MoDiff: Learning What Is Hard for Spatial Vision-Language Models

Anonymous authors
Paper under double-blind review

Abstract

Spatial VLMs are often trained with curricula whose difficulty is fixed by relation count, viewpoint change, or task stage. Such scores describe the task, but not necessarily the current learner: a simple depth relation may expose a visual blind spot, whereas a complex question may admit a language shortcut. We introduce MoDiff, a posterior curriculum correction framework for this mismatch. MoDiff maps static difficulty to an item prior and updates a shared multi-skill ability state and item difficulty from ordinary target-likelihood evidence. A posterior-averaged local information score guides sampling without extra VLM forward passes. Under a 25% token budget, MoDiff reaches 62.4% average accuracy across five spatial benchmarks, 3.1 points above static scheduling in three-seed means, with 3.8% wall-clock overhead. These results foreground sample-efficient spatial fine-tuning through online correction of a fixed curriculum.

1 Introduction

Spatial reasoning remains a persistent weakness of vision–language models. Although humans read many spatial relationships from a scene with little effort, current VLMs fail on questions as basic as whether one object is in front of or behind another (Liu et al., 2023; Kamath et al., 2023; Chen et al., 2025). Accuracy deteriorates further with viewpoint change and relational composition (Huang et al., 2025). This weakness constrains robotics, autonomous driving, and embodied agents, for which spatial competence is a prerequisite (Song et al., 2025; Chen et al., 2024).

A natural response is progressive training from perception to relational and multi-step reasoning (Li et al., 2026a). The unresolved issue is how difficulty is defined. Spatial curricula typically derive static difficulty from task metadata such as view count, relation hops, or perspective taking. We use the term human difficulty only when a score is supported by human accuracy or response time. In either case, the score is fixed before a particular model learns.

Static and model-relative difficulty can diverge systematically. Visual representations and attention leave geometric blind spots, so nominally simple depth or viewpoint relations remain hard (Chen et al., 2025); conversely, language or answer-distribution shortcuts can simplify nominally complex questions. A fixed order may therefore revisit mastered samples while postponing uncertain ones, especially when a limited budget precludes repeated coverage.

A naive solution is to replace the static score with the latest loss. This leaves three gaps: an unvisited example has no model-relative score, a visited score becomes stale as the network changes, and independent scores ignore transfer. Updating one depth example can alter every depth judgment because all items share the same VLM parameters. Static difficulty and online evidence are therefore not competing scalars; a suitable state model must separate item difficulty from shared model ability and propagate evidence across related spatial skills.

This observation connects spatial curriculum learning to item-response theory (IRT), but changes the type of learner being modeled. In educational IRT, a response reveals a student's latent ability relative to item difficulty. In our setting, the learner is a neural model that changes after every selected batch, and each image–question pair can involve coupled skills. An independent per-item tracker would miss this shared, action-dependent state. MoDiff represents the VLM by a dynamic vector of spatial abilities and each example by a static-difficulty prior with a multi-skill loading.

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.

Target likelihood from ordinary SFT provides a soft response that jointly corrects ability and item difficulty through online assumed-density filtering.

The posterior also determines what to train. A posterior-averaged local information score replaces a hand-designed window around medium difficulty. Its logistic curvature $p(1-p)$ peaks when ability matches difficulty, while posterior variance favors uncertain items. The score gives a tractable selection rule for the factorized filter.

The framework also defines misalignment directly: static difficulty initializes the prior, and posterior inference measures how much each learner corrects it. Evidence on one depth item updates the shared depth ability used for unseen depth examples. One dynamic IRT state thereby covers cold start, cross-example transfer, learner change, and uncertainty.

We make three contributions.

- We formulate curriculum mismatch as posterior disagreement between static item difficulty and a VLM's dynamic multi-skill ability, separating dataset-designed scores from human difficulty.
- We propose MoDiff, an online dynamic-IRT curriculum in which ordinary target likelihood updates shared ability and item factors and a posterior-averaged local information score guides selection.
- We characterize the local joint-Gaussian information score and its competence-matching optimum, then evaluate sample efficiency, predictive validity, and budget sensitivity.

MoDiff changes only the sampling distribution, reuses ordinary supervised-training quantities, and adds neither VLM parameters nor an auxiliary forward pass.

2 Related Work

Spatial reasoning benchmarks and diagnosis. Spatial benchmarks expose failures from basic relations to multi-step reasoning. VSR and What's Up reveal large gaps even for common spatial relations (Liu et al., 2023; Kamath et al., 2023), while BLINK, VSI-Bench, CV-Bench, OmniSpatial, RoboSpatial, and ViewSpatial-Bench extend evaluation to perception, video memory, 2D/3D understanding, manipulation, and perspective change (Fu et al., 2024; Yang et al., 2025; Tong et al., 2024; Jia et al., 2026; Song et al., 2025; Li et al., 2025). Spatial-DISE further finds a broad ceiling on compositional reasoning (Huang et al., 2025). Mechanistic evidence attributes part of the gap to misallocated attention and visual-encoder geometry, especially for depth relations (Chen et al., 2025; Radford et al., 2021); depth- or region-grounded models reduce but do not eliminate it (Cheng et al., 2024). These studies diagnose capability; we use model-specific difficulty as a training signal.

Training enhancement for spatial VLMs. SpatialLadder progressively trains from perception to reasoning (Li et al., 2026a); SpaRE adds synthetic data (Ogezi & Shi, 2025); Spatial-MLLM adds 3D representations (Wu et al., 2026); and R1-Zero-VSI and SpaceR optimize reasoning outputs (Liao et al., 2025; Ouyang et al., 2025). SpatialLadder is the closest comparison because it prespecifies a perception-to-reasoning order. MoDiff instead updates the difficulty axis from model behavior; representation and data augmentation remain orthogonal.

Curriculum learning and difficulty-aware sampling. Curriculum learning orders examples by difficulty (Bengio et al., 2009); self-paced and automated variants condition selection on the learner (Kumar et al., 2010; Graves et al., 2017; Lalor & Yu, 2020; Feng et al., 2025). PUDF estimates IRT difficulty using an artificial crowd and adapts training to model ability (Meng et al., 2026); AdaCuRL aligns difficulty with capability in RL (Li et al., 2026b). CASS combines multidimensional IRT diagnosis and Fisher-information selection for educational LLMs (Guo et al., 2026). COINCIDE selects vision-language data using concept-skill transferability estimated with a reference model (Lee et al., 2024). MoDiff's narrower distinction is online item-prior correction from the target VLM's ordinary training likelihood, coupled to a temporally updated shared skill state.

Item response and knowledge tracing. IRT separates ability from item difficulty, and knowledge tracing follows mastery through sequential responses; modern variants add multiple skills and un-

certainty (Vie & Kashima, 2019; Minn et al., 2022; Cheng et al., 2025). They primarily predict human responses. MoDiff instead addresses a neural learner changed by selected batches, uses soft target likelihood as evidence, and values examples by reduction in uncertainty about shared spatial state.

Training dynamics, memorization and forgetting. Dataset cartography and forgetting events summarize sample-level training behavior (Swayamdipta et al., 2020; Toneva et al., 2018); other selectors use gradients, influence, or reducible loss (Paul et al., 2021; Feldman & Zhang, 2020; Mindermann et al., 2022). MoDiff uses normalized target likelihood as a single observation in a probabilistic state, where uncertainty, staleness, and revisitation have distinct roles.

3 METHOD

3.1 PROBLEM SETUP

Let $\mathcal{D} = \{(x_i, y_i, h_i, \mathbf{q}_i)\}_{i=1}^{N}$ be a spatial instruction set. Here x_i contains images and a question, y_i is the target answer, $h_i \in [0,1]$ is a static difficulty score, and $\mathbf{q}_i \in \mathbb{R}^K$ is a normalized multi-hot loading over K spatial skills. The skill vocabulary is fixed from task metadata and includes 2D direction, depth order, viewpoint transformation, metric distance, counting, and relational composition. If $\mathcal{S}_i$ is the active skill set, then $q_{ik} = |\mathcal{S}_i|^{-1/2}$ for $k \in \mathcal{S}_i$ and zero otherwise, so $\|\mathbf{q}_i\|_2 = 1$. Thus multi-skill questions do not receive larger total loading merely because they have more metadata tags. A curriculum is a sequence of sampling distributions $\pi_t(i)$ used to fine-tune a VLM f_{ω_t}. We reserve ω_t for the VLM parameters and use $\boldsymbol{\theta}_t$ for the low-dimensional latent ability state maintained by MoDiff. The four difficulty objects are the fixed score h_i, latent item factor b_i, learner-relative gap $G_{t,i}$, and posterior difficulty $\bar{d}_{t,i}$ defined in Eq. (13). Let $\mathcal{H}_t$ denote the sampled items and their soft responses observed through step t.

Static training treats h_i as a model-independent order. MoDiff instead separates an item's latent difficulty b_i from the VLM's time-varying skill ability $\boldsymbol{\theta}_t \in \mathbb{R}^K$. Their difference

$$G_{t,i} = \mathbf{q}_i^\top \boldsymbol{\theta}_t - b_i \tag{1}$$

is the ability–difficulty gap. Conditional mastery is $P_{t,i} = \sigma(G_{t,i})$, where σ is the logistic function. An item is easy for the current model when $G_{t,i} \gg 0$, hard when $G_{t,i} \ll 0$, and at the competence boundary when $G_{t,i} \approx 0$.

3.2 A STATIC SCORE AS A WEAK PRIOR

The static score is useful before an item is visited but may be wrong for the current learner. We map it monotonically to an item-difficulty prior

$$b_i \sim \mathcal{N}(\beta_i^0, v_b^0), \qquad \beta_i^0 = s_h \operatorname{logit}(\varepsilon_h + (1 - 2\varepsilon_h)h_i)\,, \tag{2}$$

where s_h controls scale and ε_h clips extremes. A broad v_b^0 lets h_i initialize ordering without determining the posterior. Without a static score, $\beta_i^0 = 0$ and a larger v_b^0 give an uninformative prior.

The initial VLM ability is

$$\boldsymbol{\theta}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_0). \tag{3}$$

Together with Eq. (2), the zero location anchors IRT's shift ambiguity. Because $\boldsymbol{\theta}_t$ is shared, one depth item changes the posterior gap of every depth-loaded item. Independent trackers cannot represent this transfer.

3.3 DYNAMIC MULTI-SKILL POSTERIOR

At a visit to example i, let $\ell_t(i)$ be the length-normalized target negative log-likelihood. We define soft response evidence

$$r_t(i) = \exp[-\ell_t(i)] \in (0,1]. \tag{4}$$

For multiple-choice-formatted examples, this is still the length-normalized likelihood of the target answer sequence, without normalization across candidate options. We use the generalized Bernoulli

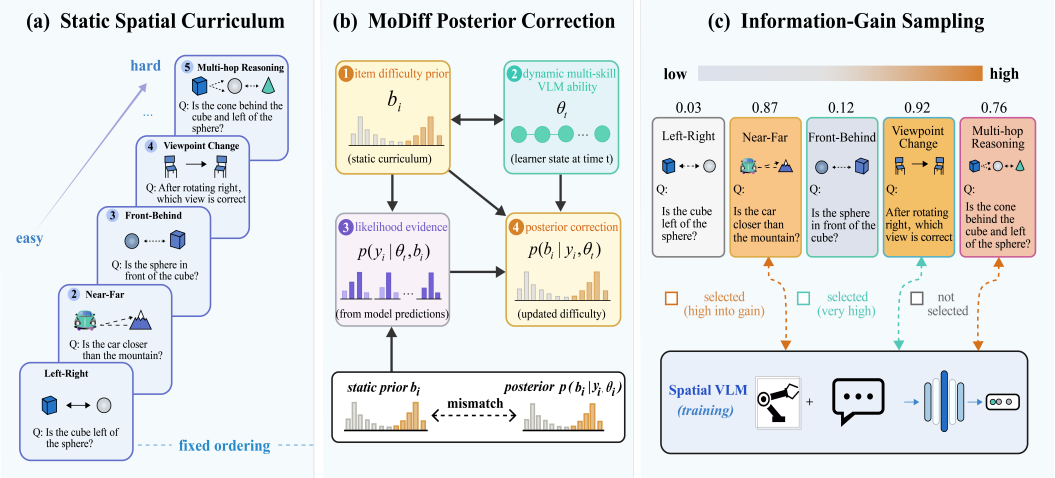

Figure 1: Overview of MoDiff. (a) Static curricula impose a task-defined easy-to-hard order. (b) MoDiff uses the VLM's evolving multi-skill state and target-likelihood evidence to correct the item-difficulty prior, yielding a learner-specific posterior whose mismatch with the static prior identifies ranking reversals. (c) Candidate examples are prioritized by a posterior-averaged local information score, so each batch targets informative skill–item pairs rather than following the fixed order.

log likelihood

$$\log p(r_t(i) \mid G_{t,i}) = r_t(i) \log \sigma(G_{t,i}) + [1 - r_t(i)] \log[1 - \sigma(G_{t,i})]. \tag{5}$$

This soft-label logistic objective is a power likelihood, and all posterior quantities below refer to a generalized Bayesian posterior. The bounded, monotone $r_t(i)$ is evidence rather than calibrated semantic correctness. Length normalization gives the geometric mean target-token probability; the score $r_t(i) - p_i \in [-1, 1]$ and curvature $p_i(1 - p_i) \leq 1/4$ bound each filtering step. The ordinary forward pass supplies the observation.

We approximate the joint posterior with a Gaussian over the shared ability and independent Gaussian item factors. Before a filtering update, ability follows a random-walk state model

$$q(\boldsymbol{\theta}_t) = \mathcal{N}(\boldsymbol{\mu}_t^-, \boldsymbol{\Sigma}_t^-), \quad \boldsymbol{\mu}_t^- = \boldsymbol{\mu}_{t-1}, \quad \boldsymbol{\Sigma}_t^- = \boldsymbol{\Sigma}_{t-1} + \mathbf{Q}_t. \tag{6}$$

Process covariance $\mathbf{Q}_t$ accounts for the preceding VLM update. We use $\mathbf{Q}_t = \sigma_0^2 \mathbf{I} + (\sigma_1^2/|B_{t-1}|) \sum_{j \in B_{t-1}} \mathbf{q}_j \mathbf{q}_j^\top$, with scales selected on development AULC and held fixed across evaluation runs. It raises uncertainty in skills represented by the previous batch. Item difficulty remains static apart from posterior correction: $q(b_i) = \mathcal{N}(\beta_i, v_i)$.

For observed item i, concatenate the relevant state as $\mathbf{w}_i = [\boldsymbol{\theta}; b_i]$, design vector $\mathbf{x}_i = [\mathbf{q}_i; -1]$, Gaussian mean $\mathbf{m}^-$, and covariance $\mathbf{S}^-$. At $z_i = \mathbf{x}_i^\top \mathbf{m}^-$, let $p_i = \sigma(z_i)$ and $c_i = p_i(1 - p_i)$. One online Laplace/assumed-density-filtering step is

$$(\mathbf{S}^+)^{-1} = (\mathbf{S}^-)^{-1} + c_i \mathbf{x}_i \mathbf{x}_i^\top, \tag{7}$$

$$\mathbf{m}^+ = \mathbf{m}^- + \mathbf{S}^+ (r_t(i) - p_i) \mathbf{x}_i. \tag{8}$$

Only the K abilities and sampled item factor participate. For a mini-batch, we apply rank-one updates sequentially over distinct item factors, then moment-match to the shared ability Gaussian and independent item marginals: retain the updated $\boldsymbol{\theta}$ mean and $K \times K$ covariance and each updated (β_i, v_i), and set ability–item and inter-item covariances to zero before the next batch. The update both corrects item priors and transfers evidence, while factorized ADF keeps cost at $O(|B_t|K^2)$.

For any item, the posterior gap is Gaussian,

$$\begin{aligned} G_{t,i} \mid \mathcal{H}_t &\sim \mathcal{N}(m_{t,i}, v_{t,i}), \\ m_{t,i} &= \mathbf{q}_i^\top \boldsymbol{\mu}_t - \beta_i, \\ v_{t,i} &= \mathbf{q}_i^\top \boldsymbol{\Sigma}_t \mathbf{q}_i + v_i. \end{aligned} \tag{9}$$

An unseen item therefore has a state through its prior and shared skills, and related evidence changes it immediately. A per-item score cannot express this transfer.

3.4 Information-Gain Curriculum

MoDiff scores items with a local joint-Gaussian information criterion. If curvature is evaluated at a fixed gap g, an observation adds scalar precision $\sigma(g)[1-\sigma(g)]$ along $\mathbf{x}_i$. By the matrix determinant lemma, the corresponding one-step entropy reduction is

$$\Delta H_i(g) = \frac{1}{2}\log(1 + v_{t,i}\,\sigma(g)[1-\sigma(g)])\,. \tag{10}$$

Because $G_{t,i}$ is uncertain, the curriculum averages this conditional score over Eq. (9):

$$U_t(i) = \mathbb{E}_{g\sim\mathcal{N}(m_{t,i},v_{t,i})}[\Delta H_i(g)]. \tag{11}$$

We use fixed Gauss–Hermite nodes. This averaging is a selection surrogate: the implemented filter evaluates curvature at $m_{t,i}$ in Eq. (7) and discards ability–item covariance after moment matching. Accordingly, $U_t(i)$ is not asserted to equal the exact expected entropy reduction of the maintained factorized posterior. Item variance affects future selection of that item, while ability covariance transfers information to related examples. The sampling distribution is

$$\pi_t(i) = \frac{\epsilon}{N} + (1-\epsilon)\frac{U_t(i)^{1/\tau}}{\sum_j U_t(j)^{1/\tau}}, \tag{12}$$

where τ controls sampling sharpness and ϵ guarantees support. No hand-designed easy-to-hard schedule remains. Difficulty matching and epistemic uncertainty enter through the same entropy objective.

Proposition 1 *For any fixed $v_{t,i} > 0$, the conditional gain $\Delta H_i(g)$ is uniquely maximized at $g = 0$, equivalently at conditional mastery $1/2$.*

Proof sketch. For fixed positive variance, Eq. (10) is strictly increasing in the logistic Fisher term $\sigma(g)[1-\sigma(g)]$, which is symmetric and uniquely maximized at $g = 0$. □

Proposition 2 (Conditional local joint entropy reduction) *For a local joint Gaussian ability–item state with gap variance $v_{t,i}$, a rank-one precision update evaluated at a fixed gap g reduces joint Gaussian entropy by Eq. (10). The implemented mean-curvature update uses $g = m_{t,i}$ before factorization.*

Proof sketch. Gaussian entropy is $\frac{1}{2}\log\det(2\pi e\mathbf{S})$. Substituting the rank-one precision update and applying the determinant lemma gives $\frac{1}{2}\log(1 + c_i\mathbf{x}_i^\top\mathbf{S}^-\mathbf{x}_i)$, where the quadratic form is $v_{t,i}$ and $c_i = \sigma(g)[1-\sigma(g)]$. □

The conditional result identifies the competence frontier; Eq. (11) averages the local score over gap uncertainty. The contribution is structural: one posterior corrects the static prior, transfers skill evidence, tracks learner drift, and supplies the next sampling distribution. The appendix gives the derivation and implementation factorization.

3.5 Difficulty Misalignment as Posterior Correction

MoDiff defines posterior model difficulty as

$$\bar{d}_{t,i} = 1 - \mathbb{E}_{G_{t,i}}[\sigma(G_{t,i})]. \tag{13}$$

The signed correction $\Delta_{t,i} = \bar{d}_{t,i} - h_i$ measures how far the current learner reverses the static assumption. Over a fixed analysis pool, global misalignment is

$$\mathrm{Mis}(t) = 1 - \rho_{\mathrm{S}}\big(h,\,\bar{d}_t\big), \tag{14}$$

where ρ_{S} is Spearman correlation. We report $\Delta_{t,i}$ with gap variance $v_{t,i}$ so that a confident reversal is not confused with missing evidence. Skill-level correction follows by aggregating items with each loading in $\mathbf{q}_i$. Misalignment is diagnostic rather than an extra term in Eq. (12): the dynamic IRT posterior already performs the correction.

Algorithm 1 MoDiff dynamic-IRT curriculum

Require: $\mathcal{D}$, skill loadings $\mathbf{q}_i$, process noise $\mathbf{Q}_t$, temperature τ, exploration ϵ, refresh R
initialize the posterior for f_{ω_0} with Eqs. (2)–(3)
for $t = 0$ to $T - 1$ **do**
  if $t = 0$ or $t \bmod R = 0$ **then**
    compute gap posteriors, utilities, and π_t
  end if
  sample $B_t \sim \pi_t$ using Eq. (12)
  forward B_t under f_{ω_t}; obtain SFT loss and $r_t(i)$
  filter ability and item factors with Eqs. (7)–(8)
  update VLM parameters $\omega_t \to \omega_{t+1}$ on B_t
  predict the next ability covariance via Eq. (6)
end for
return f_{ω_T}

Table 1: Accuracy (%) on five spatial reasoning benchmarks. The first three rows are results reported by SpatialLadder (Li et al., 2026a); the remaining rows are our three-seed means under the tight 25% token budget.

Method	VSI	SP-SI	SP-MV	CV-Bench	ViewSpatial	Avg.
Qwen2.5-VL-3B	29.4	40.3	36.6	70.6	35.6	42.5
GPT-4o	34.0	42.4	48.2	75.4	32.6	46.5
SpatialLadder-3B	45.7	70.2	70.9	73.7	44.2	60.9
Random	42.0	66.0	67.5	71.8	40.5	57.6
Static curriculum	44.5	68.5	69.0	72.5	42.0	59.3
Loss self-paced	45.5	69.5	70.5	73.0	43.5	60.4
Ability-adaptive	46.3	70.5	71.2	73.6	44.2	61.2
Difficulty–utility	47.0	71.0	71.8	74.0	45.0	61.8
MoDiff (ours)	**48.1**	**71.8**	**72.4**	**74.2**	**45.6**	**62.4**

Complexity. With $K \ll N$, factorized filtering costs $O(|B_t|K^2)$ per update and stores $O(K^2+N)$ statistics. Computing all utilities costs $O(NK^2+NL)$ for L fixed quadrature nodes and is refreshed every R optimizer steps. This overhead is independent of VLM parameter count and requires no extra forward pass.

4 Experiments

Evaluation questions. We ask whether posterior correction improves accuracy and sample efficiency (RQ1), whether the inferred state predicts later error and next-visit utility (RQ2), which state components are necessary (RQ3), and whether the gains remain consistent across budgets and model families (RQ4).

Backbone and data. We use Qwen2.5-VL (Bai et al., 2025) at the 3B and 7B scales, with LLaVA-OneVision (Li et al., 2024) as a cross-family check. The training pool is SpatialLadder-26K (Li et al., 2026a). Static score h_i is constructed from the released perception-to-reasoning metadata and rank-normalized to $[0, 1]$. Skill loadings $\mathbf{q}_i$ are deterministic from question type and relation metadata and are frozen before training.

Training and implementation. Within each backbone, methods start from the same checkpoint and share instruction formatting, preprocessing, AdamW schedule, and LoRA modules. Hyperparameters are selected once on development AULC and frozen across samplers. MoDiff uses 12 skills, 8 quadrature nodes, a refresh interval of 20, temperature 0.7, and exploration 0.05. The appendix lists the available training and numerical-stability settings.

Budgets and controls. Tight, medium, and full budgets correspond to 25%, 50%, and 100% of a complete random run's valid target tokens. Failed examples do not consume budget and are logged

by cause. Scheduler computation counts toward wall-clock overhead but grants no extra updates or validation queries.

Benchmarks. The main evaluation follows SpatialLadder on VSI-Bench, SPBench-SI, SPBench-MV, CV-Bench, and ViewSpatial-Bench. Table 1 includes results reported by SpatialLadder for context; all method comparisons use our controlled runs.

Baselines. We compare with five data-ordering strategies under identical optimization and valid-token budgets. Random samples uniformly, while Static curriculum follows the supplied easy-to-hard score. Loss self-paced uses the latest per-example loss, Ability-adaptive selects data according to estimated model ability (Lalor & Yu, 2020), and Difficulty–utility combines static difficulty with an estimated training utility. Together they cover static, loss-based, ability-based, and history-based ordering under the same forward-pass constraint as MoDiff. PUDF (Meng et al., 2026) requires an artificial crowd for item calibration, and COINCIDE (Lee et al., 2024) uses a reference model for concept-skill transfer; their published pipelines do not satisfy this no-auxiliary-model controlled setting. They are relevant methodological comparators, but this table does not establish empirical superiority over them. Gradient-, influence-, and reducible-loss selectors (Paul et al., 2021; Feldman & Zhang, 2020; Mindermann et al., 2022) require per-example backward statistics or an additional reference model and are therefore outside the matched-overhead comparison.

Metrics. We report accuracy, valid-token AULC, item coverage, and wall-clock overhead. AULC is the area under benchmark accuracy as a function of cumulative valid target tokens, normalized to a 0–100 scale. Coverage is the percentage of unique training items selected at least once by the budget endpoint. Wall-clock overhead is the percentage increase in end-to-end training time relative to the matched random-sampling pipeline. Controlled results are averaged over three seeds. Correlation with future error tests whether the inferred state contains information beyond the latest loss.

4.1 Main Results

Table 1 reports the primary tight-budget comparison. We compare MoDiff with static scheduling and the strongest model-relative baseline under matched valid-token and optimization budgets.

Overall spatial accuracy. MoDiff reaches 62.4 average accuracy, numerically 3.1 points above static, 4.8 above random, and 0.6 above Difficulty–utility. Its three-seed mean margins over Difficulty–utility are positive on all five benchmarks (1.1, 0.8, 0.6, 0.2, and 0.6 points). The largest margin occurs on VSI-Bench and the smallest on CV-Bench. Without seed-level dispersion, these means do not establish that the 0.6-point matched-budget difference exceeds run-to-run variation. The 0.2-point CV-Bench difference supports no separate benchmark-level conclusion.

Accuracy under limited training. The tight budget is most diagnostic because later exposure cannot repair an early allocation error. Relative to difficulty–utility, MoDiff gains 0.6 points in final accuracy, 1.1 AULC points, and 3.0 coverage points. The larger AULC separation indicates earlier improvement, while higher coverage indicates broader sampling rather than repeated exploitation of a small subset. The appendix shows that the final-accuracy margin remains 0.7 points at both medium and full budgets.

Learning efficiency. Final accuracy measures the learned model, AULC records when improvement occurs, and coverage reveals whether the policy neglects most of the pool. MoDiff improves all three without additional VLM parameters or optimizer steps. The result is consistent with earlier allocation of useful training examples while the exploration floor maintains support.

4.2 Assessing the Inferred State

The posterior correction $\Delta_{t,i}$ is a model-internal diagnostic, so its disagreement with the static score is not independent evidence that the initial VLM finds an item hard. Figure 2 displays the exploratory analysis. Its reversal-subset panels do not specify the thresholds or posterior checkpoint used to select that subset, so we do not treat their displayed accuracy as a controlled subgroup result. We

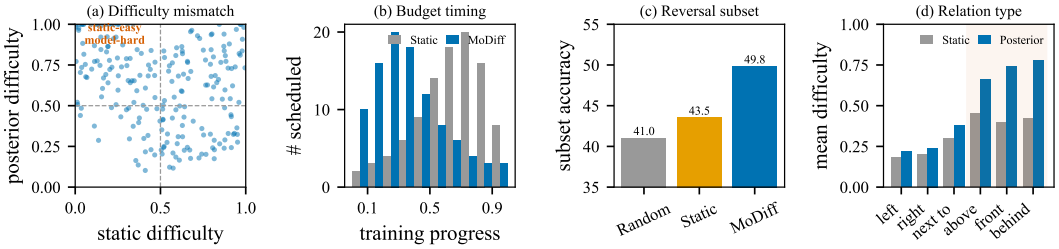

Figure 2: Static–model difficulty misalignment. (a) Static versus posterior difficulty. (b) Training allocation for the reversal quadrant. (c) Accuracy on that quadrant. (d) Prior-to-posterior correction by spatial relation.

assess the state using early-checkpoint predictions of later item error and next-visit training utility in Table 3.

4.3 Ablations

The ablations follow the probabilistic derivation. Independent items removes shared $\boldsymbol{\theta}_t$ and cross-example transfer; Point frontier ranks by $|m_{t,i}|$ and discards covariance; No dynamics sets $\mathbf{Q}_t = 0$, treating the learner as stationary. Table 2 shows these core components. The static-only and no-exploration variants appear with the complete results in Appendix Table A5.

Table 2: Core tight-budget ablations. AULC is normalized to $[0, 100]$.

Variant	Acc.	AULC	Coverage
Full MoDiff	**62.4**	**60.7**	74.0
Independent items	60.4	57.9	69.0
Point frontier	60.7	58.5	66.0
No dynamics ($\mathbf{Q} = 0$)	61.0	59.0	71.0

Table 3: Predictive validity of scores measured at an early checkpoint.

Score	Final error	Next-visit utility
Static difficulty	0.55	0.52
Latest loss	0.69	0.61
Posterior mean difficulty	**0.76**	0.70
Information gain $U_t(i)$	0.74	**0.77**

Component effects. Independent items lose 2.0 accuracy and 2.8 AULC points, showing that evidence transfer through shared skills matters most early in training. The point frontier loses 1.7 accuracy and 2.2 AULC points, so posterior variance contributes beyond matching a point estimate to the competence boundary. Setting $\mathbf{Q} = 0$ loses 1.4 accuracy points, consistent with stale certainty when the VLM changes. In the complete ablation, removing exploration reduces coverage from 74.0 to 55.0 while accuracy falls by only 0.8 points, showing that the support floor mainly protects data diversity.

4.4 Analysis

Posterior behavior. Figure 3(b) illustrates representative posterior mean-difficulty trajectories, including an item whose inferred difficulty rises after initially falling. Panel (a) shows an exploratory association on five datasets distinct from the controlled benchmarks in Table 1; we do not use it to quantify the main result. The trajectories are illustrative; Appendix Table A6 reports the available analysis categories. Forgotten/re-entered items comprise 22% of the pool, and 71% of them were initially static-easy in that analysis.

Predictive validity. Table 3 compares four early-checkpoint scores for final item error and next-visit utility, defined as positive held-out loss reduction after the item's next training visit. Posterior mean difficulty best predicts final error (0.76 AUC); information gain $U_t(i)$ best predicts next-visit utility (0.77 AUC, versus 0.61 for latest loss). In this setting, current hardness and immediate training value call for different scores. The AUC reversal indicates that selecting only the hardest predicted items could miss examples with higher expected next-visit value.

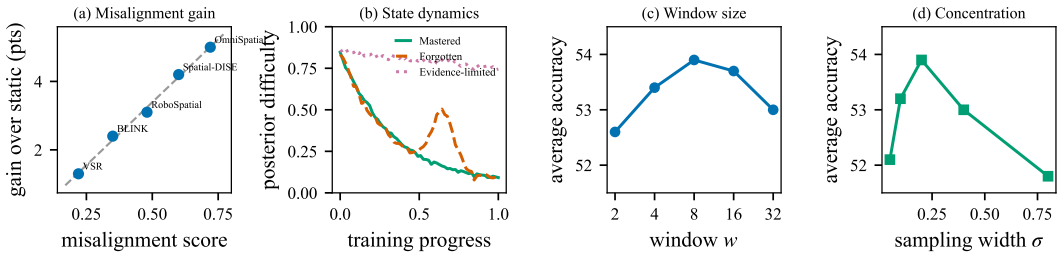

Figure 3: Additional diagnostics. (a) Misalignment versus plotted gain on five datasets distinct from Table 1. (b) Posterior mean-difficulty trajectories. (c) Window w and (d) sampling-width σ sweeps. Their settings are not documented here, so panels (c,d) cannot be interpreted as controlled sweeps of MoDiff's refresh R or temperature τ, or compared directly with Table 1.

Cross-model results and cost. Across Qwen2.5-VL-3B/7B and LLaVA-OneVision, MoDiff exceeds static scheduling by 3.1/2.0/2.0 points, respectively, in the reported means. Its 3.8% end-to-end overhead is 1.2 points above Difficulty–utility and accompanies a 1.1-point AULC difference under the tight budget. The filter and utility refresh use no auxiliary VLM pass; additional budget and backbone results appear in the appendix. Its static-baseline gain narrows from 3.1 to 2.7 points across budgets (Appendix Table A4).

Budget-dependent behavior. Across the three token budgets, MoDiff's five-benchmark mean rises from 62.4 to 62.7 to 62.9, whereas static scheduling rises from 59.3 to 59.8 to 60.2. The margin over static consequently narrows from 3.1 to 2.7 points as exposure grows. The margins over Difficulty–utility remain 0.6, 0.7, and 0.7 points. These are descriptive means; without run-level dispersion they do not establish a reliable ordering at every budget. Under the tight budget, the 1.1-point AULC difference and 3.0-point coverage difference distinguish earlier learning and wider item support from final accuracy. Larger budgets can allow other samplers to revisit missed items, so the strongest evidence here concerns allocation under constrained training rather than a universal ordering of schedulers.

When correction helps. Static scheduling beats random sampling, so its prior remains useful. MoDiff changes rankings as learner evidence arrives, including temporarily harder examples returning to the frontier (Figure 3b). The available posterior diagnostics do not independently validate the initial static-score mismatch.

5 Conclusion

MoDiff corrects a static spatial curriculum using target-likelihood evidence and shared multi-skill ability. At a 25% token budget, it reaches 62.4% average accuracy across five benchmarks, 3.1 points above static scheduling in three-seed means, with 3.8% end-to-end overhead. Ablations and posterior trajectories are consistent with learner-specific ranking correction as a useful sampling signal. By revisiting items whose inferred difficulty changes, MoDiff uses the static curriculum as a prior while adapting selection to the current learner. Its 1.1-point AULC difference in the reported means suggests earlier learning. Testing independent checkpoint measurements and larger training pools will clarify how well this behavior transfers.

Limitations

Evidence covers one 26K pool and 3B–7B VLMs from two families. Metadata-derived skills and factorized Gaussian filtering approximate learner state. The 0.6-point matched-budget mean difference lacks seed-level dispersion, so its robustness to run-to-run variation is unknown. Posterior-derived misalignment lacks independent frozen-checkpoint validation.

AI USE STATEMENT

We used generative AI tools to improve the clarity and fluency of the manuscript, suggest relevant papers and background materials, assist in drafting portions of the manuscript, and check consistency among equations, figures, and claims. The authors take responsibility for the final content, including AI-assisted text, citations, and claims.

REPRODUCIBILITY STATEMENT

Algorithm 1 specifies the training and selection order. The appendix provides the observation and filtering derivations, metadata-based prior construction, reported hyperparameters (Table A2), seed and budget rules, and extended results.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, pp. 41–48, 2009.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, pp. 14455–14465, 2024.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for VLMs? an attention mechanism perspective on focus areas. In *ICML*, pp. 9910–9932. PMLR, 2025.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded spatial reasoning in vision-language models. In *NeurIPS*, volume 37, pp. 135062–135093, 2024.

Weihua Cheng, Hanwen Du, Chunxiao Li, Ersheng Ni, Liangdi Tan, Tianqi Xu, and Yongxin Ni. Uncertainty-aware knowledge tracing. In *AAAI*, volume 39, pp. 27905–27913, 2025.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. In *NeurIPS*, volume 33, pp. 2881–2891, 2020.

Qi Feng, Yihong Liu, and Hinrich Schütze. Your pretrained model tells the difficulty itself: A self-adaptive curriculum learning paradigm for natural language understanding. In *ACL 2025 Student Research Workshop*, 2025.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *ECCV*, pp. 148–166. Springer, 2024.

Alex Graves, Marc G Bellemare, Jacob Menick, Rémi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *ICML*, pp. 1311–1320, 2017.

Yuxiang Guo, Yan Zhuang, Qi Liu, Zhenya Huang, Xianquan Wang, Liyang He, Jiatong Li, Rui Li, and Shijin Wang. From diagnosis to generalization: A cognitive approach to data selection for educational llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21522–21530, 2026. doi: 10.1609/aaai.v40i26.39300.

Xinmiao Huang, Qisong He, Zhenglin Huang, Boxuan Wang, Zhuoyun Li, Guangliang Cheng, Yi Dong, and Xiaowei Huang. Spatial-DISE: A unified benchmark for evaluating spatial reasoning in vision-language models. *arXiv preprint arXiv:2510.13394*, 2025.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. OmniSpatial: Towards comprehensive spatial reasoning benchmark for vision language models. In *ICLR*, 2026.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *EMNLP*, pp. 9161–9175, 2023.

M. Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *NeurIPS*, 2010.

John P. Lalor and Hong Yu. Dynamic data selection for curriculum learning via ability estimation. In *EMNLP*, pp. 545–555, 2020.

Jaewoo Lee, Boyang Li, and Sung Ju Hwang. Concept-skill transferability-based data selection for large vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5060–5080. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.291.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. URL https://arxiv.org/abs/2408.03326.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, et al. ViewSpatial-Bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*, 2025.

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. SpatialLadder: Progressive training for spatial reasoning in vision-language models. In *ICLR*, 2026a.

Renda Li, Hailang Huang, Fei Wei, Feng Xiong, Yong Wang, and Xiangxiang Chu. AdaCuRL: Adaptive curriculum reinforcement learning with invalid sample mitigation and historical revisiting. In *AAAI*, 2026b.

Zhenyi Liao, Qingsong Xie, Yanhao Zhang, Zijian Kong, Haonan Lu, Zhenyu Yang, and Zhijie Deng. Improved visual-spatial reasoning via R1-Zero-Like training. *arXiv preprint arXiv:2504.00883*, 2025.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.

Guangyu Meng, Qinkai Zeng, John P Lalor, and Hong Yu. A psychology-based unified dynamic framework for curriculum learning. *Computational Linguistics*, 2026. doi: 10.1162/COLI.a.584.

Sören Mindermann, Jan M. Brauner, Muhammed T. Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N. Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *ICML*, pp. 15630–15649. PMLR, 2022.

Sein Minn, Jill-Jênn Vie, Koh Takeuchi, Hisashi Kashima, and Feida Zhu. Interpretable knowledge tracing: Simple and efficient student modeling with causal relations. In *AAAI*, volume 36, pp. 12810–12818, 2022.

Michael Ogezi and Freda Shi. SpaRE: Enhancing spatial reasoning in vision-language models with synthetic data. In *ACL*, pp. 7855–7875, 2025.

Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. SpaceR: Reinforcing MLLMs in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In *NeurIPS*, volume 34, pp. 20596–20607, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763. PmLR, 2021.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. RoboSpatial: Teaching spatial understanding to 2D and 3D vision-language models for robotics. In *CVPR*, pp. 15768–15780, 2025.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *EMNLP*, pp. 9275–9293, 2020.

Mariya Toneva, Alessandro Sordoni, Rémi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *ICLR*, 2018.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *NeurIPS*, 2024.

Jill-Jênn Vie and Hisashi Kashima. Knowledge tracing machines: Factorization machines for knowledge tracing. In *AAAI*, volume 33, pp. 750–757, 2019.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-MLLM: Boosting MLLM capabilities in visual-based spatial intelligence. In *NeurIPS*, volume 38, pp. 13569–13597, 2026.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*, pp. 10632–10643, 2025.

A Model and Inference Details

A.1 Notation and Model Summary

Table A1 summarizes the notation. Item i contains input x_i, target y_i, static score h_i, and normalized skill loading $\mathbf{q}_i \in \mathbb{R}^K$. The low-dimensional posterior contains shared ability $\boldsymbol{\theta}_t$ and item difficulty b_i, not VLM parameters ω_t.

Table A1: Notation used in the main paper and appendix.

Symbol	Meaning
h_i	static dataset/curriculum difficulty
$\mathbf{q}_i$	normalized multi-skill loading
b_i	latent item difficulty
$\boldsymbol{\theta}_t$	dynamic VLM spatial ability
$G_{t,i}$	ability–difficulty gap $\mathbf{q}_i^\top \boldsymbol{\theta}_t - b_i$
$r_t(i)$	soft target-likelihood evidence
$\mathbf{Q}_t$	process covariance after VLM updates
$U_t(i)$	posterior-averaged local information score used for selection
$\pi_t(i)$	curriculum sampling probability

Static difficulty is mapped to an item prior, not treated as truth:

$$b_i \sim \mathcal{N}\big(s_h \operatorname{logit}[\varepsilon_h + (1-2\varepsilon_h)h_i], v_b^0\big)\,. \tag{A1}$$

Ability begins at $\boldsymbol{\theta}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_0)$. The anchors fix the otherwise arbitrary joint shift of ability and difficulty. For datasets without h_i, we use zero-mean item priors with larger variance.

Skill loadings are constructed from task metadata, not learned. For active set $\mathcal{S}_i$, $q_{ik} = |\mathcal{S}_i|^{-1/2}\mathbb{I}[k \in \mathcal{S}_i]$; an allocentric depth comparison, for example, loads on depth, viewpoint transformation, and relational comparison. Hence $\mathbf{q}_i^\top \mathbf{q}_i = \operatorname{tr}(\mathbf{q}_i \mathbf{q}_i^\top) = 1$: multiple tags create off-diagonal coupling without increasing total Fisher weight.

At cold start, $\boldsymbol{\mu}_0 = \mathbf{0}$ and item-prior means provide the order, while $\boldsymbol{\Sigma}_0$ and v_b^0 control uncertainty. These scales are selected once on development AULC and frozen. Each backbone initializes a new posterior; absolute ability coordinates are never transferred.

A.2 Generalized Bayesian Observation Model

For gap G, the dynamic IRT mastery probability is $p = \sigma(G)$. Ordinary SFT provides length-normalized target NLL $\ell_t(i)$, from which

$$r_t(i) = \exp[-\ell_t(i)] \tag{A2}$$

is a bounded soft response. We use

$$\mathcal{L}_i(G) = -r_i \log \sigma(G) - (1-r_i)\log[1-\sigma(G)]. \tag{A3}$$

Equation (A3) is logistic cross-entropy with a soft target and can be interpreted as a power likelihood. Its derivatives are

$$\begin{aligned} \frac{\partial \log p(r_i \mid G)}{\partial G} &= r_i - p_i, \\ -\frac{\partial^2 \log p(r_i \mid G)}{\partial G^2} &= p_i(1-p_i). \end{aligned} \tag{A4}$$

The curvature equals logistic Fisher information and is independent of the realized response. Since $r_i - p_i \in [-1, 1]$ and curvature is at most $1/4$, very low likelihood shifts the update toward lower mastery without an unbounded residual. Length normalization prevents answer length alone from exponentially shrinking the response, allowing one observation model to drive filtering and selection.

We use the same raw, length-normalized response mapping in all reported runs, including multiple-choice-formatted examples; it is not renormalized across candidate answers. We clip extreme probabilities as described below. This keeps the observation rule independent of benchmark labels and additional validation queries.

A.3 Online Filtering Derivation

State prediction. The underlying VLM changes after every optimizer update, so ability is non-stationary:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{\xi}_t, \qquad \boldsymbol{\xi}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_t). \tag{A5}$$

The mean is carried forward and covariance becomes $\boldsymbol{\Sigma}_t^- = \boldsymbol{\Sigma}_{t-1} + \mathbf{Q}_t$. We parameterize

$$\mathbf{Q}_t = \sigma_0^2 \mathbf{I} + \sigma_1^2 \sum_{i \in B_{t-1}} \mathbf{q}_i \mathbf{q}_i^\top / |B_{t-1}|. \tag{A6}$$

The isotropic term allows global drift; the low-rank term expresses that the last update most directly perturbed represented skills. This is uncertainty propagation, not an additional difficulty score.

Assumed-density update. For sampled item i, define $\mathbf{w}_i = [\boldsymbol{\theta}; b_i]$, $\mathbf{x}_i = [\mathbf{q}_i; -1]$, prior $q^-(\mathbf{w}_i) = \mathcal{N}(\mathbf{m}^-, \mathbf{S}^-)$, logit $z_i = \mathbf{x}_i^\top \mathbf{m}^-$, probability $p_i = \sigma(z_i)$, and curvature $c_i = p_i(1 - p_i)$. A second-order expansion of the generalized log likelihood around $\mathbf{m}^-$ gives posterior precision

$$(\mathbf{S}^+)^{-1} = (\mathbf{S}^-)^{-1} + c_i \mathbf{x}_i \mathbf{x}_i^\top. \tag{A7}$$

Completing the square gives mean

$$\mathbf{m}^+ = \mathbf{m}^- + \mathbf{S}^+ (r_i - p_i) \mathbf{x}_i. \tag{A8}$$

For a mini-batch B, the conceptual augmented state is $[\boldsymbol{\theta}; \{b_i\}_{i \in B}]$, with a distinct item-coordinate basis for each observation. We apply the rank-one ADF updates sequentially and then moment-match back to a shared ability Gaussian and independent scalar item marginals.

The full covariance over N items is unnecessary. We keep a $K \times K$ ability covariance and scalar (β_i, v_i) for each item. After the sequential batch updates, the factorized ADF projection retains the updated ability mean and covariance and each updated item mean and variance; it sets ability–item and inter-item covariance blocks to zero. Thus ability–item cross-covariance contributes within a local update but is not maintained across batches. The resulting gap distribution is

$$\begin{aligned} G_{t,i} \mid \mathcal{H}_t &\sim \mathcal{N}(m_{t,i}, v_{t,i}), \\ m_{t,i} &= \mathbf{q}_i^\top \boldsymbol{\mu}_t - \beta_i, \\ v_{t,i} &= \mathbf{q}_i^\top \boldsymbol{\Sigma}_t \mathbf{q}_i + v_i. \end{aligned} \tag{A9}$$

Numerical safeguards. We clip probabilities to $[10^{-5}, 1 - 10^{-5}]$, symmetrize the covariance after each update, and floor eigenvalues at 10^{-6}. Item variance is capped at its prior value so process noise affects shared ability but does not make static item properties progressively less identifiable. Utility refresh occurs every R optimizer steps; between refreshes we sample from an alias table.

A.4 Information-Gain Policy

Conditioned on gap g, the rank-one precision added along the item direction is $c(g) = \sigma(g)[1 - \sigma(g)]$. For the local joint ability–item state, Gaussian differential entropy is $H(\mathbf{S}) = \frac{1}{2} \log \det(2\pi e \mathbf{S})$. The matrix determinant lemma therefore yields

$$\Delta H_i(g) = \frac{1}{2} \log[1 + v_{t,i} c(g)]. \tag{A10}$$

MoDiff averages this conditional local score to define its selection utility

$$U_t(i) = \mathbb{E}_{g \sim \mathcal{N}(m_{t,i}, v_{t,i})}[\Delta H_i(g)]. \tag{A11}$$

We compute Eq. (A11) using $L = 8$ Gauss–Hermite nodes. In the implemented ADF update, curvature is evaluated at the predicted mean gap $m_{t,i}$, and factorization discards cross-covariance after each batch. Consequently, the posterior average in Eq. (A11) is a local information surrogate, not the exact expected entropy reduction of the maintained factorized posterior. The curriculum distribution is

$$\pi_t(i) = \epsilon / N + (1 - \epsilon) \frac{U_t(i)^{1/\tau}}{\sum_j U_t(j)^{1/\tau}}. \tag{A12}$$

The information objective is intentionally local and includes both shared ability and item-difficulty uncertainty. It does not equate posterior entropy with downstream accuracy; instead, the experiments test whether this state-based utility improves selection and predicts next-visit learning utility.

A.5 Theoretical Results

Proposition A1 (Conditional competence matching) *For any fixed $v_{t,i} > 0$, the conditional gain $\Delta H_i(g)$ is uniquely maximized at $g = 0$, equivalently at conditional mastery $1/2$.*

Proof. For fixed positive $v_{t,i}$, both multiplication by $v_{t,i}$ and the logarithm preserve ordering in $c(g)$. Since

$$c'(g) = \sigma(g)[1 - \sigma(g)][1 - 2\sigma(g)], \tag{A13}$$

the only finite stationary point is $g = 0$. Since $c(g)$ is symmetric, $c(0) = 1/4$, and $c(g) \to 0$ as $|g| \to \infty$, this stationary point is the global maximum. Equation (A11) averages the competence frontier over posterior gap uncertainty. □

Proposition A2 (Conditional local joint entropy reduction) *For a local joint Gaussian ability–item state, Eq. (A10) is its one-step entropy reduction when the rank-one curvature is evaluated at a fixed g. The implemented mean-curvature update corresponds to $g = m_{t,i}$ before factorized projection.*

Proof. Let $\mathbf{S}^+ = ((\mathbf{S}^-)^{-1} + c\mathbf{x}\mathbf{x}^\top)^{-1}$. Then

$$H(\mathbf{S}^-) - H(\mathbf{S}^+) = \tfrac{1}{2} \log \frac{\det \mathbf{S}^-}{\det \mathbf{S}^+} \tag{A14}$$

$$= \tfrac{1}{2} \log \det(\mathbf{I} + c\mathbf{S}^-\mathbf{x}\mathbf{x}^\top) \tag{A15}$$

$$= \tfrac{1}{2} \log[1 + c\mathbf{x}^\top \mathbf{S}^- \mathbf{x}], \tag{A16}$$

where the last equality is the matrix determinant lemma. The quadratic form equals the gap variance $v_{t,i}$. □

Role of exploration. The exploration floor assigns every item nonzero sampling probability. Under a finite budget, unvisited items retain their static priors, while repeated evidence corrects visited items and propagates through shared skill directions. This conservative update limits unsupported deviations from the prior.

B Implementation and Reported Experimental Setup

B.1 Periodic Refresh and Complexity

Algorithm 1 in the main paper gives the training order. The implementation computes all gap moments and quadrature utilities before the first sample ($t = 0$) and every R optimizer updates thereafter, then builds an alias table for π_t. Between refreshes, batches use the latest table while every observed response still updates the filter. After the standard VLM optimizer step, process covariance is added only to shared ability before the next selection step; item variances are not inflated by learner drift.

Filtering costs $O(|B|K^2)$ and utility refresh costs $O(NK^2+NL)$ with L quadrature nodes. Memory is $O(K^2 + N)$ beyond the VLM. For $K \leq 32$, these costs are small relative to image encoding, but all reported overhead is measured wall-clock rather than inferred asymptotically.

B.2 Difficulty Construction

The training prior uses only dataset-provided metadata. We use the released perception-to-reasoning stage as the primary ordinal key, break ties with relation-hop count, number of views, and viewpoint transformation, and then rank-normalize within the training pool. The 12-way skill mapping is likewise fixed before training and shared by every sampler and backbone; neither h_i nor $\mathbf{q}_i$ uses model responses, validation accuracy, or test performance.

B.3 Reported Experimental Protocol

The protocol follows the four questions in the main paper: RQ1 compares downstream accuracy and sample efficiency, RQ2 tests whether early posterior scores predict later outcomes, RQ3 ablates posterior components, and RQ4 checks consistency across budgets and model families.

Table A3: Published benchmark accuracy (%) reported by SpatialLadder (Li et al., 2026a).

Model	VSI-Bench	SPBench-SI	SPBench-MV	CV-Bench	ViewSpatial
Qwen2.5-VL-3B	29.4	40.3	36.6	70.6	35.6
GPT-4o	34.0	42.4	48.2	75.4	32.6
Qwen2.5-VL-7B	35.8	48.4	37.3	79.0	37.9
SpatialLadder-3B	45.7	70.2	70.9	73.7	44.2

Table A4: Performance across training budgets. Accuracy columns report the five-benchmark mean; AULC, coverage, and overhead correspond to the tight budget.

Method	Tight	Medium	Full	Tight AULC	Coverage	Overhead
Random	57.6	59.0	60.0	54.5	72.0	0.0
Static	59.3	59.8	60.2	56.2	62.0	0.4
Loss self-paced	60.4	60.7	60.9	58.0	67.0	1.2
Ability-adaptive	61.2	61.6	61.9	58.8	69.0	2.0
Difficulty–utility	61.8	62.0	62.2	59.6	71.0	2.6
MoDiff	**62.4**	**62.7**	**62.9**	**60.7**	74.0	3.8

Table A2: Default configuration.

Setting	Value
Backbones	Qwen2.5-VL-3B/7B, LLaVA-OneVision
Training pool	SpatialLadder-26K
Optimizer	AdamW
Learning rate	1×10^{-4}
Effective batch	16
LoRA rank	16
Ability dimensions K	12
Quadrature nodes L	8
Utility refresh R	20 updates
Temperature τ	0.7
Exploration ϵ	0.05
Seeds	0, 1, 2

Budgets are defined by valid target tokens and optimizer updates. A sample that fails image loading, tokenization, or length validation does not consume budget. Every failure is logged. Tight, medium, and full budgets correspond to 25%, 50%, and 100% of the valid-token budget of the full random run.

Hyperparameters are selected using development AULC and then frozen. Final benchmark evaluation is run once per seed, and controlled results are averaged across three seeds.

C Additional Results

C.1 Published Reference Results

Table A3 summarizes benchmark accuracies reported by SpatialLadder (Li et al., 2026a) for comparison with the controlled experiments in the main paper.

C.2 Extended Results

Table A5: Complete tight-budget ablations. AULC is normalized to $[0, 100]$.

Variant	Avg. Acc.	AULC	Coverage
Full MoDiff	**62.4**	**60.7**	74.0
Static only	59.3	56.2	62.0
Independent items	60.4	57.9	69.0
Point frontier	60.7	58.5	66.0
No dynamics ($\mathbf{Q} = 0$)	61.0	59.0	71.0
No exploration	61.6	59.7	55.0

Posterior-state diagnostics. Figure 3(b) in the main paper illustrates representative posterior trajectories. Table A6 reports their distribution and overlap with the static-easy region in the available analysis.

Table A6: Distribution of posterior trajectories and overlap with the static-easy region.

Trajectory type	Fraction of data	Static-easy share
Stably mastered	0.48	0.64
Forgotten/re-entered	0.22	0.71
Persistently below frontier	0.18	0.25
Evidence-limited	0.12	0.46

Forgotten/re-entered items comprise 22% of the pool, and 71% of them were initially labeled easy, so a static order can move past them before their inferred difficulty rises again. Only 25% of persistently below-frontier items are static-easy, indicating that posterior correction retains some prior structure.

Efficiency–coverage tradeoff. At the tight budget, random sampling covers 72.0% of items, static scheduling 62.0%, difficulty–utility 71.0%, and MoDiff 74.0%. The static order repeatedly favors the head of its ranking; MoDiff's exploration floor and uncertainty-sensitive utility widen support. Its 3.8% overhead includes the data pipeline, while VLM capacity and optimization tokens remain fixed.

Table A7: Average tight-budget accuracy (%) across backbones.

Backbone	Static	Diff.–utility	MoDiff
Qwen2.5-VL-3B	59.3	61.8	62.4
Qwen2.5-VL-7B	61.0	62.5	63.0
LLaVA-OneVision	55.5	57.0	57.5

Backbone consistency. MoDiff improves static scheduling by 3.1 points on Qwen2.5-VL-3B and by 2.0 points on both Qwen2.5-VL-7B and LLaVA-OneVision. Relative to difficulty–utility, the margins are 0.6, 0.5, and 0.5 points. Absolute scores vary across model families, but all three comparisons favor MoDiff by a similar amount. Each backbone constructs a fresh posterior from its own target likelihoods, as required by the learner-relative definition of difficulty. The tested 3B–7B range supports cross-backbone consistency; evaluation at larger scales remains future work.

Alternative explanations and scope. Exploration preserves coverage but does not recover the AULC deficits caused by removing shared ability, uncertainty, or dynamics. Controlled runs match target tokens and optimizer updates; MoDiff adds no VLM parameters or auxiliary forward pass. The posterior trajectories are illustrative model-internal diagnostics, and high gap variance signals limited evidence.

D Additional Limitations

The generalized Bernoulli observation assumes comparable target likelihoods across examples; absolute ability values are not transferred between backbones. Metadata-derived skills may miss latent strategies or shortcuts. The factorized Gaussian filter and posterior-averaged selection score are local approximations. Evidence covers one 26K pool and 3B–7B VLMs from two families. Independent frozen-checkpoint mismatch tests and alternate priors would sharpen the empirical picture.