SAME PROBLEM, DIFFERENT START: CONTROLLING TEXT-IMAGE MODALITY GAPS VIA FIRST-DECISION DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work has shown that presenting the same problem in Image rather than Text form can induce substantial performance and reasoning differences in multimodal large language models (MLLMs), while distilling full Text-side reasoning traces can substantially recover Image-side performance. But does visual input truly impair the model's ability to reason, or does it simply send a capable model down the wrong generation trajectory, such that minimal steering can redirect it? We find that a remarkably localized intervention can suffice: altering only the first generation decision can substantially redirect the frozen continuation. On a held-out set of Image-presented GSM8K problems, suppressing answer entry only at the first generation position raises accuracy from 41.7% to 93.0%; conversely, forcing naturally non-answer-start Text responses toward answer entry reduces accuracy from 94.8% to 28.9%. Motivated by this finding, we introduce First-Decision Distillation (FDD), which distills only the teacher's first-step decision and uses a clean handoff to leave all subsequent generation to the frozen base model. With only 2 paired training examples and 16 presentations per task, FDD-Logit reduces the Qwen3-VL-8B Text–Image accuracy gap from 12.10 to 1.03 percentage points (pp), a 91.5% reduction. Evaluations across model families, presentation directions, prompt formulations, and structured-chart inputs further test the scope of first-decision control, including an improvement of 10.92 pp from reverse Image-to-Text FDD on Gemma-3-12B. These results identify response onset as a high-leverage, context-dependent control point for reducing Text–Image modality gaps without imitating the teacher's full response.

1 INTRODUCTION

Prior work has shown that presenting the same problem as Text or Image can produce substantial performance and behavioral discrepancies in multimodal large language models (MLLMs) (Lyu et al., 2025; Zhang et al., 2024; van Sprang et al., 2026; Liu et al., 2026). More recently, Sun et al. (2026) showed that visual presentation can suppress visible reasoning, while full Text-side reasoning distillation can partially recover Image-side performance. This leaves open whether visual input impairs underlying reasoning capacity or primarily redirects generation at response onset. If the latter, sequence-wide correction may be unnecessary.

We first examine 400 paired Text–Image GSM8K problems on Qwen3-VL-8B. Image responses contain less visible reasoning than their Text counterparts in 63.8% of cases, but more in 32.2%, revealing bidirectional rather than uniformly shorter Image-side behavior. Figure 1 quantifies this paired shift, while Figure 2 illustrates representative trajectories in both directions. This heterogeneity motivates testing whether the divergence can be localized to response onset.

This motivates a direct causal test at response onset. We therefore intervene only at the first generation position and leave all subsequent generation to the frozen base model. On held-out GSM8K cohorts, suppressing answer entry raises Image accuracy from 41.7% to 93.0%, whereas forcing answer entry reduces Text accuracy from 94.8% to 28.9%. These bidirectional interventions identify response onset as a high-leverage, context-dependent control point.

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.

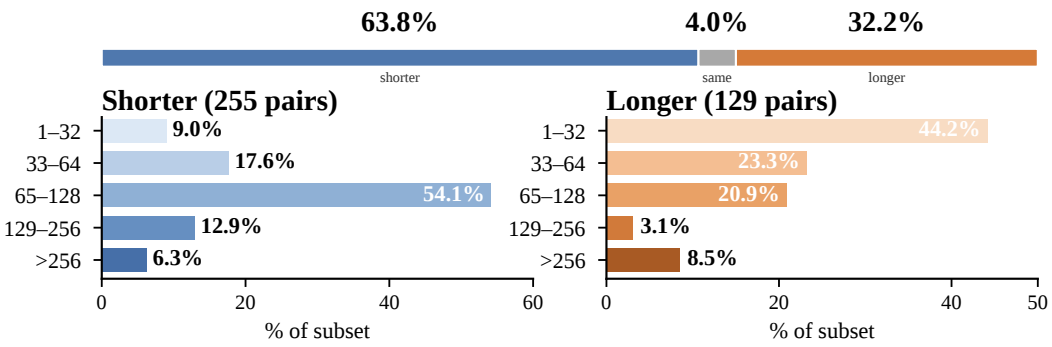


Figure 1: Paired Image-minus-Text differences in visible reasoning across 400 paired GSM8K test problems on Qwen3-VL-8B. Image responses contain fewer visible reasoning tokens in 63.8% of pairs, more in 32.2%, and the same number in 4.0%; the lower panels show the magnitude distributions within the shorter and longer subsets.

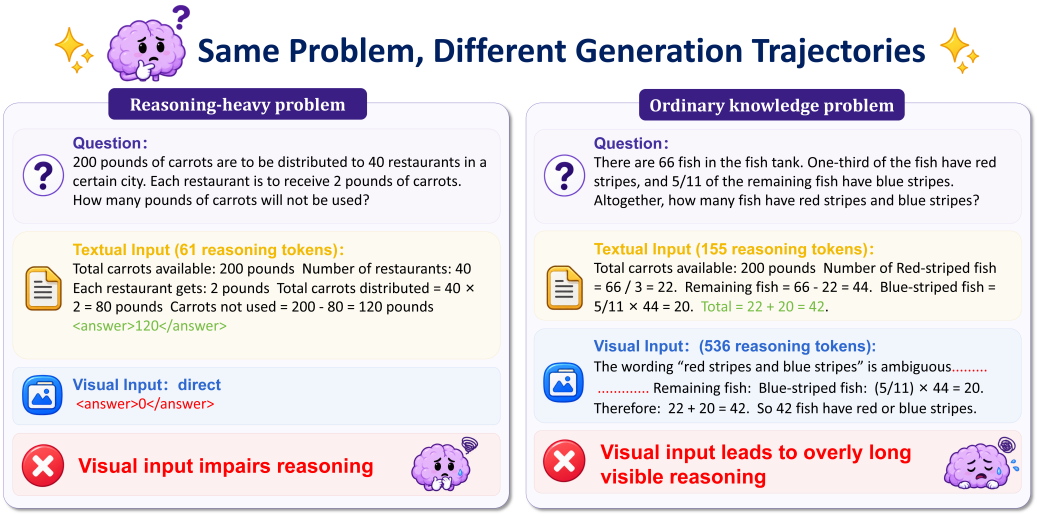


Figure 2: Examples of bidirectional Text–Image trajectory shifts. Left: the Image response enters the answer prematurely and fails. Right: another Image response produces substantially more visible reasoning than its paired Text response.

Motivated by this finding, we introduce *First-Decision Distillation* (FDD), which distills only the teacher's first generation decision. A clean handoff then returns all subsequent generation to the frozen base model, implementing a "steer once, then hand off" protocol without full-response supervision.

This localized supervision target is highly data-efficient. Across five benchmarks, with only $N = 2$ paired examples and $P = 16$ presentations per task, FDD-Logit reduces the Qwen3-VL-8B Text–Image accuracy gap from 12.10 to 1.03 ± 0.06 pp. Evaluations across model families, presentation directions, prompt formulations, and structured-chart inputs further characterize the scope and limits of first-decision control.

Our main contributions are as follows:

(1) Response onset is a high-leverage control point for Text–Image modality gaps. We show that behavioral and performance discrepancies between Text and Image presentations in MLLMs are largely governed by trajectory divergence at response onset rather than an inherent loss of reasoning capacity. Through bidirectional single-position interventions, we demonstrate that altering only the first generation decision can substantially redirect the continuation generated by the frozen base model and bridge presentation-induced performance drops. The effect is causal but context dependent, rather than a fixed reasoning-versus-answer rule.

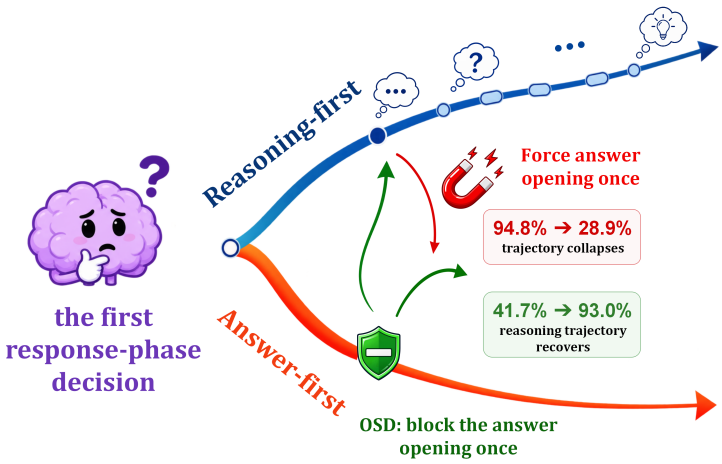


Figure 3: Changing only the first generation decision can redirect the frozen model's continuation, motivating first-decision supervision in FDD.

(2) First-Decision Distillation with a clean handoff. We propose FDD, a localized distillation framework that supervises only the first generation decision. By discarding the adapted key–value cache and resuming generation after a fresh prefill of the frozen base model, FDD implements a "steer once, then hand off" protocol that separates onset steering from continuation generation without requiring full-response supervision.

(3) Data-efficient onset control across diverse settings. We show that FDD is data-efficient: with only two paired training examples per task ($N = 2$, $P = 16$ presentations), FDD-Logit reduces the Qwen3-VL-8B Text–Image accuracy gap by 91.5%. Furthermore, evaluations across three model families (Qwen, Gemma, and InternVL), presentation directions (T $\rightarrow$ I and I $\rightarrow$ T), five prompt formulations, and structured-chart inputs on PlotQA characterize the scope and boundary conditions of first-decision control.

2 From Cross-Presentation Mismatch to First-Decision Control

We study paired Text and Image presentations of the same semantic problem. Let x denote the underlying problem, T and I its Text and Image presentations, and c the requested output format, or response contract. Within each pair, we hold the semantic content and response contract fixed and examine how presentation format changes the generation trajectory. Appendix B provides the rendering, instruction-delivery, and cohort details.

The resulting mismatch is not determined by presentation format alone. In a 200-example GSM8K control cohort, the Text–Image accuracy gap is 30.0 pp under AF-XML but only 2.0 pp under RR-text, despite identical underlying problems. Presentation effects therefore interact strongly with the response contract rather than reflecting a fixed Image-side degradation. Full controls are reported in Appendix D.

We next ask whether this context-dependent divergence can be causally influenced at the very beginning of generation. We use the term *first decision* to denote the model's decoding choice at the first generation step ($t = 0$). Operationally, this choice is represented by the first-step token scores or distribution and is realized, under the decoding rule, as the first generated token y_0. The first decision and the first token are therefore related but not identical: the former denotes the model's choice at the first decoding step, whereas the latter is the token emitted as its outcome.

Forward intervention. We first test whether preventing premature answer entry can redirect the continuation. When the first token of the designated answer opening is top-1 at $t = 0$, we mask only that token for the first decoding step and then return immediately to unmodified decoding. On a

held-out 300-example Image/AF-XML cohort, this one-position intervention raises accuracy from 41.7% to 93.0%, rescuing 155 examples while harming only one.

Reverse induction and token specificity. We then reverse the intervention. Among 194 Text/AF-XML examples whose base responses do not naturally begin with the answer opening, we promote the first token of that opening to just above the maximum logit at $t = 0$. Every resulting response enters the answer immediately: visible reasoning disappears and accuracy falls from 94.8% to 28.9%. In contrast, promoting an eligible non-answer token under the same rule leaves accuracy high. The corresponding Image/RR-text probe exhibits the same qualitative pattern; see Appendix E. Thus, the effect is not explained by an arbitrary perturbation at the first position. Changing the specific onset decision can redirect the frozen continuation even though the input representation and all later decoding steps are left unchanged.

Table 1: First-position interventions on separate GSM8K cohorts. The reverse probe uses greedy decoding on naturally non-answer-start outputs. CI denotes confidence interval.

Input / n	Intervention	Base (%)	Changed (%)	Delta Acc. (pp)	95% CI
Image / 300	Block opening	41.7	93.0	+51.3	[45.7, 57.0]
Text / 194	Promote opening	94.8	28.9	−66.0	[−72.7, −59.3]
Text / 194	Promote non-answer	94.8	96.4	+1.5	[0.0, 3.6]

Together, these interventions establish response onset as a high-leverage but context-dependent control point. Importantly, the useful decision is not a universal "reason first" or "avoid the answer" rule: suppressing answer entry is strongly beneficial in the GSM8K Image setting, whereas the same intervention has the opposite effect on PlotQA (Section 4.4). The problem is therefore not to prescribe a fixed first token, but to select an appropriate first-step decision for the current input and response context. This motivates learning that decision from paired supervision, which leads to First-Decision Distillation in Section 3.

3 First-Decision Distillation

FDD transfers the teacher's first-step decision to the student presentation while leaving all subsequent generation to the frozen base model. As defined in Section 2, the first decision occurs at generation step $t = 0$ and is realized as the first generated token y_0.

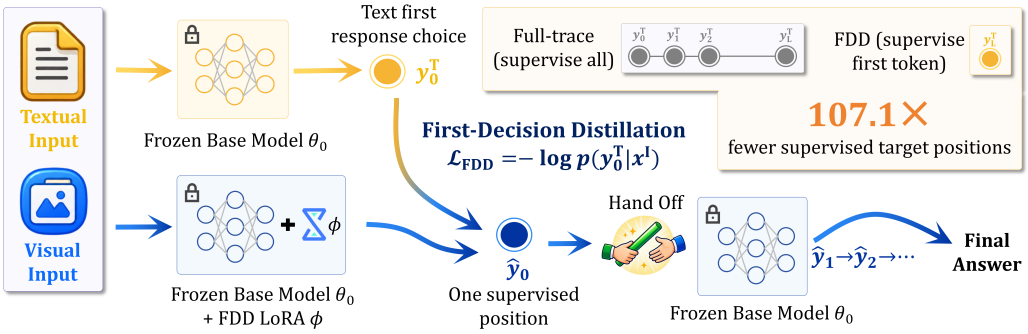


Figure 4: FDD in the Text-to-Image direction. The Image-conditioned adapter learns the paired Text teacher's first-step target. At inference, the adapted model is used for only the first decoding step to produce y_0; its key–value cache is then discarded, and the frozen base model is freshly prefilled with the original prompt and y_0 to generate the continuation. The approximately 107× reduction refers to supervised target positions in the matched full-data 8B comparison, not to an end-to-end speedup.

Hard and distributional first-step targets. Let θ_0 denote the frozen base model and ϕ the trainable adapter parameters. Consider a task with N paired training problems; the pooled objective averages

the corresponding task-level losses. For problem i, the Text presentation provides the teacher signal at the first generation step.

Hard FDD retains only the realized greedy choice of the Text teacher,

$$y^{\mathrm{T}}_{i,0} = \arg\max_{y} p_{\theta_0}(y \mid x_i, \mathrm{T}, c), \qquad \mathscr{L}_{\mathrm{hard}}(\phi) = -\frac{1}{N}\sum_{i=1}^{N} \log p_{\theta_0,\phi}\left(y^{\mathrm{T}}_{i,0} \mid x_i, \mathrm{I}, c\right). \tag{1}$$

We use rank-8 Low-Rank Adaptation (LoRA) on the language model's $q/k/v/o$ attention projections (Hu et al., 2022), while keeping the base language model and vision encoder frozen. No response position after $t = 0$ is supervised. FDD-Logit instead retains the teacher's full vocabulary distribution at the same first generation step. Let

$$q_i(v) = \left[\mathrm{softmax}\left(z^{\mathrm{T}}_i\right)\right]_v,$$

where z^{T}_i denotes the Text teacher's first-step logits and $\mathscr{V}$ the token vocabulary. The objective is

$$\mathscr{L}_{\mathrm{logit}}(\phi) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{v\in\mathscr{V}} q_i(v) \log p_{\theta_0,\phi}\left(v \mid x_i, \mathrm{I}, c\right). \tag{2}$$

For comparison, Full-Trace LoRA supervises the entire paired Text response of length L_i:

$$\mathscr{L}_{\mathrm{full}}(\phi) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{L_i-1} \log p_{\theta_0,\phi}\left(y^{\mathrm{T}}_{i,t} \mid x_i, \mathrm{I}, c, y^{\mathrm{T}}_{i,<t}\right). \tag{3}$$

Hard FDD and FDD-Logit therefore supervise the same decoding step but retain different amounts of teacher information: Hard FDD uses only the teacher's realized first-token choice, whereas FDD-Logit matches the full first-step vocabulary distribution. Full-Trace LoRA uses the same adapter family but supervises the complete response trajectory. Equations (1)–(2) describe the primary Text-to-Image direction; Reverse FDD exchanges the teacher and student presentations.

One adapted decision, then a clean handoff. At inference, the adapted model is used for one decoding step to obtain the first generated token $\widehat{y}_0$. We then discard its adapted key–value cache, disable LoRA, and freshly prefill the frozen base model with the original Image prompt followed by $\widehat{y}_0$. The resulting generation process is

$$\widehat{y}_0 \sim p_{\theta_0,\phi}(\cdot \mid x, \mathrm{I}, c), \qquad \widehat{y}_t \sim p_{\theta_0}(\cdot \mid x, \mathrm{I}, c, \widehat{y}_{<t}), \quad t \geq 1. \tag{4}$$

Fresh prefill ensures that the adapter can influence the continuation only through the emitted first token $\widehat{y}_0$; all later hidden states and logits are recomputed by the frozen base model. No paired teacher is required at inference. *One-Decision Full-LoRA* uses the same hard one-position objective but keeps the adapter active throughout decoding. Comparing it with Hard FDD under matched full-data training isolates the effect of the clean handoff (Appendix F).

Effect of the clean handoff. Under a fixed decoding rule, the sequence distribution induced by FDD factorizes as

$$P_{\mathrm{FDD}}(Y \mid x, \mathrm{I}, c) = \pi_\phi(y_0 \mid x, \mathrm{I}, c)\, K_{\theta_0}(Y_{1:} \mid x, \mathrm{I}, c, y_0), \tag{5}$$

where $Y = (y_0, Y_{1:})$, π_ϕ is the adapted distribution over the first generated token, and K_{θ_0} is the frozen base model's conditional continuation distribution. FDD therefore changes which frozen-model continuation is entered through its choice of y_0, while leaving the conditional continuation mechanism K_{θ_0} unchanged. Later tokens still depend on y_0 through the ordinary autoregressive context; the locality claim is only that the adapter's direct computational influence ends after the first decoding step. Appendix C gives the corresponding score decomposition.

4 Experiments

4.1 Protocol and Paired Metrics

We conduct our primary experiments on Qwen3-VL-2B-Instruct and Qwen3-VL-8B-Instruct (Bai et al., 2025) across GSM8K, BBH-seen, MMLU, ARC-Challenge, and CommonsenseQA. Each

benchmark has 800 training, 200 validation, and 400 test examples. Unless otherwise stated, we use the AF-XML response contract, temperature 0.1, repetition penalty 1.0, a 1024-token cap, and shared per-example decoding seeds across paired presentations. Rendering and data construction are specified in Appendix B.

The main low-data protocol uses $N = 2$ unique Text–Image pairs and $P = 16$ total training presentations per task; the selected budget and subset are frozen before test evaluation. Hard FDD and FDD-Logit use the same pairs and balanced schedule, while Full-Trace LoRA and One-Decision Full-LoRA provide full-data references (Appendix F). Detailed exposure counts, checkpoint selection, and scale-transfer settings are given in Appendix B.

We report task accuracy together with the macro absolute Text–Image gap, correctness agreement, phase agreement, onset KL (the teacher-to-student first-step divergence), teacher top-1 agreement, and strict instruction compliance. Metric definitions and aggregation rules are given in Appendix B.4.

We additionally evaluate separate adapters on InternVL3.5-8B in the Text-to-Image direction and Gemma-3-12B in the reverse Image-to-Text direction, and test pairing sensitivity, answer-only prompt formulations, and structured-chart inputs from PlotQA. Their budgets and evaluation details are reported in the corresponding appendices.

4.2 Alignment Across Tasks and Model Scales

Table 2: Five-benchmark AF-XML test results. Pairs and presentations are counted per task. Hard FDD and FDD-Logit use $N = 2$, $P = 16$; Full-Trace and One-Dec. full are full-data references. Learned results are mean ± standard deviation (SD) over three seeds. Parentheses give relative gap reduction or changes in correctness and phase agreement from Base Image, in percentage points.

Scale	Method	Pairs	Pres.	Acc. gap (pp) ↓	Corr. agree. (%) ↑	Phase agree. (%) ↑	Onset KL (nats) ↓
2B	Base Image	–	–	11.40	72.60	84.35	–
	Full-Trace	800	2400	**3.35 ± 0.73** (−70.6%)	**81.48 ± 0.98** (+8.88 pp)	98.40 ± 0.08 (+14.05 pp)	–
	One-Dec. full	800	2400	6.27 ± 0.69 (−45.0%)	79.67 ± 0.39 (+7.07 pp)	**98.92 ± 0.02** (+14.57 pp)	–
	Hard FDD	2	16	3.52 ± 0.03 (−69.1%)	78.32 ± 0.03 (+5.72 pp)	**98.92 ± 0.02** (+14.57 pp)	1.312 ± 0.035
	FDD-Logit	2	16	3.45 ± 0.05 (−69.7%)	78.52 ± 0.08 (+5.92 pp)	98.17 ± 0.21 (+13.82 pp)	**0.694 ± 0.009**
8B	Base Image	–	–	12.10	80.00	79.65	–
	Full-Trace	800	2400	3.07 ± 0.20 (−74.6%)	90.30 ± 0.36 (+10.30 pp)	95.85 ± 0.46 (+16.20 pp)	–
	One-Dec. full	800	2400	1.05 ± 0.15 (−91.3%)	90.35 ± 0.35 (+10.35 pp)	**97.73 ± 0.24** (+18.08 pp)	–
	Hard FDD	2	16	**0.65 ± 0.00** (−94.6%)	90.45 ± 0.00 (+10.45 pp)	96.57 ± 0.10 (+16.92 pp)	1.149 ± 0.043
	FDD-Logit	2	16	1.03 ± 0.06 (−91.5%)	**90.97 ± 0.06** (+10.97 pp)	97.35 ± 0.13 (+17.70 pp)	**0.335 ± 0.016**

On Qwen3-VL-8B, FDD-Logit increases Image accuracy from 73.10% to 85.17 ± 0.06% and reduces the mean absolute Text–Image gap from 12.10 to 1.03 ± 0.06 pp, a 91.5% reduction (Table 2). Hard FDD achieves a smaller gap on the original training subset, whereas FDD-Logit more closely matches the teacher's first-step distribution: onset KL is 0.335 versus 1.149. At 2B, the two objectives again yield similar accuracy gaps, but FDD-Logit achieves lower onset KL. Thus, two pairs per task suffice to learn an effective first-decision rule in this setting, while distributional supervision improves agreement with the teacher's first-step distribution. This strong low-data performance is consistent with FDD's localized supervision target: the method learns the paired first-step decision rather than a complete multi-token response trajectory.

Table 3 shows that the largest gains occur on GSM8K and BBH-seen, which have the largest initial gaps at 8B. Gains are smaller where Base Image already approaches the Text reference.

4.3 Cross-Family and Directional Generalization

On Gemma-3-12B, reverse Image-to-Text FDD increases Text accuracy from 69.70% to 80.62 ± 1.16% and reduces onset KL from 2.447 to 0.277 ± 0.087. This result extends the first-decision procedure beyond both the Qwen model family and the primary Text-to-Image direction. InternVL3.5-8B provides a contrasting result: onset KL decreases from 3.245 to 1.422 ± 0.109, but Image accuracy

Table 3: Per-benchmark AF-XML test accuracy (%) with $N = 2$, $P = 16$ per task. Learned entries are three-seed means. Parentheses give the relative change from Base Image (green) and the accuracy ratio to the Text reference (purple), both expressed as percentages.

Scale	Benchmark	Text	Base Image	Hard FDD	FDD-Logit
2B	GSM8K	81.00	64.00	**82.25** (+28.5%; 101.5% Text)	80.50 (+25.8%; 99.4% Text)
	BBH-seen	63.50	49.00	**63.00** (+28.6%; 99.2% Text)	62.58 (+27.7%; 98.6% Text)
	MMLU	57.75	52.50	**53.42** (+1.7%; 92.5% Text)	53.33 (+1.6%; 92.4% Text)
	ARC-Challenge	70.75	70.50	70.25 (−0.4%; 99.3% Text)	**70.33** (−0.2%; 99.4% Text)
	CommonsenseQA	69.50	49.50	**58.50** (+18.2%; 84.2% Text)	**58.50** (+18.2%; 84.2% Text)
8B	GSM8K	93.25	63.50	**93.08** (+46.6%; 99.8% Text)	92.58 (+45.8%; 99.3% Text)
	BBH-seen	92.50	64.25	93.00 (+44.7%; 100.5% Text)	**93.50** (+45.5%; 101.1% Text)
	MMLU	74.75	75.00	75.83 (+1.1%; 101.4% Text)	**76.50** (+2.0%; 102.3% Text)
	ARC-Challenge	82.25	81.75	**82.25** (+0.6%; 100.0% Text)	82.00 (+0.3%; 99.7% Text)
	CommonsenseQA	82.75	81.00	**81.25** (+0.3%; 98.2% Text)	**81.25** (+0.3%; 98.2% Text)

Table 4: FDD-Logit across model families and presentation directions on the held-out test sets. Direction denotes teacher → student, with T/I for Text/Image; N/P are per-task training budgets. Adapters are trained separately for each model family.

Model	Dir.	N/P	Teacher acc. (%)	Student acc. (%) Base → FDD	Δ acc.(pp)	Phase (%)	Onset KL Base → FDD
InternVL3.5-8B	T → I	2/16	73.05	72.70 → 71.02 ± 0.06	−1.68 ± 0.06	80.58 ± 0.98	3.245 → 1.422 ± 0.109
Gemma-3-12B	I → T	4/32	69.55	69.70 → 80.62 ± 1.16	**+10.92 ± 1.16**	90.27 ± 1.23	2.447 → **0.277 ± 0.087**

decreases by 1.68 ± 0.06 pp. The procedure therefore improves onset alignment in both additional settings, while the effect on final accuracy depends on the model and task.

4.4 Beyond Rendered Question Text: Structured Charts

We next test first-decision control on PlotQA, pairing each original chart image with a deterministic Text serialization of its structured annotation while keeping the question and response contract unchanged. In the validation intervention, blocking the first answer-opening token on 181 examples where it is initially top-1 reduces accuracy from 76.24% to 24.86%. The direction is opposite to the GSM8K result, demonstrating that a fixed answer-suppression rule is not universally beneficial. FDD instead learns from paired teacher decisions. On the separate 400-example held-out set, FDD-Logit reduces onset KL from 0.202 to 0.0016 ± 0.0007 and increases Image accuracy from 80.25% to 85.33 ± 0.52% (Table 5). Importantly, FDD does not pull the Image student toward the Text teacher's lower final accuracy: Base Image already outperforms the Text teacher (80.25% versus 59.5%), and FDD further improves it to 85.33% while driving first-step agreement to 100%. This pattern is consistent with transferring the teacher's onset behavior without replacing the student's Image-conditioned continuation with the teacher's full response trajectory.

Table 5: Held-out PlotQA results on 400 questions from 400 unique official-test plots. Text is a deterministic serialization of the official structured annotation; Image is the official plot. FDD uses the validation-frozen $N = 16$, $P = 128$ protocol and three frozen checkpoints.

Condition	Accuracy (%)	Δ vs. Base Image (pp)	Phase agreement (%)	Onset KL (nats)	Teacher top-1 agreement (%)	Strict compliance (%)
Text teacher	59.5	–	–	–	–	100.0
Base Image	80.25	–	92.75	0.202125	93.0	92.75
FDD-Logit	**85.33 ± 0.52**	**+5.08 ± 0.52**	**100.0 ± 0.0**	**0.001601 ± 0.000706**	**100.0 ± 0.0**	**100.0 ± 0.0**

4.5 Sensitivity to Training-Pair Selection

The $N = 2$, $P = 16$ setting achieves strong alignment on the original validation-selected training subset, but a result based on only two pairs per task may depend on which pairs are chosen. We

therefore fix the optimization seed and retrain both objectives on five independently sampled subsets, each containing two pairs from each task. This analysis isolates training-pair variation from the optimization variation reported in Table 2.

Table 6: Sensitivity to training-pair selection on Qwen3-VL-8B with $N = 2$, $P = 16$. The optimization seed is fixed; results over five independently sampled subsets are reported as mean ± SD.

Method	Original gap (pp)	Random-subset gap (pp)	Phase agreement (%)	Onset KL
Hard FDD	0.65	1.86 ± 1.42	95.57 ± 2.37	1.099 ± 0.617
FDD-Logit	1.03	$\mathbf{1.05 \pm 0.34}$	$\mathbf{97.12 \pm 0.78}$	$\mathbf{0.476 \pm 0.173}$

Across these five subsets, Hard FDD's mean absolute gap ranges from 0.80 to 3.90 pp, compared with 0.85–1.65 pp for FDD-Logit. FDD-Logit's original 1.03 pp gap also closely matches the random-subset mean of 1.05 pp. Distributional first-step supervision is therefore less sensitive to training-pair selection in this low-data comparison; the smaller original Hard FDD gap does not represent the average behavior across the sampled subsets. The stability across independently sampled two-pair subsets suggests that the low-data result is not solely an artifact of the originally selected training examples.

4.6 Paired Supervision versus Global First-Step Calibration

Table 7: Global first-step calibration versus paired onset alignment on Qwen3-VL-8B. Global Constant Bias learns one input-independent bias from the same ten training pairs; FDD-Logit uses input-conditioned first-step supervision.

Method	Image acc. (%)	Acc. gap (pp)	Phase (%)	Onset KL	Teacher top-1 agreement (%)
Base Image	73.10	12.10	79.65	1.306	72.75
Global Constant Bias	84.70	**1.00**	96.50	1.225	76.05
FDD-Logit	$\mathbf{85.17 \pm 0.06}$	1.03 ± 0.06	$\mathbf{97.35 \pm 0.13}$	$\mathbf{0.335 \pm 0.016}$	**84.92**

Global Constant Bias increases Image accuracy from 73.10% to 84.70%, recovering 96.1% of the aggregate gain achieved by Paired FDD-Logit. A shared first-step bias therefore explains much of the average Qwen accuracy recovery; the aggregate accuracy gain alone does not establish the value of example-specific pairing.

Paired supervision improves the more specific objective of matching each example's teacher: relative to Global Constant Bias, FDD reduces onset KL from 1.225 to 0.335 and increases teacher top-1 agreement from 76.05% to 84.92%. On minority-onset examples, where the teacher's first-token choice differs from the modal choice for that task, Paired FDD achieves 89.70% accuracy versus 85.57% for Global Constant Bias. These results distinguish a shared calibration shift from example-conditioned onset alignment. The accuracy gains do not imply better answer-only compliance: on the same test, strict compliance decreases from 80.95% for Base Image to 58.25 ± 0.22% for Paired FDD. Appendix H reports the teacher reference, subgroup definition, and mean-teacher and shuffled-pair controls.

4.7 Robustness to Answer-Only Prompt Paraphrases

To test whether the training procedure depends on the `<answer>` marker, we repeat the fixed $N = 2$, $P = 16$ recipe under four additional answer-only formulations. Each formulation has its own matched Text teacher and prompt-specific adapter. This one-seed validation study therefore evaluates the same training procedure under different prompts, rather than transferring a single adapter to unseen prompts.

FDD-Logit lowers onset KL under all five prompts and reduces the mean absolute Text–Image gap from 10.50 to 2.92 pp, with gap reductions under every non-XML formulation as well. Hard FDD narrows the gap under four of the five prompts but has higher onset KL than FDD-Logit throughout. Smaller gaps do not always indicate higher accuracy: under Final-label, FDD-Logit reduces the gap while Image accuracy decreases from 51.6% to 45.4%. The results support the applicability

Table 8: Prompt-paraphrase validation for Qwen3-VL-8B ($N = 2$, $P = 16$, one seed, matched teachers, prompt-specific adapters). Gap is the macro absolute Text–Image accuracy difference; KL is measured at the first generation position.

	Accuracy gap (pp) ↓			Onset KL ↓		
Prompt	Base	Hard	Logit	Base	Hard	Logit
AF-XML	13.3	**2.6** (−80.5%)	2.9 (−78.2%)	1.319	0.909 (−31.1%)	**0.293** (−77.8%)
Plain	21.7	6.9 (−68.2%)	**4.5** (−79.3%)	2.490	1.145 (−54.0%)	**0.665** (−73.3%)
No-reason	2.3	2.4 (+4.3%)	**2.1** (−8.7%)	1.215	0.677 (−44.2%)	**0.564** (−53.6%)
Final-label	6.3	3.9 (−38.1%)	**2.5** (−60.3%)	0.906	0.959 (+5.9%)	**0.328** (−63.8%)
Final-XML	8.9	2.8 (−68.5%)	**2.6** (−70.8%)	1.004	0.647 (−35.6%)	**0.274** (−72.7%)
Mean	10.50	3.72 (−64.6%)	**2.92** (−72.2%)	1.387	0.867 (−37.5%)	**0.425** (−69.4%)

of first-decision training beyond a single XML marker, while preserving the distinction between alignment and task performance.

5 Related Work

Semantically equivalent Text and Image presentations can produce substantial modality gaps and inconsistent predictions in MLLMs (Lyu et al., 2025; Zhang et al., 2024; van Sprang et al., 2026; Liu et al., 2026). Sun et al. (2026) further shows that the gap depends on the task and rendering conditions and can include a collapse of visible chain-of-thought under visual input. Existing response-level supervision shows that transferring a teacher trajectory can repair cross-modal behavior; our question is whether sequence-wide transfer is necessary when useful control can be localized to response onset. Existing control methods, including Prefix-Tuning, Prompt Tuning, Plug and Play Language Models (PPLM), and decoding-time controllers, influence multiple decoding positions through persistent conditioning or repeated logit modification (Li & Liang, 2021; Lester et al., 2021; Dathathri et al., 2020; Yang & Klein, 2021; Liu et al., 2021). FDD instead adapts one position, discards the adapted cache, and returns the continuation to the frozen base model. Classical, sequence-level, and rationale-based distillation transfer output distributions, complete sequences, or intermediate reasoning traces (Hinton et al., 2015; Kim & Rush, 2016; Hsieh et al., 2023); FDD retains supervision only at the first generation position. Appendix A provides a fuller comparison.

6 Conclusion and Limitations

Text–Image modality mismatch is not simply a uniform shortening of visible reasoning: the same problem can follow different generation trajectories across presentation formats. More broadly, in settings where response onset has causal leverage, changing how generation begins can sometimes recover useful continuations available to the frozen model without sequence-wide correction. Our single-position interventions identify the first generation decision as a high-leverage, context-dependent control point. FDD learns this decision from paired supervision and uses a clean handoff to return all subsequent generation to the frozen base model. In the primary five-benchmark Qwen setting, only 2 pairs and 16 presentations per task substantially improve Image accuracy and reduce the Text–Image gap. Additional model-family, reverse-direction, prompt-specific, and structured-chart evaluations characterize the scope and limits of this first-decision procedure.

The first decision is an operational intervention point, not a complete reasoning phase, and the results do not establish that all presentation effects are mediated at onset. Better onset alignment also need not improve final accuracy or strict instruction compliance, as the InternVL and prompt-specific results illustrate. Cross-family evaluations train separate adapters, and the prompt-paraphrase study is limited to matched-prompt validation. Finally, clean handoff requires a second prefill: localized supervision should not be interpreted as a proportional end-to-end speedup. These boundaries distinguish control of the response onset from control of the entire response.

Reproducibility Statement

Sections 2–4 describe the response contracts, interventions, training objectives, clean handoff, and main evaluation. Appendix B specifies the cohorts, decoding settings, and training protocols; Appendix L reports supervision and compute accounting. Appendix I gives the five prompt formulations and matched-teacher evaluations, Appendix J details PlotQA serialization and held-out evaluation, and Appendix H reports the calibration, pairing, and compliance controls. Upon acceptance, we will release the complete source code, training and evaluation scripts, configuration files, and stored records needed to reproduce the reported results.

Artificial intelligence (AI) Use Statement

Generative AI tools were used to assist with literature retrieval and discovery, hypothesis refinement, code development, drafting and revising portions of the manuscript, and the visual refinement of author-designed figures. All AI-assisted outputs were reviewed and verified by the authors. The authors take full responsibility for the final content of this work, including its claims, experimental results, code, and figures.

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, 2025. URL https://arxiv.org/abs/2511.21631.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300/.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1edEyBKDS.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017, Toronto, Canada, July 2023. Association for Computational Linguistics.

doi: 10.18653/v1/2023.findings-acl.507. URL https://aclanthology.org/2023.findings-acl.507/.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213. Curran Associates, Inc., 2022. doi: 10.52202/068431-1613. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. GeDi: Generative discriminator guided sequence generation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4929–4952, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.424. URL https://aclanthology.org/2021.findings-emnlp.424/.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243/.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353/.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. DExperts: Decoding-time controlled text generation with experts and anti-experts. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6691–6706, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.522. URL https://aclanthology.org/2021.acl-long.522/.

Qing'an Liu, Juntong Feng, Yuhao Wang, Xinzhe Han, Yujie Cheng, Yue Zhu, Haiwen Diao, Yunzhi Zhuge, and Huchuan Lu. Vista-bench: Do vision-language models really understand visualized text as well as pure text?, 2026. URL https://arxiv.org/abs/2602.04802.

Zhiheng Lyu, Xueguang Ma, and Wenhu Chen. Pixelworld: Towards perceiving everything as pixels. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=uY5eDN2bML.

Kaiser Sun, Xiaochuang Yuan, Hongjun Liu, Chen Zhao, Cheng Zhang, Mark Dredze, and Fan Bai. Reading, not thinking: Understanding and bridging the modality gap when text becomes pixels in multimodal llms, 2026. URL https://arxiv.org/abs/2603.09095.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.824. URL https://aclanthology.org/2023.findings-acl.824/.

Angela van Sprang, Laurens Samson, Ana Lucic, Erman Acar, Sennay Ghebreab, and Yuki M. Asano. Same content, different answers: Cross-modal inconsistency in mllms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8781–8790, June 2026.

Xuezhi Wang and Denny Zhou. Chain-of-thought reasoning without prompting. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 66383–66409. Curran Associates, Inc., 2024. doi: 10.52202/079017-2123. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/7a8e7fd295aa04eac4b470ae27f8785c-Paper-Conference.pdf.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Kevin Yang and Dan Klein. FUDGE: Controlled text generation with future discriminators. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3511–3535, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.276. URL https://aclanthology.org/2021.naacl-main.276/.

Xiang Zhang, Senyu Li, Ning Shi, Bradley Hauer, Zijun Wu, Grzegorz Kondrak, Muhammad Abdul-Mageed, and Laks V. S. Lakshmanan. Cross-modal consistency in multimodal large language models, 2024. URL https://arxiv.org/abs/2411.09273.

A Extended Related Work

Cross-presentation modality gaps. Semantically equivalent Text and Image inputs need not elicit equivalent responses. PixelWorld studies language, tables, code, and diagrams in a common pixel space and finds particularly large losses on reasoning-intensive tasks (Lyu et al., 2025). Cross-modal consistency benchmarks examine answer agreement across presentations (Zhang et al., 2024; van Sprang et al., 2026), while VISTA-Bench compares pure-text and visualized-text questions under controlled rendering conditions (Liu et al., 2026). Closest to our setting, Sun et al. (2026) documents task- and rendering-dependent changes in visible reasoning and uses Text-side reasoning traces to supervise Image responses. FDD asks a complementary question: how much of the resulting trajectory divergence can be redirected by changing only the first generation decision?

Reasoning elicitation and generation control. Chain-of-thought prompting elicits multi-step solutions through demonstrations or explicit reasoning instructions (Wei et al., 2022; Kojima et al., 2022). CoT decoding explores alternative initial tokens to expose reasoning trajectories that greedy decoding may miss (Wang & Zhou, 2024). Other controllers act throughout generation: continuous prompts and prefixes provide persistent conditioning (Li & Liang, 2021; Lester et al., 2021), PPLM modifies hidden activations, and GeDi, FUDGE, and DExperts reweight next-token distributions (Dathathri et al., 2020; Krause et al., 2021; Yang & Klein, 2021; Liu et al., 2021). FDD instead learns an input-conditioned first-step decision from paired presentations, applies the adapter only at the first decoding step, and returns every later position to the frozen model. The intervention unit is the first generation decision; the emitted first token is its realized outcome, not a binary representation of reasoning or a universally preferred response opening.

Parameter-efficient adaptation and distillation. LoRA ordinarily remains active during the complete response (Hu et al., 2022). Standard distillation matches teacher distributions, sequence-level distillation uses complete teacher outputs, and rationale distillation transfers intermediate explanations as well as labels (Hinton et al., 2015; Kim & Rush, 2016; Hsieh et al., 2023). Cross-presentation response distillation likewise supervises Text-side reasoning traces on Image inputs (Sun et al., 2026). Hard FDD and FDD-Logit retain only the teacher's realized top-1 token or full vocabulary distribution at the first generation step. Clean handoff discards the adapted key–value cache and freshly prefills the frozen base model before continuation. Table 16 in Appendix F compares the two design choices under matched full-data training: Full-Trace LoRA and One-Decision Full-LoRA differ in supervision scope, while One-Decision Full-LoRA and Hard FDD differ in whether adaptation persists during continuation.

B Experimental Protocol and Evaluation

B.1 Tasks and Paired Presentations

The primary evaluation uses Qwen3-VL-2B-Instruct and Qwen3-VL-8B-Instruct on GSM8K, BBH-seen, MMLU, ARC-Challenge, and CommonsenseQA. Each benchmark has 800 training, 200 validation, and 400 test examples. Smaller diagnostic and pilot cohorts are identified separately below. A Text–Image pair preserves the semantic problem and response contract while changing its presentation. The primary renderer uses a 1280×720 canvas, NotoSansMath, and fixed regions for the question and response instruction. In the image-delivered instruction controls, the question pixels remain unchanged when the response instruction changes.

Each teacher uses the student's frozen base checkpoint and the same response contract, with the paired presentation as input. Qwen and InternVL use Text-to-Image supervision, whereas Gemma uses Image-to-Text supervision. Each model family has its own paired teacher targets and independently trained adapters; the cross-family experiments replicate the procedure rather than transfer a shared adapter. PlotQA pairs original chart images with deterministic serializations of their structured annotations, as detailed in Appendix J.

B.2 Response Contracts

The main evaluation and diagnostic experiments use the following response contracts (instructions specifying the requested output format):

> **AF-XML:** `Give only the final answer inside <answer>...</answer>.`
> **RN-XML:** `End your response with <answer>...</answer>.`
> **RR-text:** `Show the reasoning and end with Final answer: ...`

AF-XML retains the name "Answer-first XML," although its instruction requires an answer-only response. RN-XML (reasoning-neutral XML) permits reasoning before the final answer, and RR-text (reasoning-required text) explicitly requests it. AF-XML is used for the main learned-method comparisons. Appendix I gives the additional answer-only formulations and their evaluation rules. The smaller diagnostic cohorts use their own prompt wrappers and instruction-delivery conditions and are analyzed separately from the main benchmark evaluation.

B.3 Training and Inference

All Qwen adapters use rank-8 LoRA on the language model's $q/k/v/o$ attention projections, with the base language model and vision encoder frozen. Full-data training uses 800 pairs and 2,400 presentations per task: 4,000 unique pairs, three epochs, and 1,500 optimizer updates across the five tasks. The main low-data setting uses $N = 2$ pairs and $P = 16$ total presentations per task. Each pair appears eight times, giving ten unique pairs, 80 presentations, and ten optimizer updates. Hard FDD and FDD-Logit share the same training pairs and balanced presentation schedule.

The low-data budget and training subset are chosen using validation data before test evaluation. For the primary Qwen setting, checkpoint selection uses validation Image-side negative log-likelihood on the teacher's first-token target. The Qwen configuration selected at 8B is applied at 2B without retuning. InternVL3.5-8B uses $N = 2$, $P = 16$; Gemma-3-12B uses $N = 4$, $P = 32$ in the reverse direction. Each cross-family setting is evaluated with three training seeds and 400 test examples per task. The InternVL budget is retained after a higher-exposure validation variant exhibits seed-sensitive accuracy.

Unless otherwise stated, generation uses temperature 0.1, repetition penalty 1.0, a 1024-token cap, and shared per-example decoding seeds across paired presentations. At inference, FDD uses the adapted model for only the first decoding step, discards the adapted key–value cache, and resumes from a fresh prefill of the frozen base model. The reverse-intervention probes use greedy decoding. The forward OSD comparison uses different generation caps for the base model and intervention (Appendix E.2); unseen-task caps are specified in Table 30.

B.4 Metrics and Statistical Reporting

Accuracy is scored using a fixed numeric or multiple-choice answer parser. The absolute Text–Image gap is computed within each benchmark and training seed before equal-weight averaging across benchmarks, as defined in Section 4. Accuracy and agreement rates are percentages; accuracy differences and gaps are in percentage points (pp). Averaging absolute gaps is not equivalent to taking the absolute difference between mean accuracies. Correctness agreement measures whether paired responses have the same correct-versus-incorrect outcome, while phase agreement is the fraction of pairs with matching answer-first or reasoning-first labels. Here, *phase* describes the visible response opening, not an internal reasoning state.

Answer entry requires completion of the designated opening, rather than generation of its first token alone. Visible reasoning is the generated text before the final-answer segment; its length describes observable output, not hidden computation or reasoning quality. Onset KL is the teacher-to-student Kullback–Leibler divergence at the first generation position, measured in nats from the full-vocabulary distributions before sampling-temperature scaling. Teacher top-1 agreement (exact first-token agreement) measures whether the teacher and student have the same highest-probability first token. It is distinct from both agreement between sampled tokens and phase agreement. For

AF-XML, strict compliance separately assesses the answer-only requirement, tag closure, and absence of extraneous text. Correct-and-compliant accuracy is the fraction of all evaluated responses that are both correct and strictly compliant. Conditional reasoning-length mean absolute error (MAE) measures the absolute Text–Image difference in visible reasoning-token counts only among pairs in which both responses are reasoning-first. These metrics measure different aspects of the response and need not improve together.

Learned-method results are reported as mean $\pm$ SD over three training seeds unless stated otherwise. These runs measure optimization variation for a fixed training subset. The training-pair sensitivity study instead fixes the training seed and reports variation over five independently sampled subsets. Prompt-paraphrase results and the preliminary budget screens use one training seed and validation data only. Intervention confidence intervals describe paired changes on their respective evaluation cohorts; validation and test cohorts are not pooled.

C Clean Handoff and the Score Decomposition

C.1 A Fixed Continuation Distribution

Consider the Text-to-Image setting. Fix the problem, Image presentation, response contract, generation cap, and decoding rule, and suppress these shared conditioning variables in the notation. Let $\pi_\phi(u)$ be the adapted first-token decoding distribution, $\pi_0(u)$ its unadapted counterpart, and $K_{\theta_0}(Y_{1:} \mid u)$ the frozen continuation distribution after token u. These are decoding distributions, including temperature and other decoding transformations; greedy decoding is the degenerate case. They are distinct from the unscaled distributions used to report onset KL. Under clean handoff, Equation (5) changes only the onset distribution and leaves the conditional continuation unchanged.

Fresh prefill is essential to this separation. Merely disabling LoRA while reusing its cache can retain adapted hidden states and hence alter the continuation distribution. Discarding the cache leaves the emitted first token as the adapter's only input to the frozen continuation.

C.2 Consequences for Output Scores

For any output score $S(Y)$ with finite expectation, define the frozen Image-side continuation value

$$V_{\mathrm{I}}^{S}(u) = \mathbb{E}_{Y_{1:}\sim K_{\theta_0}(\cdot|u)}[S(u, Y_{1:})]. \tag{6}$$

Conditioning on the first token gives

$$\mathbb{E}_{\mathrm{FDD}}[S] - \mathbb{E}_{\mathrm{Base}}[S] = \sum_{u\in\mathcal{V}} \big(\pi_\phi(u) - \pi_0(u)\big) V_{\mathrm{I}}^{S}(u). \tag{7}$$

The identity expresses the score change as a reweighting of frozen conditional continuations. It does not require a binary reasoning policy, nor does it imply that the first token encodes the answer or mediates every effect of presentation.

Let $C(Y)$ be the indicator of parsed correctness and $A(Y)$ the indicator of strict instruction compliance. Setting $S = C$, $S = A$, or $S = CA$ gives the corresponding change in accuracy, strict compliance, or correct-and-compliant accuracy. Since V_{I}^{C} and V_{I}^{A} can differ, the same onset shift can improve correctness while reducing compliance. Matching the Text teacher's onset distribution need not maximize either Image-side score, and agreement on a coarse phase label need not align the remaining response. The factorization follows from the clean-handoff design; the benefit of a learned onset decision is an empirical property of the model, input, and task. The same argument applies to Image-to-Text FDD after exchanging the teacher and student presentations.

D Response Onset and Instruction Sensitivity

D.1 Paired Variation in Visible Reasoning

Figure 1 compares frozen Text and Image responses on 400 paired Qwen3-VL-8B GSM8K test problems under AF-XML. The Image-minus-Text difference averages −48.66 visible reasoning tokens. Image responses contain fewer visible reasoning tokens in 255 pairs, the same number in 16, and more in 129; the magnitude distributions are normalized separately within the shorter and longer subsets. Thus, the average shortening coexists with substantial variation in both directions. Figure 2 illustrates the corresponding qualitative differences.

D.2 Response Contracts and Instruction Delivery

Table 9 compares AF-XML and RR-text on a separate set of 200 paired GSM8K problems, varying whether the response instruction is supplied as text or rendered into the image. With text-delivered instructions, the accuracy gap is 30.0 pp under AF-XML and 2.0 pp under RR-text. The change is accompanied by more visible reasoning and higher Image accuracy. It shows that the presentation gap depends on the requested response behavior, rather than being fixed by the semantic problem alone.

Table 9: Response-contract and instruction-delivery controls on 200 paired GSM8K problems per condition. Entry is the rate of completing the AF-XML answer opening at response onset. Question pixels are unchanged between the two image-delivered instruction conditions.

Question form	Instruction	Contract	Acc. (%)	Reason. tok.	Entry (%)
Text	Text	AF-XML	88.0	143.7	2.0
Text	Text	RR-text	90.0	220.2	–
Image	Text	AF-XML	58.0	87.4	60.5
Image	Text	RR-text	88.0	254.7	–
Image	Image	AF-XML	66.5	108.3	46.5
Image	Image	RR-text	86.5	264.5	–

The same pattern appears when the instruction is delivered in the image: changing from AF-XML to RR-text raises accuracy from 66.5% to 86.5% while leaving the question pixels unchanged. Because RR-text changes the requested behavior, these comparisons demonstrate instruction sensitivity rather than an improvement under an unchanged answer-only requirement.

D.3 Matched Answer-Only and Reasoning-Neutral Controls

Tables 10 and 11 compare AF-XML with RN-XML using the same frozen model, paired problems, and question-only renderer, with 100 pairs per benchmark. Under AF-XML, both presentations frequently enter the answer immediately, and the GSM8K gap is smaller than in the 200-example instruction-delivery experiment. These cohorts differ in their prompt wrappers and sampled problems; the contract label alone therefore does not determine the size of the observed gap.

Table 10: Matched AF-XML control with 100 pairs per benchmark. T/I denotes Text/Image. Entry is the rate of completing the answer opening at response onset; differences are Image minus Text.

Benchmark	Acc. T/I (%)	Δ acc. (pp)	Reason. tok. T/I	Entry T/I (%)	Δ entry (pp)
GSM8K	31 / 27	−4	4.1 / 0.0	95 / 100	+5
MMLU	72 / 70	−2	0.2 / 0.2	81 / 80	−1
ARC-Challenge	90 / 92	+2	0.3 / 0.5	66 / 55	−11
GPQA Diamond	36 / 32	−4	0.1 / 0.1	87 / 85	−2

Under RN-XML, none of the Image responses in these four cohorts begins with the designated answer opening, yet accuracy differences remain. Immediate answer entry is therefore one presentation-dependent behavior, not a complete explanation of all modality gaps. Longer visible responses likewise do not by themselves establish better reasoning.

Table 11: Matched RN-XML control with 100 pairs per benchmark. This instruction permits reasoning before the final answer. T/I denotes Text/Image; differences are Image minus Text. Accuracies are percentages, and reasoning length counts visible tokens before the answer.

Benchmark	Text acc.	Image acc.	Δ acc. (pp)	Reason. tok. T/I	Δ tokens
GSM8K	89.0	88.0	−1.0	189.3 / 215.8	+26.5
MMLU	81.0	75.0	−6.0	202.1 / 259.7	+57.6
ARC-Challenge	91.0	89.0	−2.0	137.7 / 201.3	+63.6
GPQA Diamond	34.0	29.0	−5.0	1536.3 / 1581.8	+45.5

D.4 The First Decision, First Token, and Completed Opening

The AF-XML answer opening spans three tokenizer tokens. At $t = 0$, the intervention changes only the score of the first tokenizer token; completion of the full opening remains a later model-generated behavior. The answer, reasoning-first, and full-response XML openings share this initial token but complete different sequences. On the 200-example Image cohort with text-delivered AF-XML instructions, the designated first token is top-1 at onset in 61.0% of cases, while the complete answer opening occurs in 60.5%. The first decision, its emitted token, and the completed opening are therefore closely related in this setting but are not identical.

Table 12: Opening-specific response behavior on 200 GSM8K Image examples with text-delivered instructions. Opening rates refer to each condition's complete opening and are not a common phase probability across contracts.

Contract	Acc. (%)	Reason. tok.	Observed opening
AF-XML	58.0	87.4	Answer: 60.5%
Reasoning first	88.5	224.1	Reasoning: 100.0%
Full response	85.5	218.8	Response: 6.5%
RR-text	88.0	254.7	No XML answer opening

Table 12 further distinguishes a shared first token from its context-dependent continuation. The "Reasoning first" and "Full response" conditions use their own prompt wrappers and complete openings. Their opening rates are not directly comparable phase probabilities across contracts. Completing the answer opening at response onset does not establish tag closure, the absence of extra text, or answer correctness. Appendix H.3 examines these distinctions on the main five-benchmark test set, separately from the diagnostic cohorts used here.

E First-Decision Interventions and Alternative Controls

E.1 Intervention Definitions

Let m_c denote the first tokenizer token of the answer opening designated by response contract c. Given raw logits z at the first generation position ($t = 0$), One-Shot Deflection (OSD) applies

$$z'_v = \begin{cases} -\infty, & v = m_c \text{ and } m_c = \arg\max_u z_u, \\ z_v, & \text{otherwise.} \end{cases} \tag{8}$$

The mask is removed immediately after the first decoding step. OSD neither blocks later answer openings nor enforces a minimum reasoning length.

Reverse induction instead raises the designated token's logit to 10^{-4} above the current maximum, leaving all other logits unchanged. The matched-rule control promotes an eligible non-answer token in the same way. Both interventions act only at $t = 0$. The control matches the location and promotion rule, but not necessarily the target token's original probability or the perturbation norm.

E.2 Forward Deflection and Reverse Induction

On the held-out 300-example Image/AF-XML cohort, OSD increases accuracy from 41.7% to 93.0%, rescuing 155 previously incorrect examples and harming one. The paired gain is 51.3 pp, with a 95% confidence interval of $[45.7, 57.0]$ pp. The Base Image and OSD runs use different generation caps: 1024 tokens with two limit hits for Base Image, and 512 tokens with eight hits for OSD. This comparison therefore does not hold the generation cap fixed.

Among triggered examples, the designated token's mean raw probability is 0.924 at $t = 0$ and 9.5×10^{-11} at $t = 1$ after deflection. These values are conditioned on different prefixes and syntactic positions. This probability trace describes the model's continuation after a single-position intervention; it does not isolate an additional causal mechanism.

Table 13: Reverse first-position interventions on naturally non-answer-start GSM8K responses. Base, Opening, and Non-answer denote unmodified decoding, promotion of the first answer-opening token, and promotion of a non-answer token. Accuracies are percentages; intervals describe paired changes under answer-opening promotion.

Input / contract	n	Base	Opening	Non-answer	Change (pp; 95% CI)
Text / AF-XML	194	94.8	28.9	96.4	$-66.0\ [-72.7, -59.3]$
Image / RR-text	200	93.0	33.5	95.0	$-59.5\ [-66.0, -52.5]$

In the reverse probes, promoting the first answer-opening token causes every intervened response to complete the designated opening and omit visible reasoning. Only the logits at $t = 0$ are modified; the frozen model generates the remaining opening and response. Accuracy falls sharply in both the Text/AF-XML and Image/RR-text cohorts, whereas non-answer promotion changes accuracy by only $+1.5$ pp and $+2.0$ pp. The RR-text probe uses its own contract-specific answer opening rather than the AF-XML sequence. Paired changes are calculated before rounding and may differ slightly from subtraction of the displayed accuracies.

E.3 Instruction, Prefix, and Multi-Position Controls

A separate 100-pair GSM8K diagnostic compares visible outputs under the same answer-only instruction and decoding configuration (Table 14). The Image responses are substantially shorter and less accurate. No Image response reaches the generation limit. The multi-step rate is a surface-form proxy: it detects at least two arithmetic patterns or at least two reasoning markers in the visible reasoning text, without verifying their validity. Because first-step logits and end-of-sequence token identities are unavailable for this cohort, it provides an observational comparison rather than a first-position intervention estimate.

Table 14: Paired GSM8K response behavior in a separate 100-example diagnostic cohort under AF-XML. "Short" denotes at most 16 generated tokens; all length measures describe visible output.

Presentation	Acc. (%)	Multi-step (%)	Reason. tok.	Output tok.	Short (%)
Text	96.0	78.0	163.7	172.8	2.0
Image	47.0	10.0	20.3	29.7	89.0

Table 15 evaluates broader elicitation strategies on the same 100 Image-input examples. The reasoning-instruction condition adds a generic request for reasoning; the prefix condition prefills the assistant response with `<think>`. Multi-position blocking suppresses both the end-of-sequence token and the first answer-opening token for the first k generation positions. Here k denotes the blocked-prefix length, reserving N for the number of training pairs. Native Thinking uses a separate checkpoint, not a different mode of the fixed Instruct model.

On a separate, non-overlapping 200-question cohort, accuracies are 39.0%, 95.0%, 93.0%, and 96.5% for unmodified decoding, the reasoning instruction, $k = 16$ blocking, and Native Thinking, respectively; the corresponding multi-step rates are 10.0%, 87.0%, 81.5%, and 99.0%. Several broader interventions therefore recover strong performance in this setting. Unlike FDD, they change

Table 15: Instruction, prefix, and multi-position controls on the same 100 Image-input GSM8K examples. Limit-hit and parse-failure rates are percentages; the final column reports them in that order. The blocked-prefix length k is distinct from the training-pair budget N. Native Thinking uses a separate checkpoint.

Condition	Acc. (%)	Multi-step (%)	Reason. tok.	Limit / parse fail. (%)
Base Image	47.0	10.0	20.3	0 / 0
Reasoning instruction	96.0	90.0	245.0	3 / 2
<think> prefill	96.0	86.0	243.7	2 / 1
Block first $k = 16$	94.0	82.0	162.5	5 / 5
Block first $k = 32$	93.0	82.0	163.2	6 / 6
Block first $k = 64$	93.0	84.0	167.8	6 / 6
Native Thinking	96.0	99.0	629.5	0 / 0

the requested response behavior, impose a longer prefix intervention, or use a different checkpoint rather than learning one decision followed by a clean handoff.

F Full-Data Comparisons and Continuation Behavior

F.1 Matched Supervision and Deployment Comparisons

The full-data comparison complements the low-data evaluation in Table 2. Every learned method in Table 16 uses 800 pairs and 2,400 presentations per task: 4,000 unique pairs, three epochs, and 1,500 optimizer updates in total across the five benchmarks. Hard FDD and One-Decision Full-LoRA share the same one-step hard-target objective but differ in whether the adapter remains active during continuation. Full-Trace LoRA uses the same adapter family with supervision over the complete response.

Table 16: Matched full-data comparisons with 800 training pairs and 2,400 presentations per task. Results are mean ± standard deviation (SD) over three training seeds. Accuracy gaps are in percentage points, agreement rates in percent, and conditional reasoning-length MAE in tokens, evaluated only on pairs in which both responses are reasoning-first.

Scale	Method	Acc. gap ↓	Corr. agree. ↑	Phase agreement ↑	Reasoning-length MAE ↓
2B	Full-Trace LoRA	**3.35 ± 0.73**	**81.48 ± 0.98**	98.40 ± 0.08	**97.40 ± 4.77**
	One-Decision Full-LoRA	6.27 ± 0.69	79.67 ± 0.39	**98.92 ± 0.02**	115.10 ± 5.35
	Hard FDD	3.72 ± 0.09	78.82 ± 0.12	**98.92 ± 0.02**	128.77 ± 6.47
8B	Full-Trace LoRA	3.07 ± 0.20	90.30 ± 0.36	95.85 ± 0.46	**45.38 ± 1.47**
	One-Decision Full-LoRA	1.05 ± 0.15	90.35 ± 0.35	**97.73 ± 0.24**	59.29 ± 4.57
	Hard FDD	**0.60 ± 0.08**	**90.87 ± 0.17**	**97.73 ± 0.24**	81.86 ± 3.85

Under this matched training budget, Hard FDD has a smaller accuracy gap than One-Decision Full-LoRA at both scales while leaving later generation to the frozen model. Full-Trace LoRA achieves the lowest conditional reasoning-length MAE, consistent with directly supervising the continuation. Accuracy-gap reduction and matching the teacher's visible reasoning length therefore favor different supervision and deployment choices.

F.2 Per-Benchmark Accuracy Gaps

Table 17 decomposes the full-data Hard FDD gap by benchmark. Each learned entry averages the absolute gap calculated within each training seed; it should not be reconstructed by taking an absolute difference between rounded mean accuracies. These results use full-data adapters rather than the $N = 2$, $P = 16$ adapters in the main low-data evaluation.

A smaller distance to the Text reference need not imply higher Image accuracy. In the 8B MMLU comparison, Image accuracy rises from 75.00% to 75.50%, moving farther from the 74.75% Text

Table 17: Per-benchmark absolute Text–Image accuracy gaps under full-data training, in percentage points. Hard FDD uses 800 pairs and 2,400 presentations per task. Learned entries average seed-wise absolute gaps over three runs.

	2B		8B	
Benchmark	Base Image gap	Full-data Hard FDD gap	Base Image gap	Full-data Hard FDD gap
GSM8K	17.00	**0.67**	29.75	**0.17**
BBH-seen	14.50	**1.58**	28.25	**0.25**
MMLU	5.25	**5.08**	**0.25**	0.75
ARC-Challenge	**0.25**	**0.25**	0.50	**0.33**
CommonsenseQA	20.00	**11.00**	1.75	**1.50**

reference. Conversely, the remaining 11.00 pp gap on 2B CommonsenseQA coexists with perfect measured phase agreement. These cases illustrate why accuracy, accuracy gap, and phase agreement are reported separately.

F.3 Preservation of Baseline-Correct Answers

Table 18 measures the fraction of BBH-seen examples answered correctly by Base Image that remain correct after adaptation. Full-data Hard FDD has the highest preservation at 2B, whereas Full-Trace LoRA has the highest at 8B. An intervention may rescue some previously incorrect examples while harming previously correct ones; conditional preservation complements the overall accuracy and gap measures.

Table 18: Correctness preservation on BBH-seen examples answered correctly by Base Image under the matched full-data protocol. Results are mean ± SD over three training seeds. Preservation is conditional on baseline correctness, not overall accuracy.

Method	2B (%)	8B (%)
Full-Trace LoRA	81.63 ± 0.72	**97.02 ± 0.18**
One-Decision Full-LoRA	73.98 ± 4.39	93.64 ± 1.02
Full-data Hard FDD	**90.82 ± 0.42**	96.50 ± 0.32

G Training Budget and Pair-Selection Sensitivity

G.1 Unique Training Pairs and Repeated Exposure

The Qwen3-VL-8B validation analysis separates the number of unique training pairs N from the total presentation budget P, with both budgets counted per task. Table 19 varies N at fixed $P = 800$ with one training seed and a pilot subset of 40 validation examples per benchmark. Useful accuracy is obtained with very few unique pairs, while distributional supervision yields lower onset KL throughout this screen. These validation results guide the choice of training budget rather than estimate a sample-complexity law.

Table 19: Qwen3-VL-8B validation pilot with one training seed and $P = 800$ presentations per task, averaged equally over 40 examples from each of the five benchmarks. The number of unique training pairs N varies while total exposure is fixed. Accuracy is in percent and onset KL in nats.

	Accuracy (%)		Onset KL (nats)	
N	Hard FDD	FDD-Logit	Hard FDD	FDD-Logit
1	82.5	**86.5**	2.075	**0.254**
2	**86.5**	85.5	1.750	**0.320**
8	**85.5**	85.0	0.878	**0.157**
800	**86.5**	85.0	0.235	**0.097**

Pair composition also matters. Under the same $P = 800$ budget and a fixed training seed, three predetermined orderings of the training pool produce the widest accuracy range at $N = 1$: 5.75 pp for Hard FDD and 1.50 pp for FDD-Logit on the GSM8K and BBH-seen continuation evaluation. The ranges narrow at $N = 2$ and $N = 8$.

Teacher uncertainty provides an additional descriptive comparison. Mean validation entropy is 0.256 nats on GSM8K and 0.419 on BBH-seen, compared with 0.009 on ARC-Challenge and 0.0004 on CommonsenseQA. At $N = 2$, teacher entropy correlates with the Hard-FDD-minus-FDD-Logit KL difference at $\rho = 0.445$ on GSM8K and $\rho = 0.759$ on BBH-seen. These exploratory correlations concern the higher-exposure validation setting; they were not re-estimated under the final $P = 16$ schedule.

G.2 Selection of the Balanced Low-Data Configuration

In the one-seed Qwen3-VL-8B validation screen, the balanced $N = 2$, $P = 16$ schedule presents each of the ten selected pairs eight times. Increasing P from 8 to 16 raises phase agreement from 90.3/86.2% to 97.4/97.7% for Hard FDD/FDD-Logit, while accuracy changes from 86.0/86.3% to 85.8/85.5%. A further increase to $P = 24$ does not improve all metrics consistently. The $P = 16$ choice therefore balances exposure and onset alignment; it is not evidence of a universal saturation threshold.

Table 20: Validation results for the balanced $N = 2$, $P = 16$ configuration. Results are mean ± SD over three training seeds on a fixed training subset. Accuracy and agreement rates are percentages; onset KL is in nats.

Scale	Objective	Teacher top-1 agreement	Onset KL	Accuracy	Phase agreement
2B	Hard FDD	78.70 ± 0.10	1.103 ± 0.031	**66.93 ± 0.12**	98.33 ± 0.21
	FDD-Logit	**80.47 ± 0.25**	**0.572 ± 0.007**	66.23 ± 0.06	**98.50 ± 0.10**
8B	Hard FDD	79.77 ± 0.47	0.955 ± 0.038	**85.80 ± 0.00**	97.2 ± 0.2
	FDD-Logit	**85.43 ± 0.42**	**0.291 ± 0.011**	85.57 ± 0.06	**97.6 ± 0.1**

Table 20 reports the subsequent three-seed evaluation on all 200 validation examples per benchmark, using the fixed training subset. The FDD-Logit-minus-Hard-FDD accuracy difference is -0.233 pp at 8B, with a paired-bootstrap 95% confidence interval of $[-0.567, 0.000]$ pp, and -0.700 pp at 2B, with an interval of $[-1.467, +0.033]$ pp. These intervals concern the fixed validation examples and trained models, not alternative training subsets. The budget, subset, and objectives are fixed before the held-out test and are not revised using its outcomes.

G.3 Sensitivity to Independently Sampled Training Pairs

To isolate pair-selection variation, the Qwen3-VL-8B analysis fixes the training seed and draws five independent subsets, each containing two pairs from each benchmark. Both objectives use the same balanced $P = 16$ schedule and the main test protocol.

Table 21: Training-pair sensitivity at $N = 2$, $P = 16$ on Qwen3-VL-8B. Results are mean ± SD over five independently sampled subsets, with one fixed training seed and the same test examples. Onset KL is in nats.

Method	Gap (pp)	Image acc. (%)	Corr. agree. (%)	Phase agreement (%)	Onset KL
Hard FDD	1.86 ± 1.42	83.94 ± 1.60	90.10 ± 1.20	95.57 ± 2.37	1.099 ± 0.617
FDD-Logit	**1.05 ± 0.34**	**85.01 ± 0.43**	**90.81 ± 0.28**	**97.12 ± 0.78**	**0.476 ± 0.173**

The absolute gap ranges from 0.80 to 3.90 pp for Hard FDD and from 0.85 to 1.65 pp for FDD-Logit. On the original validation-selected subset, the corresponding gaps are 0.65 pp and 1.03 pp. The original Hard FDD subset thus performs better on this metric than all five newly sampled subsets, whereas FDD-Logit's original result is close to its random-subset mean of 1.05 pp. FDD-Logit is less

sensitive to training-pair selection in this comparison, even though Hard FDD has the smaller gap on the original subset.

G.4 One-Pair and Cross-Task Pilots

One-seed pilot experiments on a fixed GSM8K-derived split further illustrate the limits of extremely sparse supervision. Across ten one-pair subsets, Hard FDD reaches a mean accuracy of 86.9% with a standard deviation of 12.4 pp, spanning 54%–93%. A leave-one-task-out pilot evaluated on GSM8K reaches 38% accuracy and begins with the answer opening on every example. These pilots show that one arbitrary pair, or a few pairs from other tasks, need not suffice. They are separate from the five-subset analysis above and from the main balanced setting, which includes two pairs from every training benchmark.

H Pairing Specificity, Calibration, and Compliance

H.1 Global Calibration and Pairing Controls

These post-hoc controls use the Qwen3-VL-8B five-benchmark test split used in the main evaluation. Their definitions and hyperparameters are fixed before their own test inference; no main FDD model is selected or retrained using the analysis. Task-level Constant Bias fits an input-independent first-step shift separately for each task from its two Text teachers. Global Constant Bias fits one shared full-vocabulary shift from all ten training pairs and does not use task identity at inference. Mean-teacher FDD replaces each paired target with the mean of the two training-teacher distributions for that task. Shuffled-pair FDD swaps those two distributions within each task. Both controls use the distributional objective of FDD-Logit. These learned controls and Paired FDD-Logit use the same low-data clean-handoff protocol and three training seeds.

Table 22: Calibration and pairing controls on the Qwen3-VL-8B five-task test. Bias methods are deterministic references; learned methods are mean ± SD over three training seeds. Accuracy, agreement, and compliance rates are percentages; onset KL is in nats.

Method	Image acc.	Acc. gap (pp)	Correctness agreement (%)	Phase agreement (%)	Onset KL	Strict compliance (%)	Correct & compliant (%)
Base Image	73.10	12.10	80.00	79.65	1.306	**80.95**	**55.90**
Task-level Constant Bias	84.85	**0.85**	90.45	96.40	1.306	59.90	47.45
Global Constant Bias	84.70	1.00	90.20	96.50	1.225	59.10	47.35
Mean-teacher FDD	80.83 ± 0.38	4.57 ± 0.38	87.37 ± 0.40	92.15 ± 0.41	0.628 ± 0.011	67.12 ± 0.51	50.50 ± 0.10
Shuffled-pair FDD	76.62 ± 0.94	8.55 ± 0.82	83.52 ± 0.85	91.72 ± 1.23	1.149 ± 0.010	67.38 ± 2.08	49.85 ± 0.83
FDD-Logit	**85.17 ± 0.06**	1.03 ± 0.06	**90.97 ± 0.06**	**97.35 ± 0.13**	**0.335 ± 0.016**	58.25 ± 0.22	46.82 ± 0.20

Global Constant Bias reaches 84.70% Image accuracy and recovers 96.1% of Paired FDD-Logit's aggregate gain, nearly matching the task-specific bias control. A shared calibration shift therefore explains much of the average accuracy recovery in this setting. Paired FDD-Logit nevertheless has the lowest onset KL and highest Image accuracy among the listed methods, while mean-teacher and shuffled-pair training perform less well on both measures. Aggregate recovery and example-specific teacher alignment are related but distinct objectives.

H.2 Example-Specific and Minority-Onset Alignment

Table 23 compares each student's onset distribution with its matched teacher, a mismatched teacher from the same task, and the mean of the two training-teacher distributions for that task. Paired FDD-Logit is closer to the matched teacher than either bias baseline, while its divergence from the task-mean teacher is similar to that of Task-level Constant Bias. Relative to Task-level Constant Bias, the paired reduction in matched-teacher KL is 0.972 nats, with Cohen's $d_z = 0.394$ and a Holm-adjusted $p = 2.35 \times 10^{-8}$.

Table 23: Example-specific onset alignment for constant-bias controls and Paired FDD-Logit. Matched-teacher KL compares each student with its own paired teacher. All KL values use the reference-to-student direction; dashes denote unreported comparisons. Minority-onset metrics are equal-task averages over the four tasks with non-empty subgroups; KL is in nats.

Metric	Task-level Constant Bias	Global Constant Bias	FDD-Logit
Matched-teacher KL	1.306	1.225	**0.335**
Mismatched same-task teacher KL	4.718	—	3.778
Task-mean teacher KL	5.107	—	5.134
Exact teacher top-1 agreement (%)	76.55	76.05	**84.92**
Minority-onset teacher top-1 agreement (%)	17.46	39.76	**58.34**
Minority-onset KL	3.930	1.821	**0.890**
Minority-onset accuracy (%)	77.51	85.57	**89.70**

Minority-onset examples are those whose teacher's top-1 token at the first generation step differs from the modal teacher top-1 token for that task. The subgroup contains 93 examples: 9 GSM8K, 31 MMLU, 8 ARC-Challenge, and 45 BBH-seen. Its metrics are equal-task averages over these four non-empty tasks. Paired FDD-Logit reaches 89.70% accuracy, compared with 85.57% for Global Constant Bias, and also improves teacher top-1 agreement and onset KL. The shared bias narrows but does not eliminate the difference on these atypical examples. The subgroup comparison reuses existing per-example results without additional training or inference.

H.3 Correctness Does Not Imply Answer-Only Compliance

Phase agreement describes how a response begins; it does not enforce the complete AF-XML instruction. As Table 22 shows, Paired FDD-Logit reaches 97.35% phase agreement but only 58.25% strict compliance. Table 24 places this result alongside the teacher's own behavior.

Table 24: Teacher and student correctness–compliance outcomes on the same five-task test. Every rate uses all evaluated examples as its denominator. Correct-and-non-compliant accuracy equals overall accuracy minus correct-and-compliant accuracy. Learned entries are three-run means.

Method	Accuracy (%)	Strict compliance (%)	Correct & compliant (%)	Correct & non-compliant (%)
Text teacher	85.10	60.70	48.25	36.85
Base Image	73.10	**80.95**	**55.90**	17.20
Task-level Constant Bias	84.85	59.90	47.45	37.40
FDD-Logit	**85.17**	58.25	46.82	38.35

The Text teacher is less compliant than Base Image (60.70% versus 80.95%). Paired FDD-Logit improves parsed accuracy but decreases correct-and-compliant accuracy from 55.90% to 46.82%. This pattern is consistent with learning an onset preference from a teacher that does not reliably follow the answer-only instruction. It also illustrates Equation (7) in Appendix C: the same change in onset can improve correctness without improving compliance.

I Robustness to Answer-Only Prompt Paraphrases

I.1 Prompts and Matched Training

This validation-only study uses Qwen3-VL-8B-Instruct, one training seed, and the same $N = 2$, $P = 16$ training subset and balanced schedule as the primary low-data configuration. Each formulation has a Text teacher evaluated under that same prompt and a prompt-specific Hard FDD or FDD-Logit adapter. The four additional formulations use fresh adapter initializations; AF-XML uses the adapter trained under the same low-data protocol. Evaluation covers 200 validation examples from each of the five tasks, without test-set use.

The five answer-only instructions are:

AF-XML: `Give only the final answer inside <answer>...</answer>.`

Plain: `Give only the final answer.`

No-reason: `Do not provide any explanation or reasoning. Output only the final answer.`

Final-label: `Give only the final answer prefixed by "Final answer:".`

Final-XML: `Give only the final answer inside <final>...</final>.`

Before applying the fixed task-answer parser, evaluation removes only the specified wrapper. For AF-XML, it removes `<answer>...</answer>`; for Final-XML, `<final>...</final>`; and for Final-label, the `Final answer:` prefix. Plain and No-reason require no wrapper removal. No additional format-recovery rule is introduced. Within each prompt, the two objectives share the same semantic pairs, presentation schedule, parser, clean handoff, and validation examples. This design evaluates the procedure with a separately trained adapter for each prompt. It does not test transfer of a single adapter to unseen prompts.

I.2 Aggregate and Per-Task Results

Table 25 reports accuracy together with onset KL and the absolute accuracy gap. FDD-Logit lowers onset KL under every formulation and reduces the mean gap from 10.50 pp to 2.92 pp. Hard FDD reduces the gap under four of the five formulations but has higher onset KL than FDD-Logit throughout. Table 26 gives the corresponding task-level accuracies and divergences.

Table 25: Prompt-level validation results on Qwen3-VL-8B with $N = 2$, $P = 16$. Accuracy is the equal-weight five-task mean in percent; gap is the mean absolute task-wise Text–Image accuracy difference in percentage points; onset KL is in nats.

	Text	Base Image			Hard FDD			FDD-Logit		
Prompt	Acc.	Acc.	Gap	KL	Acc.	Gap	KL	Acc.	Gap	KL
AF-XML	86.4	73.7	13.3	1.319	**85.8**	**2.6**	0.909	85.5	2.9	**0.293**
Plain	84.6	62.9	21.7	2.490	77.7	6.9	1.145	**80.1**	**4.5**	**0.665**
No-reason	64.6	62.7	2.3	1.215	62.2	2.4	0.677	**62.7**	**2.1**	**0.564**
Final-label	47.9	51.6	6.3	0.906	**49.8**	3.9	0.959	45.4	**2.5**	**0.328**
Final-XML	86.8	78.7	8.9	1.004	85.2	2.8	0.647	**85.4**	**2.6**	**0.274**

The relationship between alignment and accuracy depends on the prompt. Under Final-label, FDD-Logit reduces the gap while Image accuracy falls from 51.6% to 45.4%; Hard FDD reaches 49.8%. No-reason starts with a gap of only 2.3 pp, leaving less scope for gap reduction. These results support first-decision training beyond the original XML marker without implying that closer agreement with every prompt-specific teacher improves task performance.

I.3 Answer Parsing and Generation Limits

Table 27 shows substantially lower parsing success under Final-label than under the other formulations. Its scored accuracies therefore reflect the fixed output-format parser as well as the generated answers. Parsing and generation-limit rates are reported alongside the task results so that format-dependent effects are not mistaken for uniform changes in reasoning ability.

J First-Decision Control on Structured Charts

J.1 Paired Chart and Text Inputs

Each PlotQA pair contains an original chart image and a Text serialization of the official structured annotation, with the same question and AF-XML instruction. The serialization follows a fixed order:

Table 26: Per-task validation accuracy and onset KL under five answer-only prompts. Accuracy is in percent; KL compares the matched Text teacher with Base Image or the adapted Image model and is measured in nats. Each condition uses the same training pairs, schedule, and validation examples.

		Text	Base Image		Hard FDD		FDD-Logit	
Prompt	Benchmark	Acc.	Acc.	KL	Acc.	KL	Acc.	KL
AF-XML	GSM8K	94.0	65.0	4.744	**96.5**	2.111	**96.5**	**0.722**
	BBH-seen	93.5	66.5	1.713	**91.5**	2.217	91.0	**0.631**
	MMLU	80.5	73.5	0.124	**76.5**	0.164	75.5	**0.085**
	ARC-Challenge	84.0	85.5	0.017	**86.5**	0.052	**86.5**	**0.025**
	CommonsenseQA	80.0	78.0	0.000	**78.0**	**0.000**	**78.0**	0.003
Plain	GSM8K	95.5	35.0	6.753	85.0	2.572	**89.5**	**1.139**
	BBH-seen	88.5	47.0	3.151	73.0	1.644	**79.0**	**1.005**
	MMLU	76.0	71.0	1.019	71.5	0.657	**72.0**	**0.488**
	ARC-Challenge	84.0	83.0	0.497	82.0	0.273	**82.5**	**0.235**
	CommonsenseQA	79.0	78.5	1.031	77.0	0.577	**77.5**	**0.460**
No-reason	GSM8K	32.0	33.0	1.726	32.0	1.168	**32.5**	**0.887**
	BBH-seen	48.0	43.5	1.529	**42.5**	0.895	**42.5**	**0.824**
	MMLU	77.5	73.0	1.065	73.0	0.574	**74.0**	**0.454**
	ARC-Challenge	85.5	85.5	0.575	**85.5**	0.261	**85.5**	**0.215**
	CommonsenseQA	80.0	78.5	1.178	78.0	0.490	**79.0**	**0.437**
Final-label	GSM8K	0.0	6.0	3.227	**4.0**	2.279	0.0	**1.110**
	BBH-seen	7.5	26.5	1.176	**18.0**	2.340	5.0	**0.433**
	MMLU	67.5	65.0	0.115	**66.5**	0.151	61.5	**0.080**
	ARC-Challenge	83.5	82.5	0.011	**82.5**	0.026	**82.5**	**0.016**
	CommonsenseQA	81.0	78.0	0.000	**78.0**	**0.000**	**78.0**	0.001
Final-XML	GSM8K	96.0	78.5	3.473	95.0	1.673	**95.5**	**0.847**
	BBH-seen	94.5	78.0	1.393	91.5	1.471	**92.0**	**0.456**
	MMLU	80.0	72.0	0.136	**73.5**	0.067	**73.5**	**0.052**
	ARC-Challenge	84.0	86.0	0.016	**87.0**	0.023	**87.0**	**0.013**
	CommonsenseQA	79.5	79.0	0.000	**79.0**	**0.000**	**79.0**	0.002

Table 27: Answer-parsing success and generation-limit rates in the prompt-paraphrase validation study. Both measures are percentages and are averaged equally across the five tasks. Parsing follows the same fixed rules within each prompt condition.

	Text		Base Image		Hard FDD		FDD-Logit	
Prompt	Parse	Limit	Parse	Limit	Parse	Limit	Parse	Limit
AF-XML	98.4	1.6	98.6	1.3	**98.8**	**1.2**	98.7	1.3
Plain	99.9	2.0	100.0	0.2	97.1	**1.4**	**97.8**	1.7
No-reason	100.0	0.0	100.0	0.0	**100.0**	**0.0**	**100.0**	**0.0**
Final-label	59.2	1.3	68.4	1.1	**65.4**	**0.9**	58.2	1.6
Final-XML	98.9	1.1	98.6	1.4	98.3	1.7	**98.5**	**1.5**

plot type; title; x- and y-axis labels and tick-label sequences; legend series names; all ordered x/y datapoints for every series; the question; and the response contract. Numeric formatting is deterministic, and the question is not used to select datapoints. No optical character recognition, model-generated transcription, precomputed answer, summary, or explanation is used in the serialization.

The Qwen3-VL-8B study uses 16 training pairs, 200 validation examples, and a held-out cohort of 400 questions from 400 unique plots in the official PlotQA V1 test split. The held-out cohort contains 68 dot-line, 139 horizontal-bar, 57 line, and 136 vertical-bar plots. Its question types comprise 54 arithmetic, 29 comparison, 61 compound, 94 data-retrieval, 44 min/max, and 118 structural questions. Each held-out plot contributes one question.

J.2 Training and Evaluation

FDD-Logit uses $N = 16$, $P = 128$, with eight presentations per pair and three training seeds. The training examples, schedule, prompt, optimizer, and other hyperparameters remain fixed across runs; no training-budget or hyperparameter search is performed in this study. The held-out cohort is selected deterministically before test inference. The three models evaluated on validation are then evaluated on this cohort without retraining, checkpoint reselection, or changes to the configuration or cohort membership.

Scoring follows the official evaluator: textual answers require exact match, nonzero numeric answers allow a 5% tolerance, and zero-valued answers require exact equality. Validation and held-out results are reported separately.

J.3 Direction of the Onset Intervention

On validation, Text accuracy is 52.0% and Base Image accuracy is 69.5%. Among the 181 Image examples whose first answer-opening token is initially top-1, blocking that token reduces accuracy from 76.24% to 24.86%. The intervention harms 94 previously correct examples and rescues one; strict compliance after blocking is 0.0%.

This direction is opposite to the GSM8K deflection result in Appendix E.2. The comparison identifies a boundary of a fixed answer-suppression rule: an appropriate onset decision depends on the input and task. The two interventions use separate task-specific cohorts and are not pooled into a common effect estimate.

J.4 Validation and Held-Out FDD Results

Table 28: PlotQA validation results on 200 questions. FDD-Logit uses $N = 16$, $P = 128$ and is reported as mean $\pm$ SD over three training seeds. Accuracy and agreement/compliance rates are percentages; onset KL is in nats.

Condition	Accuracy (%)	Phase agreement (%)	Onset KL (nats)	Teacher top-1 agreement (%)	Strict compliance (%)
Text teacher	52.0	–	–	–	100.0
Base Image	69.5	91.0	0.251127	90.5	91.0
FDD-Logit	**78.0 ± 0.5**	**100.0 ± 0.0**	**0.001276 ± 0.000543**	**100.0 ± 0.0**	**100.0 ± 0.0**

On the common 200-example validation cohort, FDD-Logit improves Image accuracy by 8.5 ± 0.5 pp and reduces onset KL by $99.49 \pm 0.22\%$. All three runs reach 100% phase agreement, exact teacher top-1 agreement, and strict compliance. Correct-and-compliant accuracy is $78.0 \pm 0.5\%$. The Image student also exceeds the Text teacher's 52.0% accuracy. This is consistent with the clean-handoff design: the teacher supplies only onset supervision, while the Image-conditioned frozen model generates the continuation.

Table 29: Held-out PlotQA evaluation on 400 questions from 400 unique official-test plots. FDD-Logit results are mean $\pm$ SD over the same three trained adapters used in validation, without retraining or test-based selection. All rates are percentages; onset KL is in nats.

Condition	Accuracy (%)	Phase agreement (%)	Onset KL (nats)	Teacher top-1 agreement (%)	Strict compliance (%)	Correct & compliant (%)
Text teacher	59.5	–	–	–	100.0	59.5
Base Image	80.25	92.75	0.202125	93.0	92.75	80.0
FDD-Logit	**85.33 ± 0.52**	**100.0 ± 0.0**	**0.001601 ± 0.000706**	**100.0 ± 0.0**	**100.0 ± 0.0**	**85.33 ± 0.52**

On the held-out cohort, Image accuracy increases from 80.25% to $85.33 \pm 0.52\%$, a gain of 5.08 ± 0.52 pp. Onset KL decreases from 0.202125 to 0.001601 ± 0.000706, a relative reduction of $99.21 \pm 0.35\%$. All three runs again achieve 100% phase agreement, teacher top-1 agreement, and strict compliance, so correct-and-compliant accuracy equals the overall FDD-Logit accuracy.

The absolute accuracy gap to the 59.5% Text teacher nevertheless increases from 20.75 pp to a mean of 25.83 pp, because the Image student improves beyond the teacher. This result separates onset alignment from final-answer matching: closer agreement with the teacher's onset distribution, greater task accuracy, and a smaller Text–Image accuracy gap are not interchangeable outcomes.

K Same-Contract Generalization to Unseen Tasks

The full-data Hard FDD adapters are evaluated on MATH-500, BoolQ, and held-out BBH without task-specific retraining (Hendrycks et al., 2021; Clark et al., 2019; Suzgun et al., 2023). The AF-XML response contract is unchanged. These experiments assess task generalization rather than transfer to new prompts, and they use the full-data adapters rather than the low-data Hard FDD or FDD-Logit models.

Table 30: Same-contract evaluation of full-data Hard FDD on unseen tasks without task-specific retraining. Values are ranges over three training seeds. Approximate 8B phase-agreement ranges retain their reported precision; generation caps are task dependent and measured in tokens; n is the number of evaluation examples.

Scale	Task	n / cap	Full-data Hard FDD Image acc. (%)	Phase agreement (%)
2B	MATH-500	500 / 4096	53.4–53.8	99.6–100.0
2B	BoolQ	400 / 1024	75.0–82.25	90.5–91.75
2B	Held-out BBH	400 / 2048	39.5–40.25	89.5–91.0
8B	MATH-500	500 / 4096	82.6–83.0	$\approx$ 97–99
8B	BoolQ	400 / 1024	87.25	100.0
8B	Held-out BBH	400 / 2048	79.5–82.75	$\approx$ 94–99

Generation caps are 4096 tokens for MATH-500, 1024 for BoolQ, and 2048 for held-out BBH. MATH-500 and held-out BBH include generation-limit hits, but matched unadapted accuracies and exact truncation counts are unavailable for every condition. Table 30 therefore reports observed transfer performance without claiming improvement over all unadapted baselines. High phase agreement can coexist with limited accuracy, particularly at 2B; aligned response starts do not guarantee transfer of the Text presentation's full task performance.

L Supervision and Computational Cost

Table 31 compares supervised target positions under the matched full-data protocol. Each run processes 4,000 unique paired problems over three epochs, totaling 12,000 presentations. One-decision training supervises one target position per presentation, whereas Full-Trace LoRA supervises every token of the paired Text response. The comparison holds the paired training set fixed while varying the scope of output supervision.

Table 31: Supervision and computational cost in one representative full-data run per model scale. One-decision training is shared by Hard FDD and One-Decision Full-LoRA. Wall time and processed input-token counts are measured separately from supervised target positions.

Scale	Training objective	Supervised positions	Wall time (s)	Processed input tokens
2B	Full-Trace LoRA	2,283,072	4131.11	13,163,112
2B	One-decision	**12,000**	**4032.10**	**10,892,040**
8B	Full-Trace LoRA	1,285,401	6316.85	12,165,441
8B	One-decision	**12,000**	**6022.01**	**10,892,040**

Full-Trace LoRA uses 190.3$\times$ as many supervised target positions at 2B and 107.1$\times$ at 8B. The corresponding measured wall-time ratios are only 1.025 and 1.049, and the processed-input-token ratios are 1.209 and 1.117. Target-position savings therefore do not translate proportionally into training-time savings. At inference, clean handoff also requires an adapted prefill followed by a fresh

frozen-base prefill. These results quantify the reduction in supervised target positions. They do not establish a proportional reduction in end-to-end training or inference time.