mathcal{M}$, and candidate data $\mathcal{D}$, Dual Tuning screens data utility and supervision modes under a fixed training configuration. Subtask labels $c_k \in \mathcal{C}$ link training subsets to evaluation groups. In our experiments, these labels are generated uniformly through an external model application programming interface (API). The decisions concern this model, paired data, prompts, and training recipe.

3.1 Paired Dual-Mode Fine-Tuning

For each visual question and answer, we retain a chain of thought (CoT) target and a direct answer (DA) target with the same visual input, question, and final answer. We jointly fine-tune on the two formats under their respective system prompts to obtain $\mathcal{M}_{\mathrm{DT}}$. Let L denote the thinking prompt and S the direct-answer prompt. Evaluating both $\mathcal{M}$ and $\mathcal{M}_{\mathrm{DT}}$ on the same evaluation groups gives

Algorithm 1 Dual Tuning for Configuration-Conditional Screening

Input: Base model $\mathcal{M}$; paired data $\mathcal{D}_{\mathrm{CoT}}, \mathcal{D}_{\mathrm{DA}}$; fixed supervised fine-tuning (SFT) recipe; group labels and training–evaluation mapping; evaluation groups $\{\mathcal{B}_c\}_{c\in\mathcal{C}}$
Output: Group-level gains, mode gaps, and screening decisions
1: Attach thinking / direct-answer prompts to the paired formats
2: $\mathcal{M}_{\mathrm{DT}} \leftarrow \mathrm{SFT}(\mathcal{M}, \mathcal{D}_{\mathrm{CoT}} \cup \mathcal{D}_{\mathrm{DA}})$
3: **for** each evaluation group $c \in \mathcal{C}$ **do**
4: $(B_L(c), B_S(c)) \leftarrow \mathrm{Eval}(\mathcal{M}, \mathcal{B}_c, L/S)$
5: $(DT_L(c), DT_S(c)) \leftarrow \mathrm{Eval}(\mathcal{M}_{\mathrm{DT}}, \mathcal{B}_c, L/S)$
6: Compute $\mathrm{Gain}_{\mathrm{CoT}}(c)$, $\mathrm{Gain}_{\mathrm{DA}}(c)$, and $\mathrm{GAP}_{\mathrm{DT}}(c)$
7: Assign a screening category using Section 3.2
8: **end for**

B_L, B_S, DT_L, DT_S. These are prompt-conditioned scores; in particular, DT_L and DT_S are two evaluations of the same jointly trained model.

The thinking prompt is given below.

```
When the user asks a question, your response must include two parts:
first, the reasoning process enclosed in <think>...</think> tags, then
 the final answer enclosed in <answer>...</answer> tags.
```

The direct answer prompt is given below.

```
Please directly answer the following question.
```

3.2 METRICS

For each subtask, we compare chain of thought (CoT) and direct answer (DA) evaluation using the following metrics. Mode gaps are score differences. Gains over the base model are relative percentages.

$$\mathbf{GAP_B} = \mathbf{B_L} - \mathbf{B_S} \tag{1}$$

measures the base model's score difference between CoT and DA evaluation.

$$\mathbf{GAP_{DT}} = \mathbf{DT_L} - \mathbf{DT_S} \tag{2}$$

measures the same mode difference after joint training.

$$\mathbf{Gain_{CoT}} = \frac{\mathbf{DT_L} - \max(\mathbf{B_L}, \mathbf{B_S})}{\max(\mathbf{B_L}, \mathbf{B_S})} \times 100 \tag{3}$$

measures the relative percentage gain of CoT evaluation after joint training over the stronger base mode.

$$\mathbf{Gain_{DA}} = \frac{\mathbf{DT_S} - \max(\mathbf{B_L}, \mathbf{B_S})}{\max(\mathbf{B_L}, \mathbf{B_S})} \times 100 \tag{4}$$

measures the corresponding relative percentage gain of DA evaluation against the same reference. Both relative gains require a positive reference score.

$$\mathbf{Gain_{token}} = \frac{\mathbf{Gain_{CoT}}}{N_{token}} \times 100 \tag{5}$$

provides an auxiliary length normalized gain, where N_{token} is the mean CoT output length in tokens. It is not used for screening.

Screening combines gain over the base model with the trained mode gap. A group favors CoT when both $\mathbf{Gain_{CoT}}$ and $\mathbf{GAP_{DT}}$ are positive, and DA when $\mathbf{Gain_{DA}}$ is positive and $\mathbf{GAP_{DT}}$ is negative. Data associated with groups having negative gains in both modes are candidates for review or exclusion. All remaining cases, including ties and zero gains, receive no strict mode recommendation.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Tasks and candidate data. We study paired chain of thought (CoT) and direct answer (DA) supervision in two data configurations. Spatial understanding is evaluated on VSI-Bench and CV-Bench (Yang et al., 2025b; Tong et al., 2024), covering eight video and four image categories. The spatial data are constructed from open 3D datasets (Roberts et al., 2021; Dai et al., 2017; Yeshwanth et al., 2023; Baruch et al., 2021; Ray et al., 2024), with CoT traces generated using Gemini 2.5 Pro (Comanici et al., 2025). This configuration contains 30k image and 8k video question answering pairs. The other configuration uses 20k selected OneThinker image pairs (Feng et al., 2025) and is evaluated on MathVista (Lu et al., 2023) and the Massive Multi-discipline Multimodal Understanding and Reasoning (MMMU) benchmark (Yue et al., 2024). We exclude weakly related categories such as tracking and segmentation, and remove traces containing `wait` more than twice. In both configurations, each CoT and DA pair shares the visual input, question, and final answer.

Models and training. Qwen2.5-VL-7B (Bai et al., 2025b) is the primary base model. Ming-lite-omni v1.5, a mixture of experts model with 20B total and 3B active parameters (InclusionAI, 2025), provides an additional model configuration. Each configuration uses one joint supervised fine-tuning (SFT) run on the paired formats. Training uses 32 NVIDIA H800 accelerators, AdamW, four epochs, and a learning rate of 1×10^{-5}. The batch size per accelerator is four for images and one for videos, with at most 256 video frames.

Evaluation and interpretation. The base and jointly trained models are evaluated with the two prompts in Method. We report benchmark scores and results for 12 spatial categories, 7 mathematical categories, and 30 multidisciplinary subjects. We uniformly assign group labels to candidate training samples through an external model application programming interface (API). These labels link candidate data to evaluation groups. Following Section 3.2, both trained modes are compared with the stronger base mode. Gains are relative percentages, whereas mode gaps are score differences. The reported categories describe the evaluated configuration and use the strict gain and gap criteria, including cases without a mode recommendation. Appendix B gives the comparison protocol and scope.

4.2 COMPARISON WITH SEPARATE TRAINING

We first examine whether the joint model yields mode comparisons consistent with the separately trained models in our study. Each separate model is trained with chain of thought (CoT) or direct answer (DA) supervision and evaluated in its corresponding mode. Across the eight comparisons in Table 1, the absolute score difference between separate and joint training is at most 1.00 point. The mode ordering is also consistent. DA scores higher on both spatial benchmarks, whereas CoT scores higher on the mathematical and multidisciplinary benchmarks.

The comparison retains the configurations actually evaluated. Spatial joint training combines image and video data, while each separate spatial baseline uses one modality. For the selected OneThinker data, both modes use the same candidate data source. Thus, the table supplies a direct reference for the observed mode preferences in these configurations. It does not require selecting the best score across mismatched prompts.

The complete results in Appendix B also show why the requested output format matters. CoT training can produce verbose outputs under a DA prompt, and DA training can produce direct answers under a thinking prompt. Joint training supplies examples for both requested formats. Its $\mathbf{GAP_{DT}}$ therefore compares two prompted evaluations of the same trained model, providing the common comparison used in the subsequent group analysis. The value of this reference is apparent on MathVista. Separate CoT and DA training score 73.40 and 67.43 in their corresponding modes, while the joint model scores 72.87 and 67.37. The preference for CoT is preserved, but the group analysis below additionally asks whether that preference corresponds to improvement over the base model.

Table 1: Mode matched comparisons for Qwen2.5-VL-7B using chain of thought (CoT) and direct answer (DA) evaluation. MMMU denotes Massive Multi-discipline Multimodal Understanding and Reasoning. Differences are joint minus separate scores in score points. Spatial joint training mixes image and video data, whereas each separate spatial baseline uses one modality. Full configurations and cross prompt scores are in Appendix B.

Benchmark	Mode	Separate	Joint	Difference
VSI-Bench	DA	50.98 (E3)	51.35 (E5)	+0.37
	CoT	44.14 (E4)	43.26 (E5)	−0.88
CV-Bench	DA	84.15 (E1)	84.00 (E5)	−0.15
	CoT	80.55 (E2)	80.25 (E5)	−0.30
MathVista	DA	67.43 (E6)	67.37 (E8)	−0.06
	CoT	73.40 (E7)	72.87 (E8)	−0.53
MMMU	DA	52.00 (E6)	51.00 (E8)	−1.00
	CoT	53.33 (E7)	54.11 (E8)	+0.78

Table 2: Screening outcomes from the reported group scores using Section 3.2. CoT and DA denote chain of thought and direct answer. MMMU denotes Massive Multi-discipline Multimodal Understanding and Reasoning. Spatial combines VSI-Bench and CV-Bench. Counts refer to evaluation groups, excluding aggregate rows, rather than training samples. Both negative means negative gains in both modes. Boundary covers the remaining cases without a strict recommendation. Assignments describe the displayed point estimates. Full scores appear here and in Appendix A.

Evaluation groups	CoT favored	DA favored	Both negative	Boundary	Total
Qwen2.5-VL-7B					
Spatial	0	11	0	1	12
MathVista	4	1	1	1	7
MMMU	9	4	10	7	30
Ming-lite-omni v1.5					
Spatial	0	10	2	0	12
MathVista	3	0	4	0	7
MMMU	8	3	10	9	30

4.3 Screening Outcomes across Task Groups

Within benchmark diversity. On the Massive Multi-discipline Multimodal Understanding and Reasoning (MMMU) benchmark, Qwen2.5-VL-7B obtains an aggregate chain of thought (CoT) score of 54.11 and a direct answer (DA) score of 51.00. The positive mode gap of 3.11 points does not imply a single curation decision for all subjects. Table 2 identifies 9 subjects favoring CoT, 4 favoring DA, 10 with negative gains in both modes, and 7 without a strict recommendation.

Figure 1 locates these decisions using the two gains. Physics improves from a best base score of 40.00 to 53.33 under CoT, while Geography favors DA at 50.00 against 30.00 under CoT and a best base score of 43.33. Accounting has a positive mode gap of 16.66 points, but neither trained mode exceeds the best base score of 66.67. It therefore belongs to the group with two negative gains rather than the CoT recommendation group. The boundary cases carry distinct information as well. Art improves equally in both modes, whereas Finance reaches the base reference under CoT without exceeding it. Neither receives a strict mode recommendation. The full 30 subject scores are in Table 5.

These outcomes translate the same benchmark evaluation into distinct candidate actions. The 13 subjects favoring one mode nominate data and a supervision format together. The 10 subjects with two negative gains nominate associated data for review under the evaluated mixture. The remaining seven retain information about ties or gains that do not cross the base reference, rather than being forced into a preferred mode. This separation preserves distinctions hidden by the aggregate CoT advantage and organizes the signals examined in subsequent curation.

Spatial gains and the DA alternative. The complete spatial results in Table 3 show why a positive CoT gain must be considered alongside the DA alternative. Using the VSI-Bench base scores of

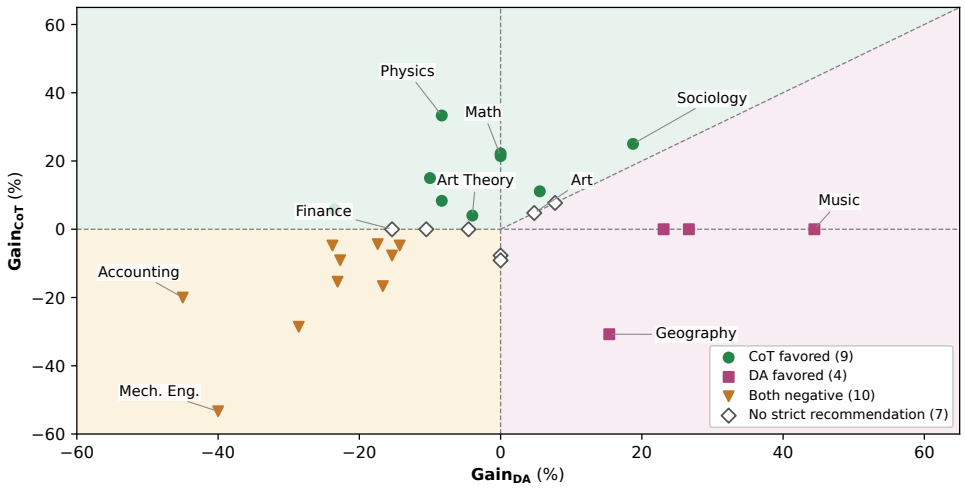

Figure 1: Subject screening on Massive Multi-discipline Multimodal Understanding and Reasoning (MMMU) for Qwen2.5-VL-7B. Axes show relative gains for direct answer (DA) and chain of thought (CoT) evaluation. All 30 subjects are plotted from Table 5. Colors and markers encode the strict screening decisions. Open markers denote cases without a strict recommendation. Selected subjects are labeled for readability.

Table 3: Spatial screening for Qwen2.5-VL-7B with chain of thought (CoT) and direct answer (DA) evaluation. B_S, B_L are base scores and DT_S, DT_L are joint model scores. Gains are relative percentages and gaps are score differences. Eleven groups favor DA. Depth has equal positive gains.

Task	$\mathbf{B_S}$	$\mathbf{B_L}$	$\mathbf{GAP_B}$	$\mathbf{DT_S}$	$\mathbf{DT_L}$	$\mathbf{Gain_{DA}}$	$\mathbf{Gain_{CoT}}$	$\mathbf{GAP_{DT}}$
VSI-Bench	37.57	24.91	-12.66	51.35	43.26	36.68	15.15	-8.09
Obj. Count	40.41	12.19	-28.22	49.65	43.47	22.87	7.57	-6.18
Abs. Dist.	22.88	1.80	-21.08	40.17	29.68	75.57	29.72	-10.49
Obj. Size	50.18	12.34	-37.84	64.30	50.97	28.14	1.57	-13.33
Room Size	38.96	22.71	-16.25	56.98	38.72	46.25	-0.62	-18.26
Rel. Dist.	39.44	40.14	0.70	58.31	48.03	45.27	19.66	-10.28
Rel. Dir.	39.66	39.57	-0.09	44.44	43.82	12.05	10.49	-0.62
Route Plan	32.47	32.47	0.00	34.02	32.47	4.77	0.00	-1.55
Appr. Order	36.57	38.03	1.46	62.94	58.90	65.50	54.88	-4.04
CV-Bench	75.17	73.81	-1.36	84.00	80.25	11.75	6.76	-3.75
Count	64.72	62.94	-1.78	70.94	61.80	9.61	-4.51	-9.14
Relation	87.69	86.46	-1.23	95.69	92.92	9.12	5.96	-2.77
Depth	70.17	72.00	1.83	87.00	87.00	20.83	20.83	0.00
Distance	80.33	76.17	-4.16	85.50	84.00	6.44	4.57	-1.50

37.57 and 24.91, joint training reaches 51.35 under DA and 43.26 under CoT. The respective relative gains are 36.68% and 15.15%. CV-Bench likewise improves in both modes, with gains of 11.75% for DA and 6.76% for CoT. Thus, the spatial result is not an absence of CoT improvement. DA yields the larger improvement under this configuration.

At group granularity, 11 of the 12 categories favor DA, and CV-Bench Depth has equal positive gains. Appearance Order illustrates substantial gains in both modes, but still favors DA at 62.94 compared with 58.90 for CoT. Room Size instead has a positive DA gain and a negative CoT gain. These cases lead to the same mode recommendation through different gain patterns. They support DA for the current spatial candidate data without assigning an intrinsic reasoning label to spatial tasks.

Mode preference and improvement over the base model. MathVista presents the complementary distinction (Table 4). CoT reaches 72.87 against 67.37 for DA, exceeding the stronger base score of 69.83. Geometry, Arithmetic, Statistical, and Algebraic Reasoning satisfy both CoT criteria.

Table 4: MathVista screening for Qwen2.5-VL-7B with chain of thought (CoT) and direct answer (DA) evaluation. Gains are relative percentages and gaps are score differences. Logical Reasoning illustrates a positive mode gap with two negative gains. Scientific Reasoning has zero gains in both modes.

Task	$\mathbf{B_S}$	$\mathbf{B_L}$	$\mathbf{GAP_B}$	$\mathbf{DT_S}$	$\mathbf{DT_L}$	$\mathbf{Gain_{DA}}$	$\mathbf{Gain_{CoT}}$	$\mathbf{GAP_{DT}}$
MathVista	69.80	69.83	0.03	67.37	72.87	-3.52	4.35	5.50
Comm.	44.44	40.51	-3.94	49.07	46.06	10.42	3.65	-3.01
Scientific	74.70	74.40	-0.30	74.70	74.70	0.00	0.00	0.00
Geometry	70.55	70.26	-0.29	64.08	74.57	-9.17	5.70	10.49
Arithmetic	78.33	80.45	2.12	71.52	82.12	-11.10	2.08	10.60
Logical	23.42	28.83	5.41	18.92	23.42	-34.37	-18.77	4.50
Statistical	84.64	84.12	-0.52	85.86	87.43	1.44	3.30	1.57
Algebraic	68.75	70.83	2.08	66.15	77.08	-6.61	8.82	10.93

Numeric Commonsense favors DA, while Scientific Reasoning has zero gains in both modes. Logical Reasoning is particularly informative. Its positive $\mathbf{GAP_{DT}} = 4.50$ points coexists with $\mathbf{Gain_{CoT}} = -18.77\%$ and a negative DA gain. A mode comparison alone would prefer CoT, whereas the joint screen identifies the associated data for review under the current configuration. Together, the spatial and mathematical results establish the need for both a base reference and an alternative supervision mode.

4.4 Subset Retraining for Data Curation

We next examine whether the screen helps nominate training subsets. The existing experiments use three partitions. The first two separate groups according to their CoT or DA screening gains and retrain the associated data with the corresponding supervision. The third separates groups with negative gains in both modes from the remaining groups and applies joint supervision. Evaluation covers the groups associated with each selected subset.

Figure 2 presents the CoT comparison. The coordinates retain the initial DA and CoT gains. Marker shapes indicate the original subset assignments, while green and red indicate positive and negative gains after retraining. The comparison therefore tests agreement in gain direction, rather than plotting the magnitude of the retraining gain on an axis. Positive CoT regions contain predominantly green markers and negative regions predominantly red markers. Physics remains positive, for example, while Accounting remains negative under CoT subset training. The positive screening gain for Materials is followed by a negative retraining gain, showing why the result supports candidate selection rather than a deterministic prediction for every group.

The DA comparison in Figure 3 tests the signal under direct supervision. The left and right halves distinguish negative and positive initial DA gains, while colors report the gain direction after DA subset retraining. The broad correspondence extends beyond CoT supervision. Figure 4 colors each group by its larger gain after joint retraining. Most groups with two negative screening gains remain negative, while Accounting becomes positive. Accounting thus illustrates how a proposed exclusion can change with the supervision used in subsequent training. Together, the three comparisons support prioritizing candidate subsets and supervision modes. They evaluate selected groups rather than a final reassembled mixture. Appendix D details the historical subset assignments, including zero gain groups.

4.5 Dependence on the Training Configuration

Base model. Ming-lite-omni v1.5 changes several screening outcomes (Table 2). On MathVista, chain of thought (CoT) reaches 78.77 compared with 70.80 for direct answer (DA), but both scores are below the stronger base mode at 79.00. Geometry, Logical, and Statistical Reasoning favor CoT, while the remaining four categories have negative gains in both modes. On the Massive Multi-discipline Multimodal Understanding and Reasoning (MMMU) benchmark, Accounting changes from two negative gains for Qwen2.5-VL-7B to a CoT recommendation for Ming-lite-omni, whose CoT score of 73.33 exceeds its best base score of 66.67. The spatial mode gaps are predominantly

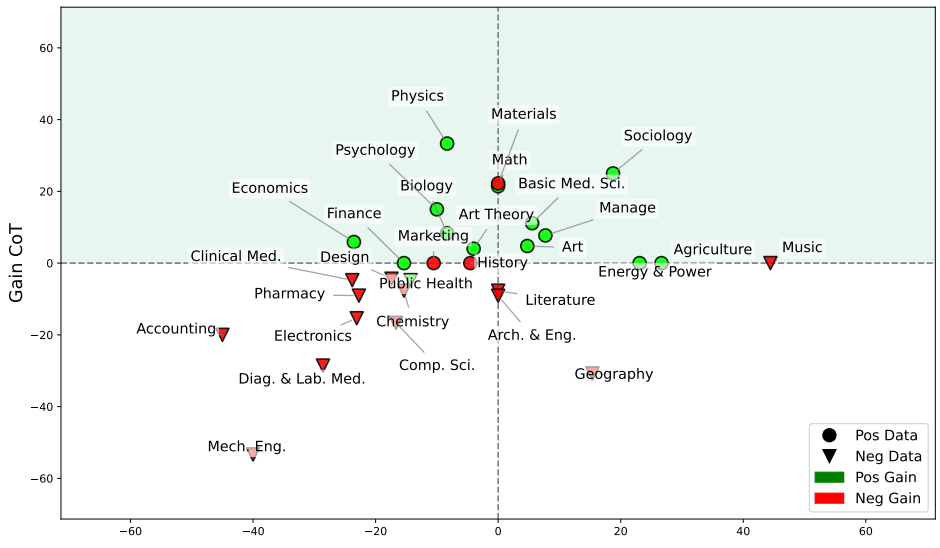

Figure 2: Subset retraining with chain of thought (CoT) supervision. Horizontal and vertical coordinates are the initial direct answer (DA) and CoT gains, respectively. Shapes retain the original subset assignments. Green and red indicate positive and negative CoT gains after retraining, not their magnitudes. Evaluation covers the groups associated with each subset.

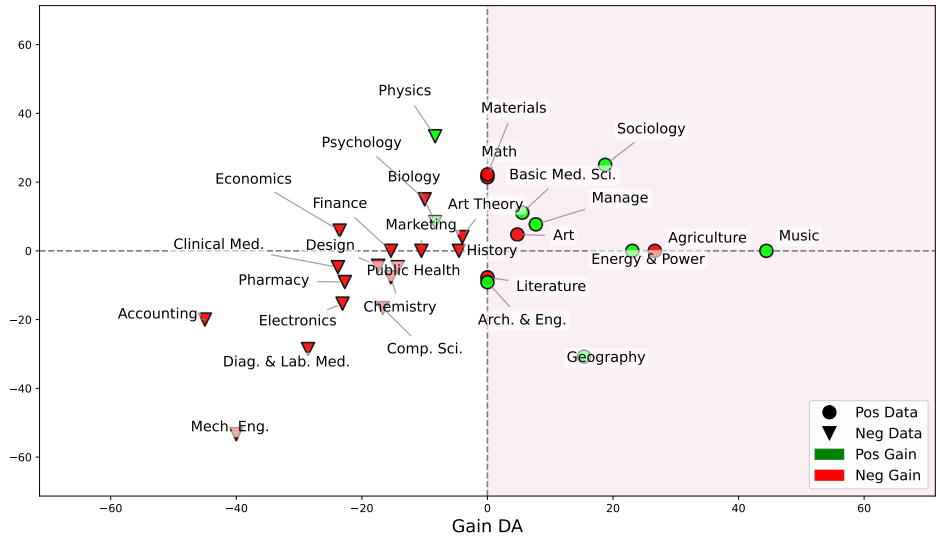

Figure 3: Subset retraining with direct answer (DA) supervision. Coordinates retain the initial DA and chain of thought (CoT) gains. Shapes show the original subset assignments, and colors show the sign of the DA gain after retraining. Evaluation covers the selected groups.

negative for both models, yet Ming-lite-omni also has two groups with negative gains in both modes. These distinctions show why a shared mode tendency does not imply identical data selection. Full Ming-lite-omni scores are in Appendix A.

The mathematical groups make the model dependence more concrete. Logical Reasoning has two negative gains for Qwen2.5-VL-7B, but its CoT score for Ming-lite-omni rises from a best base score of 27.03 to 29.73. Geometry favors CoT for both models, although the relative CoT gain is 5.70% for Qwen2.5-VL-7B and 0.32% for Ming-lite-omni. Thus, agreement on a supervision mode can coexist with different gains over the base model, while other groups change screening category

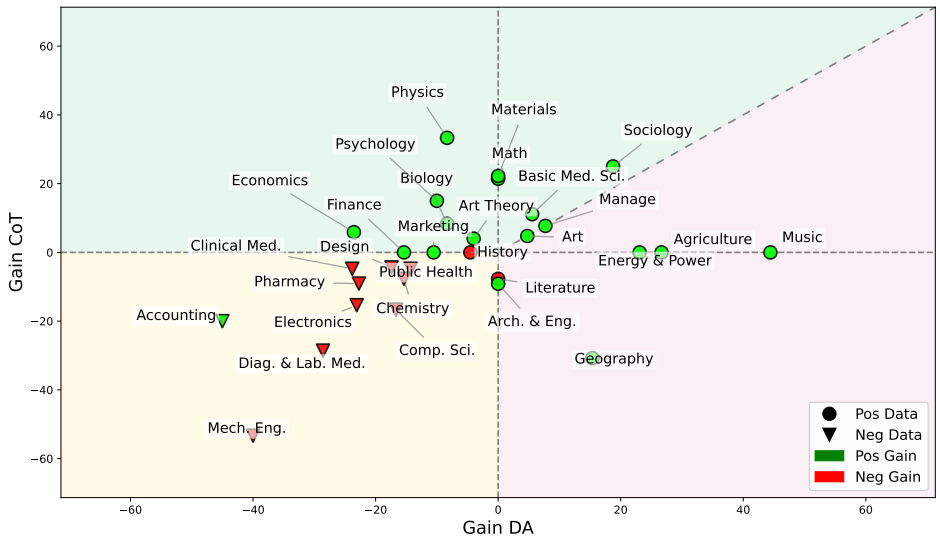

Figure 4: Joint subset retraining with chain of thought (CoT) and direct answer (DA) supervision. Coordinates are the initial screening gains. Triangles identify the subset with two negative gains and circles the complementary subset. Colors show the sign of the larger CoT or DA gain after retraining. The complementary subset includes boundary cases.

altogether. The group view distinguishes these cases from the shared aggregate preference for CoT on MathVista.

Candidate reasoning data. An additional configuration uses CoT data constructed with Qwen3-VL-235B (Bai et al., 2025a). The MMMU gain distributions differ from those of the original dataset (Figure 5 in Appendix E). The comparison makes the candidate CoT data part of the screening specification, alongside the base model and task group. When those data change, the relevant question is again which groups improve and which mode is better supported. Appendix E additionally reports the output length normalized metric as a descriptive measure.

Subsequent training stage. An additional reinforcement learning (RL) stage also changes some group decisions. The aggregate mode gaps on the two spatial benchmarks move closer to zero, while Object Count, Object Size, Route Plan, and CV-Bench Depth satisfy both CoT criteria after RL. The Object Size gap is 0.07 points, close to a tie. The mathematical results likewise include category changes despite a continuing aggregate CoT preference. Appendix C provides the full results and training recipe. These observations locate the screening decision in its current training configuration rather than treating it as a permanent label for the task or dataset.

5 Conclusion

Dual Tuning jointly screens data utility and supervision modes for multimodal supervised fine-tuning (SFT). One joint run per configuration supplies gains over the stronger base mode and comparisons between chain of thought (CoT) and direct answer (DA) evaluation across labeled task groups. The experiments show why both measurements matter. Positive CoT gains can coexist with larger DA gains, and a CoT advantage can coexist with negative gains in both modes. Subject level heterogeneity, changes across model and data configurations, and subset retraining together establish the value of assessing the available model and candidate data jointly. Dual Tuning provides a common empirical basis for choosing candidate subsets and their supervision modes before subsequent training.

AI USAGE DISCLOSURE

Generative AI tools were used during manuscript preparation solely to assist with the retrieval and discovery of relevant literature. The authors reviewed and verified all AI-assisted literature findings, retained full control over the scientific content and citation selection, and take responsibility for the final manuscript.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

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*, 2025b.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

Kaituo Feng, Manyuan Zhang, Hongyu Li, Kaixuan Fan, Shuang Chen, Yilei Jiang, Dian Zheng, Peiwen Sun, Yiyuan Zhang, Haoze Sun, et al. Onethinker: All-in-one reasoning model for image and video. *arXiv preprint arXiv:2512.03043*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

InclusionAI. Ming-omni: A unified multimodal model for perception and generation, 2025. URL https://arxiv.org/abs/2506.09344.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Qwen, An Yang, Baosong Yang, et al. Qwen2.5 technical report, 2024.

Qwen Team. Qwq-32b: Embracing the power of reinforcement learning. Qwen Blog, 2025. URL https://qwenlm.github.io/blog/qwq-32b/.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, et al. Sat: Dynamic spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10892–10902. IEEE, 2021.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, 2022. URL https://arxiv.org/abs/2201.11903.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting Influential Data for Targeted Instruction Tuning, 2024. URL https://arxiv.org/abs/2402.04333.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining, 2023. URL https://arxiv.org/abs/2305.10429.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025b.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–22, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal Chain-of-Thought Reasoning in Language Models, 2023. URL https://arxiv.org/abs/2302.00923.

A Full Multidisciplinary and Additional Model Scores

Chain of thought (CoT) and direct answer (DA) use prompt subscripts L and S. Base scores are B_L, B_S and joint model scores are DT_L, DT_S. Gains are relative percentages over the stronger base mode, while gaps are score differences. Red and green metric cells indicate negative and positive values. The full Qwen2.5-VL-7B spatial and mathematical scores appear in Tables 3 and 4 in the main text. Here we give its 30 subjects on Massive Multi-discipline Multimodal Understanding and Reasoning (MMMU), followed by all Ming-lite-omni v1.5 results. Counts in Table 2 exclude aggregate rows and apply the strict criteria in Section 3.2.

Table 5: Full Massive Multi-discipline Multimodal Understanding and Reasoning (MMMU) results for Qwen2.5-VL-7B. CoT and DA denote chain of thought and direct answer. Gains are relative percentages and gaps are score differences. The subject scores yield 9 CoT recommendations, 4 DA recommendations, 10 groups with two negative gains, and 7 boundary cases.

Task	$\mathbf{B_S}$	$\mathbf{B_L}$	$\mathbf{GAP_B}$	$\mathbf{DT_S}$	$\mathbf{DT_L}$	$\mathbf{Gain_{DA}}$	$\mathbf{Gain_{CoT}}$	$\mathbf{GAP_{DT}}$
MMMU	53.33	51.56	-1.77	51	54.11	-4.37	1.46	3.11
Art	70.00	70.00	0.00	73.33	73.33	4.76	4.76	0.00
Design	76.67	76.67	0.00	63.33	73.33	-17.40	-4.36	10.00
Music	30.00	26.67	-3.33	43.33	30.00	44.43	0.00	-13.33
Art Theory	83.33	76.67	-6.66	80.00	86.67	-4.00	4.01	6.67
Accounting	56.67	66.67	10.00	36.67	53.33	-45.00	-20.01	16.66
Finance	40.00	43.33	3.33	36.67	43.33	-15.37	0.00	6.66
Economics	56.67	53.33	-3.34	43.33	60.00	-23.54	5.88	16.67
Manage	43.33	36.67	-6.66	46.67	46.67	7.71	7.71	0.00
Marketing	63.33	60.00	-3.33	56.67	63.33	-10.52	0.00	6.66
Math	46.67	46.67	0.00	46.67	56.67	0.00	21.43	10.00
Biology	36.67	40.00	3.33	36.67	43.33	-8.33	8.33	6.66
Chemistry	33.33	43.33	10.00	36.67	40.00	-15.37	-7.69	3.33
Geography	43.33	40.00	-3.33	50.00	30.00	15.39	-30.76	-20.00
Physics	33.33	40.00	6.67	36.67	53.33	-8.33	33.33	16.66
Basic Med. Sci.	60.00	56.67	-3.33	63.33	66.67	5.55	11.12	3.34
Clinical Med.	60.00	70.00	10.00	53.33	66.67	-23.81	-4.76	13.34
Diag. & Lab. Med.	46.67	43.33	-3.34	33.33	33.33	-28.58	-28.58	0.00
Public Health	70.00	66.67	-3.33	60.00	66.67	-14.29	-4.76	6.67
Pharmacy	73.33	56.67	-16.66	56.67	66.67	-22.72	-9.08	10.00
History	73.33	60.00	-13.33	70.00	73.33	-4.54	0.00	3.33
Literature	86.67	80.00	-6.67	86.67	80.00	0.00	-7.70	-6.67
Psychology	66.67	60.00	-6.67	60.00	76.67	-10.00	15.00	16.67
Sociology	53.33	50.00	-3.33	63.33	66.67	18.75	25.01	3.34
Agriculture	50.00	50.00	0.00	63.33	50.00	26.66	0.00	-13.33
Arch. & Eng.	36.67	33.33	-3.34	36.67	33.33	0.00	-9.11	-3.34
Comp. Sci.	60.00	56.67	-3.33	50.00	50.00	-16.67	-16.67	0.00
Electronics	33.33	43.33	10.00	33.33	36.67	-23.08	-15.37	3.34
Energy & Power	43.33	36.67	-6.66	53.33	43.33	23.08	0.00	-10.00
Materials	23.33	30.00	6.67	30.00	36.67	0.00	22.23	6.67
Mech. Eng.	50.00	33.33	-16.67	30.00	23.33	-40.00	-53.34	-6.67

Table 6: Spatial results for Ming-lite-omni v1.5 under chain of thought (CoT) and direct answer (DA) evaluation. Gains are relative percentages and gaps are score differences. Ten groups favor DA. Relative Direction and Route Plan have negative gains in both modes despite positive trained mode gaps.

Task	$\mathbf{B_S}$	$\mathbf{B_L}$	$\mathbf{GAP_B}$	$\mathbf{DT_S}$	$\mathbf{DT_L}$	$\mathbf{Gain_{DA}}$	$\mathbf{Gain_{CoT}}$	$\mathbf{GAP_{DT}}$
VSI-Bench	36.07	36.69	0.62	49.40	38.98	34.64	6.24	-10.42
Obj. Count	40.23	40.32	0.09	51.98	29.04	28.92	-27.98	-22.94
Abs. Dist.	13.68	17.22	3.54	37.53	25.22	117.94	46.46	-12.31
Obj. Size	43.89	45.62	1.73	66.14	54.39	44.98	19.22	-11.75
Room Size	31.77	29.06	-2.71	53.16	29.93	67.33	-5.79	-23.23
Rel. Dist.	45.07	44.79	-0.28	52.54	42.82	16.57	-4.99	-9.72
Rel. Dir.	43.63	45.61	1.98	41.12	45.21	-9.84	-0.88	4.09
Route Plan	29.38	33.51	4.13	30.93	31.96	-7.70	-4.63	1.03
Appr. Order	40.94	37.38	-3.56	61.81	53.24	50.98	30.04	-8.57
CV-Bench	74.53	76.69	2.16	84.99	81.05	10.82	5.69	-3.94
Count	68.91	65.23	-3.68	72.21	66.50	4.79	-3.50	-5.71
Relation	81.23	89.08	7.85	95.38	90.15	7.07	1.20	-5.23
Depth	84.17	78.33	-5.84	90.67	87.67	7.72	4.16	-3.00
Distance	65.00	76.67	11.67	84.83	83.67	10.64	9.13	-1.16

Table 7: MathVista results for Ming-lite-omni v1.5 under chain of thought (CoT) and direct answer (DA) evaluation. Gains are relative percentages and gaps are score differences. Geometry, Logical, and Statistical Reasoning favor CoT. The aggregate CoT score exceeds DA but remains below the stronger base mode.

Task	$\mathbf{B_S}$	$\mathbf{B_L}$	$\mathbf{GAP_B}$	$\mathbf{DT_S}$	$\mathbf{DT_L}$	$\mathbf{Gain_{DA}}$	$\mathbf{Gain_{CoT}}$	$\mathbf{GAP_{DT}}$
MathVista	72.17	79.00	6.83	70.80	78.77	-10.38	-0.29	7.97
Comm.	52.31	49.77	-2.54	50.46	48.61	-3.54	-7.07	-1.85
Scientific	72.62	77.68	5.06	75.89	75.60	-2.30	-2.68	-0.29
Geometry	82.33	87.79	5.46	79.60	88.07	-9.33	0.32	8.47
Arithmetic	73.48	89.85	16.37	69.55	89.70	-22.59	-0.17	20.15
Logical	12.61	27.03	14.42	15.32	29.73	-43.32	9.99	14.41
Statistical	84.47	88.31	3.84	82.90	89.70	-6.13	1.57	6.80
Algebraic	72.40	80.21	7.81	76.04	76.56	-5.20	-4.55	0.52

Table 8: Full Massive Multi-discipline Multimodal Understanding and Reasoning (MMMU) results for Ming-lite-omni v1.5. CoT and DA denote chain of thought and direct answer. Gains are relative percentages and gaps are score differences. The subject scores yield 8 CoT recommendations, 3 DA recommendations, 10 groups with two negative gains, and 9 boundary cases.

Task	$\mathbf{B_S}$	$\mathbf{B_L}$	$\mathbf{GAP_B}$	$\mathbf{DT_S}$	$\mathbf{DT_L}$	$\mathbf{Gain_{DA}}$	$\mathbf{Gain_{CoT}}$	$\mathbf{GAP_{DT}}$
MMMU	52.89	56.44	3.55	53.11	59.11	-5.9	4.73	6
Art	66.67	70.00	3.33	66.67	66.67	-4.76	-4.76	0.00
Design	83.33	76.67	-6.66	70.00	70.00	-16.00	-16.00	0.00
Music	40.00	26.67	-13.33	46.67	16.67	16.68	-58.33	-30.00
Art Theory	83.33	70.00	-13.33	80.00	80.00	-4.00	-4.00	0.00
Accounting	46.67	66.67	20.00	46.67	73.33	-30.00	9.99	26.66
Finance	33.33	56.67	23.34	36.67	63.33	-35.29	11.75	26.66
Economics	46.67	70.00	23.33	50.00	73.33	-28.57	4.76	23.33
Manage	40.00	60.00	20.00	46.67	50.00	-22.22	-16.67	3.33
Marketing	53.33	80.00	26.67	63.33	86.67	-20.84	8.34	23.34
Math	50.00	56.67	6.67	26.67	50.00	-52.94	-11.77	23.33
Biology	40.00	40.00	0.00	46.67	46.67	16.68	16.68	0.00
Chemistry	30.00	43.33	13.33	36.67	53.33	-15.37	23.08	16.66
Geography	63.33	43.33	-20.00	56.67	50.00	-10.52	-21.05	-6.67
Physics	50.00	60.00	10.00	43.33	60.00	-27.78	0.00	16.67
Basic Med. Sci.	56.67	66.67	10.00	63.33	70.00	-5.01	4.99	6.67
Clinical Med.	63.33	56.67	-6.66	60.00	63.33	-5.26	0.00	3.33
Diag. & Lab. Med.	43.33	33.33	-10.00	30.00	40.00	-30.76	-7.69	10.00
Public Health	66.67	83.33	16.66	63.33	83.33	-24.00	0.00	20.00
Pharmacy	66.67	80.00	13.33	63.33	73.33	-20.84	-8.34	10.00
History	63.33	66.67	3.34	70.00	83.33	4.99	24.99	13.33
Literature	80.00	90.00	10.00	80.00	83.33	-11.11	-7.41	3.33
Psychology	63.33	56.67	-6.66	56.67	56.67	-10.52	-10.52	0.00
Sociology	66.67	53.33	-13.34	66.67	60.00	0.00	-10.00	-6.67
Agriculture	40.00	53.33	13.33	53.33	46.67	0.00	-12.49	-6.66
Arch. & Eng.	36.67	26.67	-10.00	36.67	33.33	0.00	-9.11	-3.34
Comp. Sci.	53.33	56.67	3.34	63.33	60.00	11.75	5.88	-3.33
Electronics	40.00	33.33	-6.67	43.33	43.33	8.33	8.33	0.00
Energy & Power	46.67	40.00	-6.67	60.00	53.33	28.56	14.27	-6.67
Materials	36.67	40.00	3.33	33.33	40.00	-16.68	0.00	6.67
Mech. Eng.	36.67	36.67	0.00	33.33	43.33	-9.11	18.16	10.00

B Data and Comparison Protocol

Data and group association. The spatial configuration contains 30k image and 8k video question answering pairs constructed from the sources listed in Section 4.1. The selected OneThinker configuration contains 20k image pairs. Chain of thought (CoT) and direct answer (DA) targets share their visual inputs, questions, and final answers. An external model application programming interface (API) uniformly generates group labels for the candidate training samples. These labels associate candidate training subsets with evaluation groups. These associations are inputs to the protocol rather than labels inferred from score gaps. A new application requires a group mapping suited to its data and evaluation targets.

Complete training comparisons. Table 9 retains both corresponding and alternative evaluation prompts for all reported separate and joint configurations. I and V denote image and video spatial data, and O denotes selected OneThinker data. Training formats S and L correspond to DA and CoT. The joint spatial configuration combines modalities, while the separate spatial baselines use one modality each. Main text Table 1 selects the corresponding training and evaluation modes, without choosing maxima from the alternative prompt scores. The VSI-Bench base scores are consistently 37.57 for DA and 24.91 for CoT.

Scope of the screening. The screen is conditioned on the base model, paired data, group mapping, prompts, evaluator, and training recipe. It reports group outcomes after shared training rather than causal effects of individual examples. The current study covers spatial, mathematical, and multidisciplinary tasks. The subset experiments evaluate their associated groups, and a final reassembled mixture defines another training configuration. Output mode comparisons concern the specified prompts and evaluators. These conditions define the practical use of the recommendations.

Table 9: Full separate and joint training comparisons for Qwen2.5-VL-7B. I and V denote image and video spatial data, and O denotes selected OneThinker data. S and L denote direct answer and chain of thought formats. Eval-S and Eval-L use their corresponding prompts. VSI and CV denote VSI-Bench and CV-Bench. MMMU denotes Massive Multi-discipline Multimodal Understanding and Reasoning. Alternative prompt evaluations are retained but are not used as matching mode comparisons.

(a) Spatial Tasks

Model	VSI (video)		CV (image)	
	Eval-S	Eval-L	Eval-S	Eval-L
Baseline	37.57	24.91	75.17	73.81
E1: I-S	37.94	37.34	84.15	84.27
E2: I-L	12.96	37.38	75.51	80.55
E3: V-S	50.98	50.50	78.20	77.82
E4: V-L	0.00	44.14	37.72	73.09
E5: IV-SL	51.35	43.26	84.00	80.25

(b) Disciplinary Tasks

Model	MathVista		MMMU	
	Eval-S	Eval-L	Eval-S	Eval-L
Baseline	69.80	69.83	53.33	51.56
E6: O-S	67.43	67.43	52.00	51.00
E7: O-L	59.13	73.40	25.89	53.33
E8: O-SL	67.37	72.87	51.00	54.11

C Additional Training Stage

We examine reinforcement learning (RL) initialized from the jointly trained model using Group Relative Policy Optimization (GRPO) (Guo et al., 2025). The RL data comprise 10k image and 10k video spatial question answering pairs and 10k OneThinker image pairs. Training uses 32 NVIDIA H800 accelerators, AdamW, four epochs, a learning rate of 1×10^{-6}, a batch size of one per accelerator, and at most 64 video frames. GRPO uses a group size of eight and a Kullback–Leibler divergence coefficient of 0.01.

Superscript RL denotes scores after this additional stage, with gains computed against the same base model reference. On the spatial benchmarks, the chain of thought (CoT) gap over direct answer (DA) evaluation changes from -8.09 to -1.62 points on VSI-Bench and from -3.75 to -0.08 points on CV-Bench (Table 10). Under the same strict criteria, Object Count, Object Size, Route Plan, and CV-Bench Depth favor CoT after RL. The Object Size gap is 0.07 points, close to a tie. On MathVista, the aggregate mode gap increases from 5.50 to 6.74 points (Table 11). Numeric Commonsense and Scientific Reasoning become CoT favored, while Logical Reasoning reaches a zero CoT gain. These results show that an additional training stage can change group decisions even when the aggregate mode ordering persists.

Table 10: Spatial gains and mode gaps before and after reinforcement learning (RL) with Group Relative Policy Optimization. CoT and DA denote chain of thought and direct answer. Superscript RL indicates the additional stage, with the base model reference unchanged. Gains are relative percentages and gaps are score differences. Near zero differences are descriptive point estimates.

Task	$\mathbf{Gain_{DA}}$	$\mathbf{Gain_{CoT}}$	$\mathbf{GAP_{DT}}$	$\mathbf{Gain^{RL}_{DA}}$	$\mathbf{Gain^{RL}_{CoT}}$	$\mathbf{GAP^{RL}_{DT}}$
VSI-Bench	36.68	15.15	-8.09	36.81	32.50	-1.62
Obj. Count	22.87	7.57	-6.18	23.29	29.82	2.64
Abs. Dist.	75.57	29.72	-10.49	76.75	75.31	-0.33
Obj. Size	28.14	1.57	-13.33	28.36	28.50	0.07
Room Size	46.25	-0.62	-18.26	40.81	13.91	-10.48
Rel. Dist.	45.27	19.66	-10.28	47.01	33.68	-5.35
Rel. Dir.	12.05	10.49	-0.62	12.48	10.77	-0.68
Route Plan	4.77	0.00	-1.55	9.55	22.24	4.12
Appr. Order	65.50	54.88	-4.04	64.24	56.59	-2.91
CV-Bench	11.75	6.76	-3.75	11.85	11.75	-0.08
Count	9.61	-4.51	-9.14	8.24	7.26	-0.63
Relation	9.12	5.96	-2.77	9.31	8.60	-0.62
Depth	20.83	20.83	0.00	21.99	23.61	1.17
Distance	6.44	4.57	-1.50	7.06	7.06	0.00

Table 11: MathVista gains and mode gaps before and after reinforcement learning (RL) with Group Relative Policy Optimization. CoT and DA denote chain of thought and direct answer. Superscript RL indicates the additional stage, with the base model reference unchanged. Gains are relative percentages and gaps are score differences. Numeric Commonsense, Scientific Reasoning, and Logical Reasoning change screening status.

Task	$\mathbf{Gain_{DA}}$	$\mathbf{Gain_{CoT}}$	$\mathbf{GAP_{DT}}$	$\mathbf{Gain^{RL}_{DA}}$	$\mathbf{Gain^{RL}_{CoT}}$	$\mathbf{GAP^{RL}_{DT}}$
MathVista	-3.52	4.35	5.50	-3.58	6.07	6.74
Comm.	10.42	3.65	-3.01	11.48	11.99	0.23
Scientific	0.00	0.00	0.00	0.40	0.80	0.30
Geometry	-9.17	5.70	10.49	-10.19	7.53	12.50
Arithmetic	-11.1	2.08	10.60	-11.49	2.26	11.06
Logical	-34.37	-18.77	4.50	-34.37	0.00	9.91
Statistical	1.44	3.30	1.57	0.63	3.92	2.79
Algebraic	-6.61	8.82	10.93	-1.47	8.09	6.77

D SUBSET RETRAINING PROTOCOL

Figures 2, 3, and 4 report subset retraining with chain of thought (CoT), direct answer (DA), and joint supervision, respectively. All three plots use the initial DA and CoT screening gains as coordinates. Marker shapes encode the original subset assignments, while colors encode the sign of the gain after subset retraining. They do not show the numerical magnitude of the retraining gain. In the DA comparison, colors describe the DA gain. In the joint comparison, colors describe the larger gain across the two modes.

The original CoT and DA partitions are retained as depicted, including their assignments of zero gain groups. These historical assignments are distinct from the strict mode recommendation categories used in Table 2. In the joint experiment, groups with two negative gains form one subset and all remaining groups form its complement. The complement contains boundary cases as well as groups with positive gains. Each model is evaluated on the groups associated with its selected training data.

The predominantly matching signs support prioritizing subsets for further curation. Exceptions are visible in the plots. Materials has a positive initial CoT gain but a negative gain after CoT subset retraining, while Accounting has two negative screening gains but a positive best mode gain after joint subset retraining. These observations characterize the effect of changing the training mixture without converting the screen into a claim about fixed sample quality.

E CANDIDATE DATA AND DESCRIPTIVE MEASURES

An alternative candidate dataset uses chain of thought (CoT) traces constructed with Qwen3-VL-235B (Bai et al., 2025a). Figure 5 compares its gains on Massive Multi-discipline Multimodal Understanding and Reasoning (MMMU) with the original dataset. Circles and triangles represent the original and alternative data, while colors identify broad subject domains. The axes show direct answer (DA) and CoT gains. Their distributions describe the two candidate data configurations considered in Section 4.5.

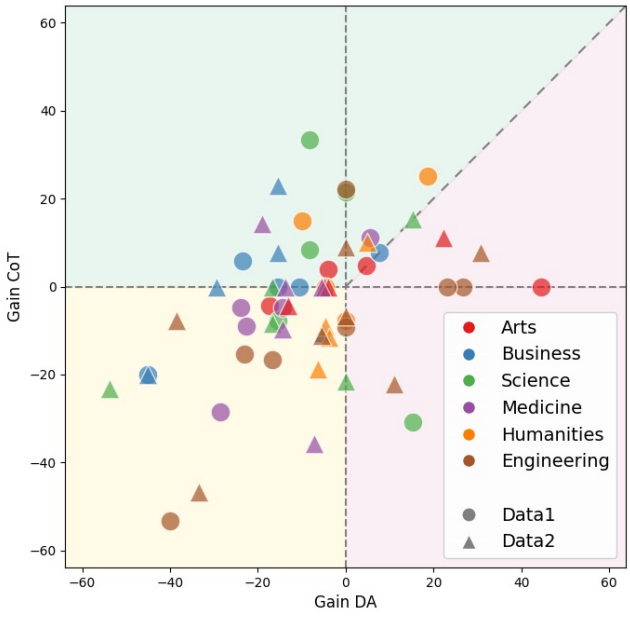

Figure 5: Subject gains on Massive Multi-discipline Multimodal Understanding and Reasoning (MMMU) for the original and alternative chain of thought (CoT) datasets, shown as circles and triangles. Axes are relative gains under direct answer and CoT evaluation. Colors identify subject domains. The comparison describes the two candidate data configurations.

The base model difference $\mathbf{GAP_B}$ compares CoT with DA evaluation before training (Figure 6). The length normalized measure $\mathbf{Gain_{token}}$ divides the relative CoT gain by the mean CoT output

length, with the scaling defined in Section 3.2 (Figure 7). It provides an auxiliary description of the candidate configurations and is not part of the screening rule. We use it as a length normalized gain rather than a measure of training cost or reasoning trace quality.

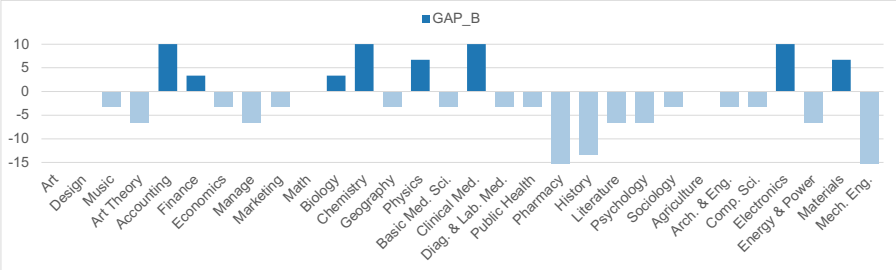

Figure 6: Base model differences between chain of thought (CoT) and direct answer (DA) scores across subjects of Massive Multi-discipline Multimodal Understanding and Reasoning (MMMU). Positive $\mathbf{GAP_B}$ values indicate higher CoT scores before joint training.

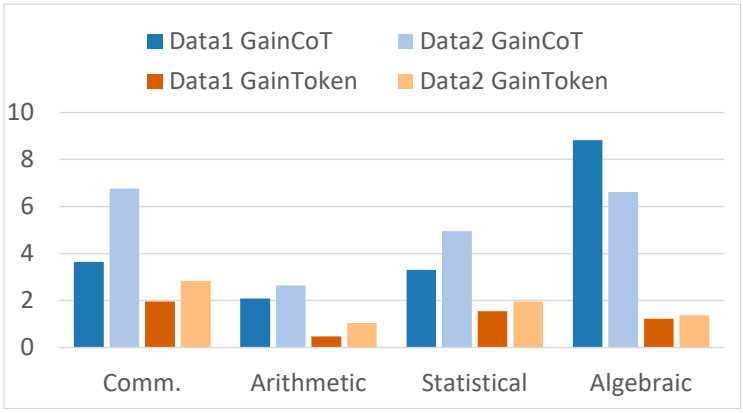

Figure 7: Relative gain under chain of thought evaluation, $\mathbf{Gain_{CoT}}$, and its output length normalized counterpart, $\mathbf{Gain_{token}}$, for the two candidate datasets. The two metrics differ in their ordering for Algebraic Reasoning. The length normalized metric is auxiliary and is not used for screening.