RESIGHT-SMC: TWO-STAGE POWER SAMPLING VIA ISLAND SMC WITH VISUAL SCOUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Power sampling has emerged as a powerful training-free approach to LLM reasoning, eliciting capabilities comparable to reinforcement learning by sharpening the model distribution over complete responses. Despite this success, power sampling remains underexplored in large vision-language models (LVLMs). We first transfer Power-SMC to LVLM decoding by defining a sequence-power target over complete responses conditioned on both the image and the prompt. This direct multimodal transfer provides a strong training-free baseline, but leaves two aspects of finite-particle multimodal inference unaddressed. At the particle level, global resampling can collapse genealogies, while particle-based power sampling does not diversify trajectories through distinct visual cues in multimodal decoding, limiting exploration under a finite particle budget. At the answer level, sequence-level sharpening makes distinct reasoning trajectories compete even when they support the same answer. We introduce ReSight-SMC, a verifier-free two-stage power sampler for LVLM inference. Its first stage uses ancestry-isolated SMC islands to preserve independent trajectory families and routes a bounded set of prefix-conditioned visual scouts to prefix-relevant image regions while discouraging redundant overlap. Each scout temporarily increases attention to the image tokens and further emphasizes its routed region. Exact importance correction preserves the base LVLM sequence-power target. The second stage aggregates terminal importance mass by canonical answer, powers the resulting answer marginal, and samples an answer together with a supporting trajectory. Across four LVLM backbones and five benchmarks, ReSight-SMC achieves stronger aggregate performance than Power-SMC over both the reasoning and perception benchmark groups. Without post-training, it remains competitive in aggregate with backbone-matched models trained using reinforcement learning.

1 INTRODUCTION

Test-time scaling improves LLM reasoning without parameter updates by allocating additional inference compute to sampling, aggregation, search, or refinement (Wei et al., 2022; Wang et al., 2023; Sammani et al., 2026). Sequence-level power sampling provides a model-internal target for test-time scaling. Given a base distribution $p_\theta(y \mid x)$, it samples complete responses from the sharpened distribution $\pi_\alpha(y \mid x) \propto p_\theta(y \mid x)^\alpha$, where $\alpha > 1$. This sharpens complete-response probabilities rather than independently lowering each next-token temperature. Reasoning with Sampling (Karan & Du, 2025) shows that this target can approach reinforcement-learning performance without training or external verifiers, while Power-SMC (Azizi et al., 2026) implements it efficiently with parallel weighted trajectories. Despite its success in LLMs, sequence-level power sampling remains underexplored for open-ended LVLM decoding. Existing multimodal applications focus on embodied control or iterative visual-reasoning refinement (Park et al., 2026; Chen et al., 2026; Jiang et al., 2026b). We transfer Power-SMC (Azizi et al., 2026) to LVLM decoding by defining its target over complete responses conditioned jointly on the image and prompt. As our experiments show, this direct multimodal transfer already provides a strong training-free baseline.

However, global resampling can repeatedly duplicate high-weight trajectories, reducing genealogical coverage and causing the classical path-degeneracy problem in SMC (Del Moral et al., 2006; Doucet & Johansen, 2011). During long visual reasoning, visual reliance can decline as the textual trace grows (Sammani et al., 2026; Favero et al., 2024), while direct Power-SMC lacks a mechanism for

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.

leveraging visual attention early in generation to promote trajectory diversity. Moreover, sequence-level power sharpens individual trajectories before answer aggregation. Trajectories that support the same answer therefore compete through their separate sequence-level weights, allowing one or a few high-probability paths to outweigh collective support distributed across multiple moderate-probability paths (Yang et al., 2026).

We introduce ReSight-SMC, a verifier-free two-stage power sampler for LVLM inference. Its first stage improves finite-particle coverage of the base LVLM sequence-power target. Ancestry-isolated islands resample independently to preserve distinct trajectory families. At a prescribed checkpoint, current prefixes route a bounded set of visual scouts to image regions in the input image while discouraging redundant spatial overlap. Each scout temporarily reactivates attention to all image tokens and gives its assigned region an additional boost. Unfinished base-proposal anchors remain active under the original proposal. Exact importance correction preserves the base LVLM target, so scouting changes exploration rather than the distribution being estimated.

The second stage aggregates normalized terminal trajectory mass by canonical answer, raises the resulting answer masses to a finite power γ, and samples an answer together with a supporting trajectory. Applying the second power after aggregation balances a few high-weight trajectories against broader same-answer support. Across four LVLM backbones and five benchmarks spanning mathematical, logical, and general multimodal reasoning, perception, and real-world spatial understanding, ReSight-SMC outperforms Power-SMC in aggregate on both reasoning and perception-focused tasks. Without post-training, it remains competitive with backbone-matched TRACE-RL and surpasses Game-RL across all benchmarks at the 7B scale.

Our contributions are:

- We extend Power-SMC (Azizi et al., 2026), which approximates the sequence-power target with parallel weighted particles, to open-ended LVLM reasoning.
- We formulate two-stage sequence-to-answer power sampling, combining a sequence-power population with answer-marginal sharpening.
- We formulate an autoregressive island SMC sampler that combines ancestry-isolated resampling with exact sequential importance correction for the sequence-power target.
- We design prefix-conditioned visual scouts that reactivate attention to routed image regions to diversify finite-particle exploration.

2 Related work

LVLM test-time scaling. LVLM test-time scaling spans structured multimodal reasoning and the transfer of sampling, aggregation, and refinement strategies (Mitra et al., 2024; Sammani et al., 2026). Longer traces can improve difficult problems but also weaken attention to image tokens (Sammani et al., 2026). AVIS (Jeddi et al., 2026) jointly adapts visual-token retention and the reasoning-rollout budget through key-based pruning and a learned difficulty predictor. TTAdapt (Kaya et al., 2025) updates model parameters at inference from consensus pseudolabels. ReSight-SMC keeps its parameters fixed and introduces scout proposals within a verifier-free weighted sampler.

Sequence-level power sampling and island SMC. Reasoning with Sampling (Karan & Du, 2025) targets $p(y \mid x)^{\alpha}$ with suffix-resampling Metropolis–Hastings (MH). Later work improves cut selection and scalability and connects power distributions to self-reward objectives (Zhou et al., 2026; Ji et al., 2026; Tomihari & Sato, 2026). MH is serial and requires 16–28$\times$ the latency of standard decoding on MATH500. Power-SMC (Azizi et al., 2026) instead advances weighted particles in parallel and resamples when the ESS falls below a threshold, reducing MATH500 latency to 1.44–$3.25\times$ standard decoding across four models. Global resampling can duplicate a few high-weight trajectories, a classical particle-impoverishment failure mode (Del Moral et al., 2006). We use an island particle system (Vergé et al., 2015) that confines resampling ancestry within each island. Island parallelism also appears in SMCEvolve (Jiang et al., 2026a), which uses migratory SMC chains for reward-tilted program search, whereas our non-migrating islands approximate a common sequence-power target in token-level autoregressive decoding. Multimodal power sampling has so far focused on embodied action and planning (Park et al., 2026; Chen et al., 2026) or MCMC-based refinement of open-ended LVLM reasoning (Jiang et al., 2026b). We transfer Power-SMC to open-ended LVLM

decoding under a base LVLM sequence-power target and extend it with ancestry-isolated islands and prefix-conditioned visual proposals.

Answer aggregation and marginal sharpening. Self-consistency (Wang et al., 2023) samples multiple reasoning paths and returns the most frequent answer as a deterministic decision rule. Marginal sharpening (Arzhantsev et al., 2026) instead defines a powered answer marginal and approximately samples from it through multi-trace decoding. Our readout starts from a sequence-power population: it aggregates corrected SMC mass by canonical answer, applies a second finite power, and samples from the resulting distribution. This composes sequence- and answer-level sharpening without unweighted voting. Yang et al. (2026) modify the trajectory target to preserve support and control per-problem deformation. Our second-stage readout leaves the sequence target unchanged and applies finite-power sharpening after aggregating trajectory mass by answer.

Visual conditioning during inference. Training-free methods modulate visual influence through image-signal amplification (Favero et al., 2024), contrastive decoding (Leng et al., 2024), attention-driven region selection (Gao et al., 2025), or external visual tools (Wang et al., 2025). ReSight-SMC introduces prefix-conditioned visual scouts to diversify particle proposals: each reactivates attention to all image tokens and emphasizes a routed region, while exact importance correction preserves the base LVLM sequence-power target.

3 METHOD

3.1 OVERVIEW

ReSight-SMC separates sharpening into a trajectory stage and a readout stage. The trajectory stage uses island SMC to construct a weighted population for a base LVLM sequence-power target. Islands resample independently so that a single successful lineage cannot eliminate all genealogical diversity. At a prespecified checkpoint, a bounded subset of particles becomes visual scouts: their current prefixes guide routing among candidate image regions under an overlap penalty, and each scout temporarily samples from a proposal that increases attention to image tokens and further emphasizes its routed region. Exact importance correction preserves the same base LVLM sequence-power target.

After generation, the readout stage pools terminal particle mass by canonical answer and applies a finite power to this answer marginal. This lets an answer draw strength from support distributed across multiple moderate-weight trajectories. By contrast, a single-stage sequence-power sampler sharpens trajectories individually and can concentrate mass on a few dominant paths, even when other paths support the same answer.

Together, these components form a coverage-to-decision pipeline. Island-local resampling preserves distinct lineages, so visual scouts operate over more diverse reasoning prefixes. The scouts can discover region-guided continuations that the base proposal may miss in a finite population. Importance correction assigns them substantial weight only when they are also supported by the base LVLM sequence-power target. The answer readout then pools corrected mass across trajectories supporting the same answer and sharpens the resulting answer marginal, allowing these trajectories to contribute jointly to answer selection. Figure 1 summarizes the pipeline, and Appendix G gives the complete inference procedure.

3.2 LVLM SEQUENCE-POWER TARGET AND ISLAND SMC

Let I be an image, x a prompt, H the decoding horizon, and $y = (y_1, \dots, y_T)$ a response completed at step $T \le H$. The fixed LVLM distribution factorizes as $p_F(y \mid I, x) = \prod_{s=1}^{T} p_F(y_s \mid y_{<s}, I, x)$, and stage one targets

$$\pi_\alpha^F(y \mid I, x) = \frac{p_F(y \mid I, x)^\alpha}{Z_\alpha^F(I, x)}, \qquad \alpha > 1. \tag{1}$$

We apply Power-SMC (Azizi et al., 2026) to the autoregressive response tokens while leaving the model's native image conditioning unchanged. It approaches the LVLM sequence-power target through intermediate exponents $1 = \beta_0 \le \cdots \le \beta_H = \alpha$. At step t, the unnormalized target is $\varphi_t(y_{1:t}) = p_{F,t}(y_{1:t} \mid I, x)^{\beta_t}$ and the base proposal locally powers the native next-token conditional,

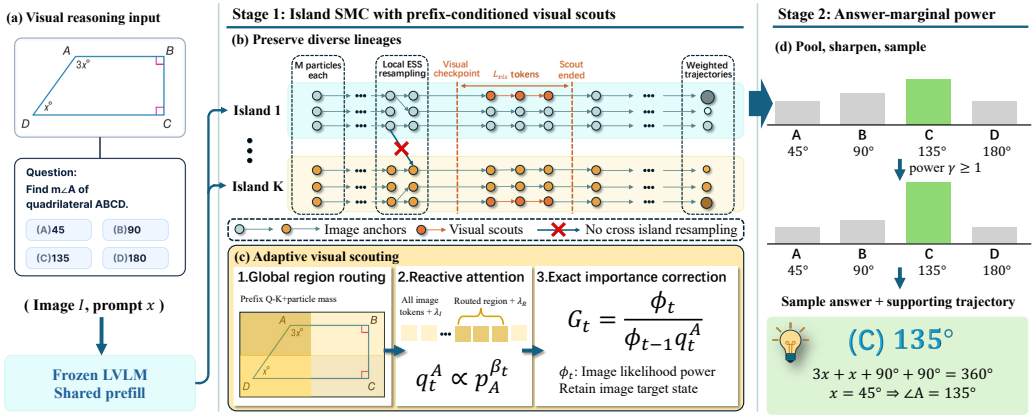


Figure 1: **ReSight-SMC.** Stage one maintains ancestry-isolated SMC islands for the base LVLM sequence-power target. Islands test for ESS-triggered resampling at regular token intervals. At a prespecified visual checkpoint, the remaining unfinished particles continue with the base proposal, while selected scouts use prefix-conditioned routing and attention-reactivated image tokens to form scout proposals. Exact importance correction preserves the base LVLM target. Stage two aggregates terminal mass by answer, applies a finite answer power, and samples an answer with a supporting trajectory.

$q_t^F(v \mid y_{<t}, I, x) \propto p_F(v \mid y_{<t}, I, x)^{\beta_t}$. Local normalization depends on the prefix, so these conditionals alone do not induce Eq. (1). The generic importance weight in Eq. (3) supplies the required sequence-level correction. Its expansion and terminal-sequence treatment are derived in Appendix B.

We maintain K isolated islands with M particles each. A single multimodal prefill of (I, x) produces an image-conditioned decoder state from which all KM particles are initialized. Let $\mathcal{F}_{t-1}$ denote the particle-system history available before step t, including particle prefixes, weights, ancestry, completion states, and any active scout assignments. At step t, the sampler uses $\mathcal{F}_{t-1}$ to select the decoder branch and its settings. For particle (k, m), the selected branch runs the LVLM autoregressively on $(I, x, y_{<t}^{k,m})$ and produces next-token logits. Their softmax is the branch-specific next-token distribution. At step t, particle (k, m) draws its next token from the locally powered proposal

$$q_t^{F,k,m}(v \mid y_{<t}^{k,m}, I, x) = \frac{p_F(v \mid y_{<t}^{k,m}, I, x)^{\beta_t}}{\sum_{u \in \mathcal{V}} p_F(u \mid y_{<t}^{k,m}, I, x)^{\beta_t}}. \tag{2}$$

The next token is drawn as $y_t^{k,m} \sim q_t^{k,m}$. The incremental importance ratio below corrects this branch-specific proposal toward the intermediate target:

$$\begin{aligned} G_t^{k,m} &= \frac{\varphi_t(y_{1:t}^{k,m})}{\varphi_{t-1}(y_{<t}^{k,m}) q_t^{k,m}(y_t^{k,m} \mid y_{<t}^{k,m}, \mathcal{F}_{t-1})}, \\ w_0^{k,m} &= \frac{1}{M}, \qquad w_t^{k,m} = w_{t-1}^{k,m} G_t^{k,m}. \end{aligned} \tag{3}$$

Here $G_t^{k,m}$ is the incremental correction for the sampled token, and $w_t^{k,m}$ is the resulting unnormalized particle weight. A particle that emits EOS remains fixed while the remaining bridge corrections bring its completed response to the common terminal exponent α. Within island k, the weights are normalized as $\bar{w}_t^{k,m} = w_t^{k,m} / \sum_{j=1}^{M} w_t^{k,j}$. The corresponding effective sample size is $\mathrm{ESS}_{k,t} = (\sum_{m=1}^{M} (\bar{w}_t^{k,m})^2)^{-1}$. We evaluate this criterion every L_{SMC} generated tokens. At each checkpoint, island k resamples only when $\mathrm{ESS}_{k,t} < \rho_w M$. When triggered, island k independently applies stratified resampling:

$$A_{k,t}^{1:M} \sim \mathrm{StratResample}(\bar{w}_t^{k,1:M}), \qquad A_{k,t}^{m} \in \{1, \ldots, M\}. \tag{4}$$

Here, $A_{k,t}^m$ is the within-island ancestor selected for descendant m. Power-SMC (Azizi et al., 2026) uses systematic resampling. We instead use stratified resampling because it is not only

conditionally unbiased but also consistent as $M \to \infty$ without imposing an ordering condition on the particles (Gerber et al., 2019). Descendants receive weight $1/M$, and all ancestors remain within island k, preventing population-wide genealogical collapse.

Before any resampling at step t, each island updates its normalizing-constant estimate as

$$c_{k,t} = \sum_{m=1}^{M} \bar{w}_{t-1}^{k,m} G_t^{k,m}, \qquad \widehat{Z}_{k,t} = \widehat{Z}_{k,t-1} c_{k,t}, \qquad \widehat{Z}_{k,0} = 1. \tag{5}$$

The terminal-sequence convention and derivation of these recursions are given in Appendix B.1.

At termination, island mass and normalized within-island weight combine into the pooled first-stage particle mass

$$\widetilde{W}_H^{k,m} = \frac{\widehat{Z}_{k,H} \bar{w}_H^{k,m}}{\sum_{j=1}^{K} \widehat{Z}_{j,H}}. \tag{6}$$

Because particle weights are normalized separately within each island, $\bar{w}_H^{k,m}$ only compares particles within island k. Multiplying by $\widehat{Z}_{k,H}$ yields pooled particle masses used by the answer-level readout.

3.3 Prefix-conditioned visual scouts

The visual-scout episode is a bounded intervention invoked once at the prescribed checkpoint τ_{vis}. Each eligible island receives a scout quota of $\lceil \rho_V M \rceil$, capped to retain at least one unfinished base-proposal anchor. A population router pairs these scouts with regions from a fixed multiscale grid bank $\mathcal{R} = \{R_1, \ldots, R_G\}$. Each scout emphasizes its assigned region for L_{vis} tokens and resumes the base proposal before the next resampling checkpoint. This introduces region-guided proposal variation across selected particles while retaining base-proposal anchors.

Routing scouts to evidence. At τ_{vis}, we compute a visual-only Q–K softmax for the current prefix, normalizing separately within each head over the original image tokens. For each candidate region, we sum this relevance over its tokens, average across heads, and apply region-size normalization:

$$A_g^{k,m} = \frac{1}{N_h |R_g|^{\zeta}} \sum_{a=1}^{N_h} \sum_{i \in R_g} \xi_{i,a}^{k,m}, \qquad 0 \le \zeta \le 1. \tag{7}$$

Here $\xi_{i,a}^{k,m}$ is the visual-only Q–K softmax weight assigned by attention head a at the final decoder layer to image token i for particle (k, m), normalized over all original-image tokens. N_h is the number of attention heads, $|R_g|$ is the number of visual tokens covered by region R_g, and ζ controls the degree of region-size normalization. We combine this score with the particle's pooled SMC mass $\widetilde{W}_{\tau_{\text{vis}}}^{k,m}$ to define the particle–region utility:

$$u_{k,m,g} = \widetilde{W}_{\tau_{\text{vis}}}^{k,m} \frac{A_g^{k,m}}{\max(\sum_{g'=1}^{G} A_{g'}^{k,m}, \epsilon)}. \tag{8}$$

Here $\epsilon = 10^{-12}$ prevents division by zero when a particle assigns zero total relevance to the candidate region bank. We min–max normalize these utilities over the eligible particle–region candidates to obtain $\widetilde{u}_{k,m,g} \in [0, 1]$. The router then greedily selects

$$(k_b, m_b, g_b) = \arg \max_{(k,m,g) \in \mathcal{C}_b} \left[\widetilde{u}_{k,m,g} - \mu \max_{r<b} \text{IoU}(R_g, R_{g_r}) \right]. \tag{9}$$

Exact score ties favor larger k, then m, and finally g. The candidate set $\mathcal{C}_b$ enforces the per-island quota and assigns at most one region to each particle. The utility directs scouts toward prefix-relevant evidence carried by high-mass hypotheses. Here μ is the global soft-overlap coefficient; it discourages redundant assignments without forbidding overlapping evidence. Exact Q–K aggregation, mass and utility normalization, quota edge cases, and routing complexity are given in Appendix C.

Attention reactivation and proposal correction. Let $\mathcal{V}_I$ denote the image-token positions and $\mathcal{V}_I(R_g) \subseteq \mathcal{V}_I$ the positions covered by a routed region. For a scout assigned to R_g, we add the following bias to the causal self-attention logits:

$$b_j(R_g) = \lambda_I \mathbf{1}[j \in \mathcal{V}_I] + \lambda_R \mathbf{1}[j \in \mathcal{V}_I(R_g)]. \tag{10}$$

At every decoder layer, the same visual-token logit bias is added to the current query's attention logits across all heads. It raises attention to all image tokens while giving the routed region an additional boost. Let $p_A^{k,m}$ be the frozen LVLM conditional under this intervention. The scout samples directly from its locally powered proposal

$$q_t^{A,k,m}(v) = \frac{p_A^{k,m}(v \mid y_{<t}, I, x)^{\beta_t}}{\sum_u p_A^{k,m}(u \mid y_{<t}, I, x)^{\beta_t}}. \tag{11}$$

To make the intervention effective from the scout's first newly generated token, we fork the selected particle's decoder state, roll back its cached latest-token computation, and recompute that fixed token once under Eq. (10). This updates the next-token proposal without duplicating or resampling any response token. During the bounded episode, every token sampled from the scout proposal is also passed through an unmodified persistent base target state. This parallel state provides the base-model likelihood required for importance correction and maintains a bias-free KV cache along the same sampled trajectory. The biased state is advanced only when another scout proposal is required and is discarded after L_{vis} tokens, after which decoding continues from the persistent base target state.

The scout branch changes only the proposal. Its exact probability enters the incremental weight:

$$G_t^{A,k,m} = \frac{\varphi_t(y_{1:t}^{k,m})}{\varphi_{t-1}(y_{<t}^{k,m}) q_t^{A,k,m}(y_t^{k,m})}. \tag{12}$$

Since φ_t is evaluated under the base LVLM, this ratio supplies exact importance correction for every scout transition to the sequence target in Eq. (1). Routing therefore alters finite-population exploration and subsequent ancestry while maintaining proper weighting for the same sequence target. The importance-weight derivation is given in Appendix B, and the finite-sample proper-weighting argument is given in Appendix D.

3.4 Answer-marginal power readout

Let $\mathrm{Ans}(y)$ be the canonical answer of response y. We first pool the terminal masses in Eq. (6) by answer and then apply a finite answer-level power:

$$\widehat{\mu}_\alpha(a) = \sum_{k=1}^{K} \sum_{m=1}^{M} \widetilde{W}_H^{k,m} \mathbf{1}[\mathrm{Ans}(y^{k,m}) = a], \tag{13}$$

$$\widehat{q}_{\alpha,\gamma}(a) = \frac{\widehat{\mu}_\alpha(a)^\gamma}{\sum_b \widehat{\mu}_\alpha(b)^\gamma}, \qquad \gamma \geq 1. \tag{14}$$

We draw the final answer from $\widehat{q}_{\alpha,\gamma}$. At the population level, $(\alpha, \gamma) = (2, 2)$ produces a fourth-order score in the base trajectory probabilities. Unlike direct $\alpha = 4$ sequence sampling, the second square is applied after trajectories supporting the same answer have pooled their mass. It therefore preserves cross-trajectory support, reducing domination by a few high-weight paths without reverting to unweighted majority voting. A formal comparison, a separating example, the optional recovery of a supporting trajectory, and the corresponding population-level output distribution are given in Appendices E and E.1.

4 Experiments

4.1 Experimental setup

Our experiments examine whether sequence-power sampling is effective when transferred to open-ended LVLM decoding and whether ReSight-SMC further improves this direct-transfer baseline. We evaluate Qwen2.5-VL-3B-Instruct, Qwen2.5-VL-7B-Instruct, Qwen3-VL-4B-Instruct, and Qwen3-VL-8B-Instruct (Bai et al., 2025b;a). The benchmarks are MathVista, LogicVista, MMStar, and

RealWorldQA (Lu et al., 2023; Xiao et al., 2024; Chen et al., 2024; xAI, 2024). We partition MMStar into reasoning (MMStar-R) and perception (MMStar-P) subsets (see Appendix I.1 for details). Together, these benchmarks cover visual mathematical reasoning, visual logical reasoning, broad perception, and real-world spatial understanding. Accuracy is measured on each released evaluation split.

All systems use the same image preprocessing, answer evaluator, response horizon, and task-specific prompt for a given backbone. MathVista, MMStar-R, and LogicVista use chain-of-thought prompting. MMStar-P and RealWorldQA use direct answering. We evaluate the ESS criterion every 32 generated tokens and resample an island only when its ESS falls below the threshold. Visual scouting is invoked once at $\tau_{\text{vis}} = 40$ and is skipped when generation terminates earlier. Population methods use 32 particles. We report pass@1 as the mean $\pm$ sample standard deviation over four runs with random seeds $\{0, 1, 2, 3\}$. Complete hyperparameters, benchmark protocols, and numerical implementation details are provided in Appendix I.

Compared systems. For every backbone, we compare the base model, low-temperature autoregressive sampling, our direct LVLM transfer of Power-SMC (Azizi et al., 2026), and ReSight-SMC. The transferred Power-SMC baseline applies sequence-power SMC to autoregressive response tokens conditioned on the backbone's native full-image prefill and uses the same backbone, prompt, and total particle budget as ReSight-SMC. All sharpening baselines use the same sequence-level exponent $\alpha = 2$. Low-temperature sampling uses temperature $1/\alpha = 0.5$. ReSight-SMC uses $K = 4$ islands with $M = 8$ particles each and answer exponent $\gamma = 2$.

At the Qwen2.5-VL scales, we additionally evaluate the released TRACE-RL 3B and 7B checkpoints, each trained with GRPO from the corresponding Instruct backbone on procedural visual-reasoning data (Alam, 2026). At 7B we include Game-RL-Qwen2.5-VL-7B, which applies GRPO on GameQA (Tong et al., 2025). These rows provide backbone-matched post-training references.

4.2 Results across backbones

Table 1 compares all systems across backbones and benchmarks.

Table 1: Pass@1 accuracy (%) across backbones and benchmarks, reported as mean $\pm$ sample standard deviation over four seeds. The all-data average is question-weighted. Bold denotes the best non-RL result within each backbone.

System	LogicVista	MathVista	MMStar-R	MMStar-P	RealWorldQA	All-data avg.
Qwen2.5-VL-3B-Instruct						
Base	31.8 ± 2.7	46.2 ± 1.8	42.9 ± 0.5	49.8 ± 4.2	53.4 ± 4.7	45.5 ± 1.8
Low-temp. sampling	33.5 ± 1.2	54.4 ± 1.0	$\mathbf{48.3 \pm 2.1}$	54.9 ± 1.8	58.2 ± 0.9	51.1 ± 0.4
Power-SMC	34.9 ± 1.1	$\mathbf{54.6 \pm 0.6}$	47.0 ± 1.7	53.5 ± 1.5	56.1 ± 1.5	50.3 ± 0.5
ReSight-SMC (Ours)	$\mathbf{35.8 \pm 1.1}$	54.2 ± 1.1	47.2 ± 1.2	$\mathbf{56.3 \pm 1.2}$	$\mathbf{58.8 \pm 0.5}$	$\mathbf{51.3 \pm 0.5}$
TRACE-RL	37.9 ± 2.7	56.4 ± 0.8	53.7 ± 0.4	53.2 ± 3.1	55.5 ± 1.9	52.8 ± 0.8
Qwen2.5-VL-7B-Instruct						
Base	40.3 ± 0.9	66.5 ± 0.3	61.0 ± 1.0	60.2 ± 1.3	66.0 ± 2.5	60.9 ± 0.5
Low-temp. sampling	41.2 ± 1.3	69.5 ± 0.3	62.6 ± 1.4	61.5 ± 1.0	69.5 ± 1.9	63.1 ± 0.3
Power-SMC	42.8 ± 1.0	70.2 ± 0.7	63.7 ± 0.8	61.7 ± 0.9	69.2 ± 0.4	63.8 ± 0.4
ReSight-SMC (Ours)	$\mathbf{43.5 \pm 2.1}$	$\mathbf{71.7 \pm 0.8}$	$\mathbf{65.3 \pm 1.0}$	$\mathbf{63.1 \pm 0.4}$	$\mathbf{70.0 \pm 0.7}$	$\mathbf{65.0 \pm 0.3}$
TRACE-RL	44.0 ± 1.2	74.3 ± 1.5	66.6 ± 0.8	62.6 ± 0.0	67.6 ± 1.8	65.6 ± 0.6
Game-RL	41.4 ± 3.3	66.4 ± 1.2	61.1 ± 0.4	61.0 ± 1.9	66.1 ± 2.4	61.1 ± 0.2
Qwen3-VL-4B-Instruct						
Base	33.9 ± 1.4	62.7 ± 0.7	57.0 ± 0.6	64.0 ± 1.8	69.0 ± 1.1	59.2 ± 0.4
Low-temp. sampling	36.2 ± 1.0	63.5 ± 0.9	57.3 ± 0.9	64.5 ± 0.5	68.9 ± 0.6	59.8 ± 0.3
Power-SMC	45.2 ± 1.0	72.5 ± 0.6	65.4 ± 0.9	67.7 ± 0.8	69.8 ± 0.5	66.1 ± 0.5
ReSight-SMC (Ours)	$\mathbf{48.8 \pm 0.6}$	$\mathbf{73.6 \pm 0.4}$	$\mathbf{67.9 \pm 0.8}$	$\mathbf{68.2 \pm 0.7}$	$\mathbf{70.8 \pm 0.6}$	$\mathbf{67.8 \pm 0.2}$
Qwen3-VL-8B-Instruct						
Base	40.0 ± 0.9	66.1 ± 1.6	60.8 ± 0.6	65.5 ± 1.5	69.9 ± 1.4	62.2 ± 0.4
Low-temp. sampling	43.2 ± 1.0	66.9 ± 1.0	60.6 ± 0.6	63.0 ± 1.8	70.1 ± 0.9	62.5 ± 0.4
Power-SMC	49.8 ± 1.0	74.8 ± 0.7	68.1 ± 0.5	67.7 ± 0.5	$\mathbf{70.5 \pm 0.6}$	68.1 ± 0.2
ReSight-SMC (Ours)	$\mathbf{53.3 \pm 1.1}$	$\mathbf{76.0 \pm 0.6}$	$\mathbf{69.9 \pm 0.6}$	$\mathbf{67.8 \pm 0.5}$	70.1 ± 0.7	$\mathbf{69.3 \pm 0.1}$

Power-SMC improves the all-data average over base sampling on all four backbones and over low-temperature sampling on three of four. ReSight-SMC further achieves the best training-free average on

every backbone, outperforming Power-SMC on 18 of the 20 benchmark results. Without post-training, ReSight-SMC surpasses Game-RL across all reported benchmarks and remains competitive with TRACE-RL. More specifically, ReSight-SMC performs better on perception-focused benchmarks, while TRACE-RL performs better on reasoning benchmarks.

4.3 Component ablation

All ablations use Qwen2.5-VL-7B-Instruct on five benchmarks. Relative to ReSight-SMC, w/o islands uses $K = 1, M = 32$, w/o visual scouts disables scouting, and w/o answer power sets $\gamma = 1$. Additional scout ablations appear in Appendix A.4. All accuracy analyses use the same seeded categorical readout as the main results.

Table 2: Pass@1 (%) for component ablations at a matched 32-particle budget. All-data averages are question-weighted.

Variant	LogicVista	MathVista	MMStar-R	MMStar-P	RealWorldQA	All-data avg.
ReSight-SMC	$\mathbf{43.53 \pm 2.09}$	$\mathbf{71.70 \pm 0.77}$	$\mathbf{65.25 \pm 1.04}$	$\mathbf{63.05 \pm 0.41}$	$\mathbf{70.00 \pm 0.74}$	$\mathbf{65.05 \pm 0.33}$
w/o islands	40.40 ± 1.06	70.68 ± 0.73	64.10 ± 1.76	62.55 ± 0.82	69.97 ± 0.67	64.01 ± 0.63
w/o visual scouts	41.57 ± 1.04	71.10 ± 0.81	64.65 ± 0.51	62.55 ± 0.72	$\mathbf{70.00 \pm 0.74}$	64.42 ± 0.26
w/o answer power	43.02 ± 1.46	71.13 ± 1.19	65.13 ± 0.99	62.05 ± 1.34	68.04 ± 0.51	64.26 ± 0.33

All three components improve the all-data average in Tab. 2. Island isolation improves every benchmark and provides the largest aggregate gain. Answer-marginal power also improves all five benchmarks. At the default scout fraction, visual scouting improves all four benchmarks on which it activates. RealWorldQA is unchanged because no evaluated run reaches the scout phase (see Tab. 7 in Appendix A.3).

5 Discussion

5.1 From island coverage to pass@1

We report pass@1 for each system and readout, pass@4 across four independently seeded executions, and Coverage@32, the fraction of terminal populations containing a correct answer. On tasks that reach an intervention checkpoint, the first stage generally raises Coverage@32 and pass@4. The second stage further raises pass@1 by concentrating answer mass, reducing pass@4 on four of five benchmarks as probability mass becomes more concentrated on the leading answers. Cross-benchmark and generation-horizon analyses are given in Appendices A.2 and A.3.

Reasoning. Reasoning traces remain active beyond the first resampling checkpoint, allowing preserved islands to develop different continuations. Across all three reasoning benchmarks, the first stage raises Coverage@32 and pass@4, while answer-level sharpening further improves pass@1 (Tab. 3 and Appendix A.2). The genealogy statistics in Appendix A.1 associate this gain with more surviving roots and less mass concentrated in the largest ancestral family.

Perception. RealWorldQA is the limiting no-intervention case: every response finishes before the first checkpoint, and no run activates visual scouting (Tab. 7). On RealWorldQA, the first stages of ReSight-SMC and Power-SMC therefore induce the same output distribution. The small observed differences in pass@1, pass@4, and Coverage@32 reflect finite-sample variation rather than a first-stage mechanism. The resulting pass@1 gain mainly comes from the second-stage readout (Tab. 3).

5.2 Trajectory power and answer-level support

The two exponents act at different levels. Increasing α favors individually likely trajectories, whereas applying γ after answer aggregation rewards their collective terminal mass. Our answer-marginal second stage therefore mediates between a few high-weight trajectories and broader support from many medium- or low-weight trajectories. Table 4 compares answer-marginal power with one-stage sequence power at $\alpha = 2$ and $\alpha = 4$.

Table 3: Representative reasoning and perception behavior with Qwen2.5-VL-7B-Instruct. The $\gamma = 1$ rows replay the ReSight-SMC populations without answer-level sharpening.

Task	System/readout	Pass@1	Pass@4	Coverage@32
LogicVista	Power-SMC	42.8 ± 1.0	65.2	63.1 ± 0.8
	ReSight-SMC, $\gamma = 1$	43.02 ± 1.46	66.3	$\mathbf{77.4 \pm 1.4}$
	ReSight-SMC, $\gamma = 2$	$\mathbf{43.53 \pm 2.09}$	**67.9**	$\mathbf{77.4 \pm 1.4}$
RealWorldQA	Power-SMC	69.2 ± 0.4	81.2	96.5 ± 0.2
	ReSight-SMC, $\gamma = 1$	68.04 ± 0.51	**81.7**	$\mathbf{96.6 \pm 0.2}$
	ReSight-SMC, $\gamma = 2$	$\mathbf{70.00 \pm 0.74}$	78.3	$\mathbf{96.6 \pm 0.2}$

Table 4: Pass@1 (%) under alternative sharpening rules. The all-data average is question-weighted. Rows sharing α replay the same token-40 populations; changing α requires separate sampling.

(α, γ)	LogicVista	MathVista	MMStar-R	MMStar-P	RealWorldQA	All-data avg.
$(2, 1)$	43.02 ± 1.46	71.13 ± 1.19	65.13 ± 0.99	62.05 ± 1.34	68.04 ± 0.51	64.26 ± 0.33
$(4, 1)$	42.58 ± 1.20	71.25 ± 0.66	64.08 ± 1.05	$\mathbf{63.10 \pm 0.89}$	$\mathbf{70.59 \pm 0.56}$	64.62 ± 0.35
$(2, 2)$	$\mathbf{43.53 \pm 2.09}$	$\mathbf{71.70 \pm 0.77}$	$\mathbf{65.25 \pm 1.04}$	63.05 ± 0.41	70.00 ± 0.74	$\mathbf{65.05 \pm 0.33}$

Relative to one-stage $\alpha = 2$, answer-marginal second power improves pass@1 on all five benchmarks. It also outperforms one-stage $\alpha = 4$ on all reasoning benchmarks. As shown in Appendix E.1, squaring the aggregated answer mass introduces cross terms between trajectories supporting the same answer, allowing their evidence to reinforce one another. On MMStar-P and RealWorldQA, $(2, 2)$ achieves slightly lower pass@1 than one-stage $(4, 1)$. Appendix F analyzes this behavior in a simplified direct-answer setting.

5.3 Efficiency

The principal additional cost is bounded decoding on the temporary scout branches. Answer aggregation requires no additional model forward pass. On a single NVIDIA RTX 5090 GPU with a common Transformers runtime, per-example latency and peak VRAM are 4.11 seconds and 15.61 GB for base sampling, 7.85 seconds and 17.27 GB for Power-SMC, and 10.37 seconds and 17.35 GB for ReSight-SMC. The complete cost table and asymptotic analysis are provided in Appendix H.

5.4 Limitations

The answer readout cannot repair a wrong or missing first-stage answer mode and depends on a task-appropriate canonicalizer. For unconstrained free-form responses, semantically equivalent strings that cannot be merged deterministically are treated as different answers, fragmenting their marginal mass. Merging them requires an external semantic-equivalence model or LLM judge. Without one, distinct strings remain separate. When every trajectory produces a unique string, answer-level power reduces to particle-wise trajectory sharpening rather than a separate aggregation stage. Access to model logits and decoder states restricts the method to open-weight LVLMs.

6 Conclusion

We first transfer Power-SMC to open-ended LVLM decoding by formulating a sequence-power target over responses conditioned on the image and prompt, establishing a strong training-free baseline for power sampling in LVLMs. We then introduce ReSight-SMC to improve finite-particle exploration and answer-marginal aggregation and sharpening. Island-local resampling preserves trajectory families that seed prefix-conditioned visual scouts, whose routed proposals are corrected to the base LVLM target. The second stage aggregates terminal mass by answer before a finite answer-level power and stochastic draw. Overall, ReSight-SMC delivers aggregate improvements over Power-SMC on both reasoning and perception tasks and remains competitive with post-trained systems, without parameter updates or a learned verifier.

AI USE STATEMENT

Generative AI tools were used to assist with literature search, drafting and editing portions of the manuscript, code inspection, and mathematical derivations and proof checking. All AI-assisted content was reviewed and verified against the implementation, experimental results, and cited sources. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The algorithm and its target distributions are specified in Sec. 3. The experimental setup, matched decoding conditions, and evaluation protocol are described in Sec. 4. Code is available in the supplementary material. Complete hyperparameters, implementation details, benchmark protocols, and complexity analysis are provided in the appendix.

REFERENCES

Md Tanvirul Alam. Trace: A taxonomy-guided environment for multidomain visual reasoning. *arXiv preprint arXiv:2607.19790*, 2026.

Aleksei Arzhantsev, Otmane Sakhi, and Nicolas Chopin. Self-consistency via marginal sharpening. *arXiv preprint arXiv:2605.28142*, 2026.

Seyedarmin Azizi, Erfan Baghaei Potraghloo, Minoo Ahmadi, Souvik Kundu, and Massoud Pedram. Power-smc: Low-latency sequence-level power sampling for training-free llm reasoning. *arXiv preprint arXiv:2602.10273*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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

Kewei Chen, Yayu Long, and Mingsheng Shang. Drift is a sampling error: Snr-aware power distributions for long-horizon robotic planning. *arXiv preprint arXiv:2605.09537*, 2026.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, 2024.

Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B*, 68(3):411–436, 2006.

Arnaud Doucet and Adam M. Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. In Dan Crisan and Boris Rozovskii (eds.), *Handbook of Nonlinear Filtering*, pp. 656–704. Oxford University Press, 2011.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14303–14312, 2024.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19520–19529, 2025.

Mathieu Gerber, Nicolas Chopin, and Nick Whiteley. Negative association, ordering and convergence of resampling methods. *The Annals of Statistics*, 47(4):2236–2260, 2019.

Ahmadreza Jeddi, Minh Ngoc Le, Amirhossein Kazerouni, Hakki Can Karaimer, Hue Nguyen, Iqbal Mohomed, Michael Brudno, Alex Levinshtein, Konstantinos G. Derpanis, Babak Taati, and Radek Grzeszczuk. Avis: Adaptive test-time scaling for vision-language models. *arXiv preprint arXiv:2606.11576*, 2026.

Xiaotong Ji, Rasul Tutunov, Matthieu Zimmer, and Haitham Bou Ammar. Scalable power sampling: Unlocking efficient, training-free reasoning for llms via distribution sharpening. *arXiv preprint arXiv:2601.21590*, 2026.

Jiachen Jiang, Huminhao Zhu, and Zhihui Zhu. SMCEvolve: Principled scientific discovery via sequential monte carlo evolution. *arXiv preprint arXiv:2605.15308*, 2026a.

Tianbao Jiang, Weicong Ni, Gerard de Melo, and Linlin Wang. Aligning large vision–language models at test time: A trajectory-guided structured sampling approach. *arXiv preprint arXiv:2608.03204*, 2026b.

Aayush Karan and Yilun Du. Reasoning with sampling: Your base model is smarter than you think. *arXiv preprint arXiv:2510.14901*, 2025.

Mehmet Onurcan Kaya, Desmond Elliott, and Dim P. Papadopoulos. Efficient test-time scaling for small vision-language models. *arXiv preprint arXiv:2510.03574*, 2025.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Jimin Park, Wonjeong Choi, and Jaekyun Moon. Policy-only power sampling for vision-language-action control. *ICML Workshop on Decision-Making from Offline Datasets to Online Adaptation*, 2026.

Fawaz Sammani, Tzoulio Chamiti, and Nikos Deligiannis. On test-time scaling for vision-language models. *arXiv preprint arXiv:2606.28864*, 2026.

Akiyoshi Tomihari and Issei Sato. Power distribution bridges sampling, self-reward rl, and self-distillation. *arXiv preprint arXiv:2605.04542*, 2026.

Jingqi Tong, Jixin Tang, Hangcheng Li, Yurong Mou, Ming Zhang, Jun Zhao, Yanbo Wen, Fan Song, Jiahao Zhan, Yuyang Lu, et al. Game-rl: Synthesizing multimodal verifiable game data to boost vlms' general reasoning. *arXiv preprint arXiv:2505.13886*, 2025.

Christelle Vergé, Cyrille Dubarry, Pierre Del Moral, and Eric Moulines. On parallel implementation of sequential monte carlo methods: The island particle model. *Statistics and Computing*, 25(2): 243–260, 2015.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Yikun Wang, Siyin Wang, Qinyuan Cheng, Zhaoye Fei, Liang Ding, Qipeng Guo, Dacheng Tao, and Xipeng Qiu. Visuothink: Empowering LVLM reasoning with multimodal tree search. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21707–21719, 2025.

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 *Advances in Neural Information Processing Systems*, 2022.

xAI. Realworldqa. `https://huggingface.co/datasets/xai-org/RealworldQA`, 2024.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

Haohui Yang, Jiaxing Sun, and Xiujun Ma. More correct mass, worse answers: Why power sampling can fail and how to fix it. *arXiv preprint arXiv:2608.14420*, 2026.

Felix Zhou, Anay Mehrotra, and Quanquan C. Liu. Reasoning with sampling: Cutting at decision points. *arXiv preprint arXiv:2605.30327*, 2026.

A Additional analysis

A.1 Island genealogy

We report terminal-ancestry statistics to characterize the genealogical effect of island-local resampling. Let ρ_r denote the normalized terminal mass carried by descendants of initial particle r. We report the effective number of roots $N_{\text{root}}^{\text{eff}} = \exp(-\sum_r \rho_r \log \rho_r)$ and the mass of the largest root $\max_r \rho_r$. These statistics use the first-stage terminal weights, before answer-level sharpening.

Table 5: Terminal genealogy on the reasoning benchmarks with Qwen2.5-VL-7B-Instruct. The two rows for each dataset share the same 32-particle budget, visual proposal, and power parameters.

Dataset	Population	Effective roots ↑	Largest-root mass ↓
LogicVista	1×32 (w/o islands)	1.625 ± 0.048	0.900 ± 0.006
	4×8 (ReSight-SMC)	$\mathbf{1.859 \pm 0.051}$	$\mathbf{0.844 \pm 0.010}$
MathVista	1×32 (w/o islands)	3.790 ± 0.030	0.671 ± 0.004
	4×8 (ReSight-SMC)	$\mathbf{4.005 \pm 0.050}$	$\mathbf{0.648 \pm 0.005}$
MMStar-R	1×32 (w/o islands)	2.628 ± 0.027	0.789 ± 0.004
	4×8 (ReSight-SMC)	$\mathbf{2.832 \pm 0.033}$	$\mathbf{0.754 \pm 0.008}$

Across all three reasoning benchmarks, island isolation increases the effective number of terminal roots and reduces the largest ancestral family's mass. The effect is obtained at the same total particle budget and with the same visual and power parameters, so it directly measures the consequence of replacing global resampling with island-local resampling. The result supports the intended mechanism: a high-weight family can expand within its island without erasing promising families maintained by the other islands.

A.2 Coverage, pass@1, and pass@4 across benchmarks

Table 6 extends the comparison in Tab. 3 across the remaining benchmarks.

Table 6: Realized pass@1, pass@4, and Coverage@32 (%) with Qwen2.5-VL-7B-Instruct. Pass@1 and Coverage@32 are mean ± sample standard deviation; pass@4 measures success across four independently seeded complete executions.

Benchmark	System	Pass@1	Pass@4	Coverage@32
MathVista	Power-SMC, 1×32	70.2 ± 0.7	80.4	81.3 ± 0.4
MathVista	ReSight-SMC, $\gamma = 1$	71.13 ± 1.19	**81.9**	$\mathbf{86.6 \pm 0.3}$
MathVista	ReSight-SMC, $\gamma = 2$	$\mathbf{71.70 \pm 0.77}$	81.3	$\mathbf{86.6 \pm 0.3}$
MMStar-R	Power-SMC, 1×32	63.7 ± 0.8	78.4	76.3 ± 0.9
MMStar-R	ReSight-SMC, $\gamma = 1$	65.13 ± 0.99	**78.9**	$\mathbf{85.5 \pm 1.0}$
MMStar-R	ReSight-SMC, $\gamma = 2$	$\mathbf{65.25 \pm 1.04}$	78.2	$\mathbf{85.5 \pm 1.0}$
MMStar-P	Power-SMC, 1×32	61.7 ± 0.9	**72.2**	89.3 ± 1.0
MMStar-P	ReSight-SMC, $\gamma = 1$	62.05 ± 1.34	72.0	$\mathbf{90.0 \pm 0.4}$
MMStar-P	ReSight-SMC, $\gamma = 2$	$\mathbf{63.05 \pm 0.41}$	68.8	$\mathbf{90.0 \pm 0.4}$

Together, these tables show that, before answer-level sharpening, ReSight-SMC attains higher empirical pass@4 than Power-SMC on four of five benchmarks while also increasing Coverage@32. Applying $\gamma = 2$ subsequently improves pass@1 over the $\gamma = 1$ readout on every benchmark, but lowers pass@4 on four of the five benchmarks as probability mass becomes more concentrated on the leading answers. LogicVista is the exception, improving both pass@1 and pass@4.

A.3 Generation horizon and scout activation

The fraction active at a checkpoint counts particles that have not emitted EOS. Completed particles continue through the exponent bridge and checkpoint calculation, but resampling them cannot create new continuations. As shown in Tab. 7, nearly every reasoning particle remains active through both

checkpoints, whereas direct-answer responses are usually complete before resampling and visual scouting.

Table 7: Generation activity and scout activation. "Scout questions" is the fraction of question–seed runs in which at least one particle activates visual scouting.

Benchmark	Active at 32 (%)	Active at 40 (%)	Scout questions (%)
LogicVista	100.0	100.0	100.0
MathVista	100.0	99.9	100.0
MMStar-R	100.0	100.0	100.0
MMStar-P	5.9	4.7	31.2
RealWorldQA	0.0	0.0	0.0

A.4 Additional visual-stage ablations

Table 8 tests the region-specific attention bias and the scope of the overlap penalty. The global-attention-only variant removes the region-specific second bias. The island-local penalty variant changes only the scope of the overlap penalty, retaining the default coefficient $\mu = 1$ and all other settings.

Table 8: Pass@1 (%) for additional visual-stage ablations with Qwen2.5-VL-7B-Instruct.

Variant	LogicVista	MathVista	MMStar-R	MMStar-P	RealWorldQA	All-data avg.
ReSight-SMC	43.53 ± 2.09	71.70 ± 0.77	65.25 ± 1.04	63.05 ± 0.41	70.00 ± 0.74	65.05 ± 0.33
global attention only	42.69 ± 1.28	70.75 ± 0.76	64.65 ± 0.56	62.65 ± 0.84	70.00 ± 0.74	64.48 ± 0.19
island-local overlap penalty	42.80 ± 1.36	71.58 ± 1.35	65.00 ± 1.28	63.20 ± 0.59	70.00 ± 0.74	64.88 ± 0.69

The region-specific bias improves pass@1 over global attention only on every benchmark where scouts activate. With $\mu = 1$, global overlap penalization improves the all-data average over island-local penalization from 64.88% to 65.05% and gives higher means on three of the four scout-active benchmarks. RealWorldQA is unchanged because no scout activates there.

Overlap-penalty coefficient. Table 9 varies the global overlap coefficient μ while retaining the biased route and all other settings. Among the evaluated values, $\mu = 1$ gives the highest question-weighted average; $\mu = 0.5$ gives the highest MathVista and MMStar-P means.

Table 9: Global overlap-penalty sensitivity in realized pass@1 (%) with Qwen2.5-VL-7B-Instruct. Values are mean $\pm$ sample standard deviation over four seeds; the all-data average is question-weighted.

Overlap coefficient	LogicVista	MathVista	MMStar-R	MMStar-P	RealWorldQA	All-data avg.
$\mu = 0$	43.14 ± 0.71	71.38 ± 1.76	65.03 ± 1.35	62.65 ± 0.77	70.00 ± 0.74	64.80 ± 0.66
$\mu = 0.5$	43.02 ± 1.79	$\mathbf{71.80 \pm 0.62}$	64.45 ± 0.29	$\mathbf{63.25 \pm 0.53}$	70.00 ± 0.74	64.83 ± 0.37
$\mu = 1$	$\mathbf{43.53 \pm 2.09}$	71.70 ± 0.77	$\mathbf{65.25 \pm 1.04}$	63.05 ± 0.41	70.00 ± 0.74	$\mathbf{65.05 \pm 0.33}$

Region-size normalization. Table 10 compares the area exponent ζ, which sets the balance between total attention mass and average per-token relevance in region scoring. The intermediate value $\zeta = 0.75$ gives the highest question-weighted average and is used by default.

Scout fraction. Table 11 varies the per-island scout fraction while keeping the particle budget and all other visual settings fixed. The default $\rho_V = 0.25$ gives the strongest LogicVista, MathVista, and MMStar-R results and the highest all-data average. At $\rho_V = 0.5$, these three scores decline slightly but remain above the visual-off variant.

Table 10: Area-exponent sensitivity in realized pass@1 (%).

Area exponent	LogicVista	MathVista	MMStar-R	MMStar-P	RealWorldQA	All-data avg.
$\zeta = 0.5$	42.47±1.76	70.85±0.87	64.60±1.04	62.60±0.85	70.00±0.74	64.46±0.65
$\zeta = 0.75$	**43.53±2.09**	**71.70±0.77**	**65.25±1.04**	**63.05±0.41**	70.00±0.74	**65.05±0.33**
$\zeta = 1$	42.13±1.76	70.48±0.58	64.63±0.90	62.75±0.75	70.00±0.74	64.34±0.38

Table 11: Scout-fraction sensitivity in realized pass@1 (%).

Scout fraction	LogicVista	MathVista	MMStar-R	MMStar-P	RealWorldQA	All-data avg.
visual-off	41.57±1.04	71.10±0.81	64.65±0.51	62.55±0.72	70.00±0.74	64.42±0.26
$\rho_V = 0.125$	43.08±0.87	71.18±0.15	64.13±0.75	**63.05±1.22**	70.00±0.74	64.55±0.21
$\rho_V = 0.25$	**43.53±2.09**	**71.70±0.77**	**65.25±1.04**	**63.05±0.41**	70.00±0.74	**65.05±0.33**
$\rho_V = 0.5$	43.36±0.80	71.28±0.57	64.80±0.55	62.10±1.09	70.00±0.74	64.66±0.42

B IMPORTANCE-WEIGHT DERIVATION

Suppressing the fixed conditioning pair (I, x), let $p_{F,t}(y_{1:t})$ denote the base-model likelihood accumulated by step t, with the likelihood held fixed after a trajectory ends. For $\varphi_t = p_{F,t}^{\beta_t}$, the generic SMC increment, with proposal-history dependence suppressed, is

$$G_t = \frac{\varphi_t(y_{1:t})}{\varphi_{t-1}(y_{<t}) q_t(y_t \mid y_{<t})}. \tag{15}$$

Meaning of the proposal. Let $\mathcal{F}_{t-1}$ denote the information available to the sampler immediately before token t is drawn. It contains all particle prefixes, weights, island masses, resampling ancestry, completion states, and any scout routes and schedules selected by that time. Thus, $\mathcal{F}_{t-1}$ is an information state maintained by the SMC controller. For particle (k, m), the model is evaluated with the standard autoregressive context $(I, x, y_{<t}^{k,m})$, represented computationally by its decoder KV cache and attention mask. Population-level information in $\mathcal{F}_{t-1}$ is used outside the LVLM to select the proposal branch and, for a scout, its routed attention bias. We display $y_{<t}^{k,m}$ separately in the conditioning notation to emphasize autoregressive dependence, although it is contained in $\mathcal{F}_{t-1}$.

Let $\mathcal{S}_{t-1}$ be the particles whose scout window is active before step t. The next token is drawn from the realized proposal

$$q_t^{k,m}(v \mid y_{<t}^{k,m}, \mathcal{F}_{t-1}) = \begin{cases} q_t^{A,k,m}(v \mid y_{<t}^{k,m}, I, x), & (k,m) \in \mathcal{S}_{t-1}, \\ q_t^{F}(v \mid y_{<t}^{k,m}, I, x), & \text{otherwise.} \end{cases} \tag{16}$$

We write q_t^F when the particle indices are clear from the conditioned prefix. It is the powered native next-token conditional, whereas $q_t^{A,k,m}$ is the powered conditional produced by the routed attention bias. See Eqs. (11) and (26). An active scout has $q_t^{A,k,m}$ as its next-token proposal. The realized-token probability appears in the denominator of G_t, ensuring that a continuation does not receive excess target mass merely because its proposal samples it more often.

From a path weight to an incremental weight. For a fixed ancestral path, suppress the particle indices and let

$$Q_t(y_{1:t}) = Q_{t-1}(y_{<t}) q_t(y_t \mid y_{<t}, \mathcal{F}_{t-1}), \qquad Q_0 = 1, \tag{17}$$

be its joint proposal probability. Iterating this recursion gives $Q_t(y_{1:t}) = \prod_{s=1}^{t} q_s(y_s \mid y_{<s}, \mathcal{F}_{s-1})$. Let $\pi_t(y_{1:t}) = \varphi_t(y_{1:t})/Z_t$ denote the normalized target, where $Z_t = \sum_{y_{1:t}} \varphi_t(y_{1:t})$. Under standard importance sampling (Doucet & Johansen, 2011), a prefix drawn from Q_t receives weight

$$\frac{\pi_t(y_{1:t})}{Q_t(y_{1:t})} = \frac{\varphi_t(y_{1:t})}{Z_t Q_t(y_{1:t})}. \tag{18}$$

The normalizer Z_t is generally unavailable, but it is common to every particle and therefore cancels when the weights are normalized. Ignoring this common factor and the initial particle weight $1/M$,

the cumulative unnormalized importance weight is

$$\mathcal{W}_t(y_{1:t}) := \frac{\varphi_t(y_{1:t})}{Q_t(y_{1:t})}. \tag{19}$$

This ratio has a direct interpretation. A prefix is sampled with probability $Q_t(y_{1:t})$ and receives weight $\mathcal{W}_t(y_{1:t}) = \varphi_t(y_{1:t})/Q_t(y_{1:t})$. Multiplying sampling probability by weight gives $Q_t(y_{1:t})\mathcal{W}_t(y_{1:t}) = \varphi_t(y_{1:t})$: the proposal probability cancels path by path, leaving exactly the unnormalized target mass.

Here $\mathcal{W}_t$ is only an auxiliary cumulative weight used to derive G_t. The implemented particle weights $w_t^{k,m}$ follow the same incremental update and are reset after resampling. Hence,

$$\begin{aligned} \frac{\mathcal{W}_t(y_{1:t})}{\mathcal{W}_{t-1}(y_{<t})} &= \frac{\varphi_t(y_{1:t})}{\varphi_{t-1}(y_{<t})} \frac{Q_{t-1}(y_{<t})}{Q_t(y_{1:t})} \\ &= \frac{\varphi_t(y_{1:t})}{\varphi_{t-1}(y_{<t}) q_t(y_t \mid y_{<t}, \mathcal{F}_{t-1})} = G_t. \end{aligned} \tag{20}$$

At the one-step level, the same cancellation is immediate. A token v is sampled with probability $q_t(v \mid y_{<t}, \mathcal{F}_{t-1})$, while its incremental weight contains the reciprocal of this probability. Their product therefore cancels q_t token by token. This correction is valid for any proposal chosen from the available population history, provided that it assigns positive probability wherever the target increment is positive.

Before the response ends, the intermediate targets satisfy

$$\varphi_t(y_{1:t}) = p_F(y_{1:t} \mid I, x)^{\beta_t}, \qquad \varphi_{t-1}(y_{<t}) = p_F(y_{<t} \mid I, x)^{\beta_{t-1}}.$$

Using the autoregressive factorization

$$p_F(y_{1:t} \mid I, x) = p_F(y_{<t} \mid I, x) p_F(y_t \mid y_{<t}, I, x)$$

in the incremental-weight definition yields

$$G_t = \frac{p_F(y_t \mid y_{<t}, I, x)^{\beta_t} p_F(y_{<t} \mid I, x)^{\beta_t - \beta_{t-1}}}{q_t(y_t \mid y_{<t})}. \tag{21}$$

The two corrections under the base proposal. Equation (21) follows by writing $p_F(y_{1:t} \mid I, x) = p_F(y_{<t} \mid I, x) p_F(y_t \mid y_{<t}, I, x)$ and dividing the target at exponent β_t by the preceding target at exponent β_{t-1}. To see what the importance weight corrects, define the prefix-dependent local normalizer

$$\begin{aligned} Z_t^{\mathrm{loc}}(y_{<t}) &:= \sum_{u \in \mathcal{V}} p_F(u \mid y_{<t}, I, x)^{\beta_t}, \\ q_t^F(y_t \mid y_{<t}, I, x) &= \frac{p_F(y_t \mid y_{<t}, I, x)^{\beta_t}}{Z_t^{\mathrm{loc}}(y_{<t})}. \end{aligned} \tag{22}$$

Substituting $q_t = q_t^F$ into Eq. (21) cancels the current-token power and gives

$$G_t = Z_t^{\mathrm{loc}}(y_{<t}) p_F(y_{<t} \mid I, x)^{\beta_t - \beta_{t-1}}. \tag{23}$$

The two remaining factors have distinct roles. First, independently sampling each token from the locally normalized q_t^F would place the prefix-dependent factor $1/Z_t^{\mathrm{loc}}(y_{<t})$ in the joint proposal. Multiplying the particle weight by $Z_t^{\mathrm{loc}}(y_{<t})$ cancels this factor. Otherwise, local temperature sampling would generally define a different distribution from the desired power of the complete sequence. Second, the prefix was evaluated under exponent β_{t-1} at the preceding step, whereas the new intermediate target evaluates the entire extended prefix under β_t. The factor $p_F(y_{<t} \mid I, x)^{\beta_t - \beta_{t-1}}$ supplies exactly this missing exponent difference to all previously generated tokens.

For an active scout, $y_t^{k,m}$ is sampled from $q_t^{A,k,m}(\cdot \mid y_{<t}^{k,m})$, while φ_t remains defined by the base LVLM. Along the particle's ancestral path, the joint proposal probability therefore extends as

$$Q_t(y_{1:t}^{k,m}) = Q_{t-1}(y_{<t}^{k,m}) q_t^{A,k,m}(y_t^{k,m} \mid y_{<t}^{k,m}).$$

Substituting this factorization into the ratio of consecutive cumulative importance weights gives

$$\begin{aligned} G_t^{A,k,m} &= \frac{\mathcal{W}_t(y_{1:t}^{k,m})}{\mathcal{W}_{t-1}(y_{<t}^{k,m})} \\ &= \frac{\varphi_t(y_{1:t}^{k,m})/Q_t(y_{1:t}^{k,m})}{\varphi_{t-1}(y_{<t}^{k,m})/Q_{t-1}(y_{<t}^{k,m})} \\ &= \frac{\varphi_t(y_{1:t}^{k,m})}{\varphi_{t-1}(y_{<t}^{k,m}) q_t^{A,k,m}(y_t^{k,m} \mid y_{<t}^{k,m})}. \end{aligned} \tag{24}$$

The last equality cancels the preceding-path proposal probability Q_{t-1}. Consequently,

$$\varphi_{t-1}(y_{<t}^{k,m}) q_t^{A,k,m}(y_t^{k,m} \mid y_{<t}^{k,m}) G_t^{A,k,m} = \varphi_t(y_{1:t}^{k,m}).$$

Thus attention reactivation changes how often each continuation is proposed, while importance weighting maps its mass back to the same base LVLM intermediate target.

Completed trajectories. Suppose a response emits EOS at step $e < t$. The completed sequence $y_{1:e}$ then remains fixed: no new token is drawn, so there is no new proposal probability to divide out. The update therefore reduces to

$$\begin{aligned} G_t &= \frac{\varphi_t(y_{1:e})}{\varphi_{t-1}(y_{1:e})} \\ &= p_F(y_{1:e} \mid I, x)^{\beta_t - \beta_{t-1}}. \end{aligned} \tag{25}$$

Multiplying these post-EOS increments from $e+1$ through H telescopes the completed sequence's exponent from β_e to $\beta_H = \alpha$. Thus every terminal response is ultimately weighted by its base-model likelihood raised to the same exponent α, regardless of when it emitted EOS. Response length affects the model likelihood, but not the terminal target exponent.

B.1 Island resampling and terminal mass

For completeness, the normalized base proposal used outside the scout episode is

$$q_t^F(v \mid y_{<t}, I, x) = \frac{p_F(v \mid y_{<t}, I, x)^{\beta_t}}{\sum_u p_F(u \mid y_{<t}, I, x)^{\beta_t}}. \tag{26}$$

Within island k, weights are normalized as $\bar{w}_t^{k,m} = w_t^{k,m} / \sum_{j=1}^M w_t^{k,j}$. At a checkpoint, the island resamples if

$$\mathrm{ESS}_{k,t} = \left(\sum_{m=1}^M (\bar{w}_t^{k,m})^2 \right)^{-1} < \rho_w M, \tag{27}$$

where $\rho_w \in (0, 1]$ is the threshold fraction. Stratified resampling law draws M parents from $\bar{w}_t^{k,1:M}$ and resets their weights to $1/M$. If the test is not triggered, the particles and their weights pass through unchanged. In either case, ancestry is confined to the island:

$$\mathrm{Anc}(k, m, t) \subseteq \{k\} \times \{1, \ldots, M\}. \tag{28}$$

Confining ancestry within each island preserves K separate genealogies.

Each island maintains the standard SMC normalizing-constant estimate. With the pre-update normalized weights, define

$$c_{k,t} = \sum_{m=1}^M \bar{w}_{t-1}^{k,m} G_t^{k,m}, \tag{29}$$

$$\widehat{Z}_{k,t} = \widehat{Z}_{k,t-1} c_{k,t}, \qquad \widehat{Z}_{k,0} = 1. \tag{30}$$

At the terminal bridge exponent, the normalized island mass is

$$\Omega_{k,H} = \frac{\widehat{Z}_{k,H}}{\sum_{j=1}^K \widehat{Z}_{j,H}}. \tag{31}$$

An ordinary hierarchical Island-SMC output first draws an island and then a particle within it:

$$k^\star \sim \mathrm{Cat}(\Omega_{1:K,H}), \tag{32}$$

$$m^\star \mid k^\star \sim \mathrm{Cat}(\bar{w}_H^{k^\star,1:M}). \tag{33}$$

Multiplying these two probabilities gives the pooled terminal mass in Eq. (6). Thus the normalizer estimates compare the target mass represented by different islands without allowing cross-island resampling.

C VISUAL-SCOUT ROUTING SPECIFICATION

The visual intervention occurs after token τ_{vis} has been weighted and before the next response token is sampled. We permit $\tau_{\mathrm{vis}} \geq 1$. The implementation requires a generated token whose query can condition the route. The L_{vis} scout transitions finish before the next island-local resampling checkpoint. If no island contains at least two unfinished particles, routing is skipped.

Let $\mathcal{U}_k$ be the unfinished particles in island k at the checkpoint. Its scout quota is

$$B_k = \min\left\{\lceil \rho_V M \rceil, \max(|\mathcal{U}_k| - 1, 0)\right\}, \qquad B = \sum_{k=1}^{K} B_k. \tag{34}$$

Hence an island with enough unfinished particles receives the common quota, whereas a smaller active island retains at least one base-proposal anchor. The quota is independent of island ESS.

For a fixed decoder layer ℓ, the current text query and original-image visual keys define

$$\xi_{i,a}^{k,m} = \frac{\exp(\langle q_a^{\ell,k,m}, k_{i,a}^{\ell}\rangle/\sqrt{d_h})}{\sum_{j\in\mathcal{V}_I}\exp(\langle q_a^{\ell,k,m}, k_{j,a}^{\ell}\rangle/\sqrt{d_h})}. \tag{35}$$

Here $\mathcal{V}_I$ is the visual-token index set, $a \in \{1,\ldots,N_h\}$ indexes attention heads, and d_h is the per-head key dimension. The softmax is restricted to visual tokens, so $\xi_{i,a}^{k,m}$ records where the current response prefix looks within the image. The score in Eq. (7) aggregates this visual-only relevance over heads and the visual tokens covered by each candidate region. The exponent $\zeta = 0$ uses total relevance mass and favors larger regions, while $\zeta = 1$ uses average relevance per visual token. We use $\zeta = 0.75$ to partially normalize area, balancing the large-region preference of total relevance against the sensitivity of full averaging to compact high-attention regions.

At the routing checkpoint, the pooled mass of particle (k, m) is

$$\widetilde{W}_{\tau_{\mathrm{vis}}}^{k,m} = \frac{\widehat{Z}_{k,\tau_{\mathrm{vis}}}\bar{w}_{\tau_{\mathrm{vis}}}^{k,m}}{\sum_{j=1}^{K}\widehat{Z}_{j,\tau_{\mathrm{vis}}}}. \tag{36}$$

This is the checkpoint analogue of Eq. (6). After computing Eq. (8), let $u_{\min}$ and $u_{\max}$ be the extrema over all initially eligible particle–region pairs. We use

$$\widetilde{u}_{k,m,g} = \frac{u_{k,m,g} - u_{\min}}{u_{\max} - u_{\min} + \epsilon}, \qquad \epsilon = 10^{-12}, \tag{37}$$

where the constant only prevents division by zero when every utility is equal.

Let $\mathcal{P}_{b-1}$ be the particles selected in the first $b-1$ routing steps, with $\mathcal{P}_0 = \varnothing$, and let $n_k(\mathcal{P})$ count selected particles from island k. The candidates for step b are

$$\mathcal{C}_b = \{(k,m,g) : (k,m) \in \mathcal{U}_k \setminus \mathcal{P}_{b-1}, \quad n_k(\mathcal{P}_{b-1}) < B_k\}. \tag{38}$$

This set permits at most one view per particle and at most B_k scouts from island k. Applying Eq. (9) and updating $\mathcal{P}_b = \mathcal{P}_{b-1} \cup \{(k_b, m_b)\}$ produces B assignments. Maintaining each region's maximum overlap with previously selected regions gives $O(BKMG)$ greedy selection.

D UNBIASEDNESS AND EXACT TARGET CORRECTION

Let $\mathcal{F}_t^-$ denote the sigma-field generated by the complete algorithmic history through propagation and weighting at step t, including the current particles, normalized weights, and normalizing-constant estimates, but before the step-t resampling decision. Let $\mathcal{F}_t$ additionally include the resulting ancestry and proposal decisions. At τ_{vis}, particle selection, region assignment, attention-bias construction, cache branching, and one-token replay are included in $\mathcal{F}_{\tau_{\text{vis}}}$ and affect proposals beginning at $\tau_{\text{vis}}+1$.

We assume that each proposal is normalized and has support wherever the target increment is positive, and that token draws are conditionally independent given the population history. For the analysis, a response that emits EOS at step e is padded at every later step by a deterministic absorbing EOS transition. This transition has proposal probability one, leaves the completed response $y_{1:e}$ unchanged, and uses the bridge increment $\varphi_t(y_{1:e})/\varphi_{t-1}(y_{1:e})$. All other transitions from that completed state have zero target mass. Thus the token-step sums below cover active and completed particles within the same recursion. Write $\mathcal{V}_t(y)$ for the available transitions from state y: it is the full vocabulary for an active prefix and the singleton absorbing EOS transition for a completed response.

The implementation applies stratified resampling separately within each island. When the ESS rule triggers resampling, the stratified-resampling procedure constructs the ancestor indices as follows:

$$\begin{aligned} &C_0 = 0, \qquad C_j = \sum_{\ell=1}^{j} \bar{w}_t^{k,\ell}, \qquad I_j = [C_{j-1}, C_j), \quad j = 1, \dots, M, \\ &S_m = [(m-1)/M, m/M), \qquad U_m \overset{\text{ind}}{\sim} \text{Unif}(S_m), \quad m = 1, \dots, M, \\ &A_{k,t}^m = j \quad \text{when } U_m \in I_j, \\ &N_j = \sum_{m=1}^{M} \mathbf{1}\{A_{k,t}^m = j\}. \end{aligned} \tag{39}$$

Here I_j is the cumulative-weight interval assigned to particle j, S_m is the m-th equal stratum, and N_j is the number of offspring of particle j. Conditional on $\mathcal{F}_t^-$,

$$\mathbb{E}[N_j \mid \mathcal{F}_t^-] = \sum_{m=1}^{M} M|S_m \cap I_j| = M|I_j| = M\bar{w}_t^{k,j}, \tag{40}$$

where $|\cdot|$ denotes interval length. Consequently, for every bounded function f,

$$\begin{aligned} &\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} f\left(y_{1:t}^{k,A_{k,t}^m}\right)\middle|\mathcal{F}_t^-\right] \\ &\qquad = \sum_{j=1}^{M} \bar{w}_t^{k,j} f(y_{1:t}^{k,j}). \end{aligned} \tag{41}$$

This proves the conditional unbiasedness of stratified resampling. Because $A_{k,t}^m \in \{1, \dots, M\}$ indexes only particles in island k, resampling cannot introduce cross-island ancestors. When resampling is not triggered, the current particles and normalized weights are carried forward unchanged, so their weighted empirical measure is preserved deterministically. When $\bar{w}_t^{k,m}$ subsequently appears as the weight entering step $t+1$, it therefore equals $1/M$ for a resampled descendant and remains the current normalized weight otherwise.

For particle $i = (k, m)$, write its proposal as $q_t^i(v \mid \mathcal{F}_{t-1})$. For any bounded function f,

$$\begin{aligned} &\mathbb{E}\big[G_t^i(Y_t^i) f((y_{<t}^i, Y_t^i)) \mid \mathcal{F}_{t-1}\big] \\ &= \sum_{v \in \mathcal{V}_t(y_{<t}^i)} \frac{\varphi_t((y_{<t}^i, v))}{\varphi_{t-1}(y_{<t}^i)} f((y_{<t}^i, v)). \end{aligned} \tag{42}$$

The proposal cancels conditionally even though visual routing depends on the joint population at the preceding checkpoint. Between resampling events, each particle therefore follows ordinary sequential importance sampling: it extends its current prefix according to its actual sequence of token

proposals and multiplies its weight by the corresponding incremental importance ratios. Conditional on the incoming weighted population, the expected updated weighted measure equals the target update of that empirical measure. Combined with conditionally unbiased resampling and the island normalizing-mass recursion, this establishes the unbiasedness of the pooled unnormalized estimator in Eq. (44).

After propagation and importance-weight updating at step t, and before any optional resampling, define the pooled unnormalized empirical measure

$$\widehat{\Gamma}_t^{K,M}(f) = \frac{1}{K}\sum_{k=1}^{K} \widehat{Z}_{k,t} \sum_{m=1}^{M} \bar{w}_t^{k,m} f(y_{1:t}^{k,m}) \tag{43}$$

and $\Gamma_t(f) = \sum_{y_{1:t}} \varphi_t(y_{1:t}) f(y_{1:t})$.

Proposition 1 (unbiased unnormalized measure). For every bounded f,

$$\mathbb{E}\Big[\widehat{\Gamma}_t^{K,M}(f)\Big] = \Gamma_t(f). \tag{44}$$

Consequently, $K^{-1}\sum_{k=1}^{K} \widehat{Z}_{k,t}$ is an unbiased estimator of the normalizing constant Z_t.

Proof. For each island k, write

$$\widehat{\Gamma}_{k,t}^{M}(f) := \widehat{Z}_{k,t} \sum_{m=1}^{M} \bar{w}_t^{k,m} f(y_{1:t}^{k,m}),$$

so that $\widehat{\Gamma}_t^{K,M}(f) = K^{-1}\sum_{k=1}^{K} \widehat{\Gamma}_{k,t}^{M}(f)$.

Induction statement. We prove that, for every island k and every bounded test function f on the step-t prefix space,

$$\mathbb{E}\Big[\widehat{\Gamma}_{k,t}^{M}(f)\Big] = \Gamma_t(f) \tag{45}$$

by induction on t.

Base case ($t = 0$). Every island contains the empty prefix with normalized weights $1/M$ and $\widehat{Z}_{k,0} = 1$. Hence $\widehat{\Gamma}_{k,0}^{M}(f) = f(\varnothing) = \Gamma_0(f)$.

Inductive step ($t-1$ to t). Assume Eq. (45) holds at step $t-1$. Before optional resampling at step t, the particle-weight and island-mass recursions give

$$\begin{aligned}\widehat{\Gamma}_{k,t}^{M}(f) &= \widehat{Z}_{k,t-1} c_{k,t} \sum_{m=1}^{M} \frac{\bar{w}_{t-1}^{k,m} G_t^{k,m}}{c_{k,t}} f(y_{1:t}^{k,m}) \\ &= \widehat{Z}_{k,t-1} \sum_{m=1}^{M} \bar{w}_{t-1}^{k,m} G_t^{k,m} f(y_{1:t}^{k,m}).\end{aligned} \tag{46}$$

Thus the normalization factor $c_{k,t}$ cancels from the unnormalized island measure. For a bounded f, define the one-step target update

$$(\mathcal{T}_t f)(y_{<t}) := \sum_{v \in \mathcal{V}_t(y_{<t})} \frac{\varphi_t((y_{<t}, v))}{\varphi_{t-1}(y_{<t})} f((y_{<t}, v)).$$

This function is bounded on the finite set of reachable states. Conditioning on the complete population history and expanding the next-token draw yields

$$\begin{aligned}&\mathbb{E}\Big[\widehat{\Gamma}_{k,t}^{M}(f) \mid \mathcal{F}_{t-1}\Big] \\ &= \widehat{Z}_{k,t-1} \sum_{m=1}^{M} \bar{w}_{t-1}^{k,m} \sum_{v \in \mathcal{V}_t(y_{<t}^{k,m})} q_t^{k,m}(v \mid \mathcal{F}_{t-1}) G_t^{k,m}(v) f((y_{<t}^{k,m}, v)) \\ &= \widehat{Z}_{k,t-1} \sum_{m=1}^{M} \bar{w}_{t-1}^{k,m} (\mathcal{T}_t f)(y_{<t}^{k,m}).\end{aligned} \tag{47}$$

The last equality uses Eq. (42): the actual proposal cancels with the proposal denominator in the incremental importance ratio. Taking expectation again, applying the induction hypothesis to the bounded function $\mathcal{T}_t f$, and then expanding its definition gives

$$\begin{aligned}\mathbb{E}\Big[\widehat{\Gamma}^M_{k,t}(f)\Big] &= \mathbb{E}\left[\widehat{Z}_{k,t-1}\sum_{m=1}^{M}\bar{w}^{k,m}_{t-1}(\mathcal{T}_t f)(y^{k,m}_{<t})\right]\\ &= \Gamma_{t-1}(\mathcal{T}_t f)\\ &= \sum_{y_{<t}}\varphi_{t-1}(y_{<t})\sum_{v\in\mathcal{V}_t(y_{<t})}\frac{\varphi_t((y_{<t},v))}{\varphi_{t-1}(y_{<t})}f((y_{<t},v))\\ &= \sum_{y_{1:t}}\varphi_t(y_{1:t})f(y_{1:t}) = \Gamma_t(f). \end{aligned} \tag{48}$$

The first equality is the tower property together with Eq. (47); the second applies the induction hypothesis at step $t-1$. In the final equality, $\varphi_{t-1}(y_{<t})$ cancels and $(y_{<t}, v)$ ranges over all step-t states. The absorbing-EOS convention makes the same calculation valid for completed responses.

If resampling is not triggered, the weighted empirical measure is unchanged. If it is triggered, $\widehat{Z}_{k,t}$ remains unchanged and Eq. (41) gives

$$\mathbb{E}\left[\widehat{Z}_{k,t}\frac{1}{M}\sum_{m=1}^{M}f(y^{k,A^m_{k,t}}_{1:t})\,\middle|\,\mathcal{F}^-_t\right] = \widehat{Z}_{k,t}\sum_{m=1}^{M}\bar{w}^{k,m}_t f(y^{k,m}_{1:t}) = \widehat{\Gamma}^M_{k,t}(f).$$

Both cases therefore preserve the induction statement at step t, completing the induction. The proof conditions on the joint population history, so the adaptive proposal and resampling decisions are measurable and independence among islands is not required.

Averaging Eq. (45) over the K islands gives Eq. (44). Finally, take $f \equiv 1$. Since the normalized weights in each island sum to one,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\widehat{Z}_{k,t}\right] &= \mathbb{E}\Big[\widehat{\Gamma}^{K,M}_t(1)\Big]\\ &= \Gamma_t(1) = \sum_{y_{1:t}}\varphi_t(y_{1:t}) =: Z_t.\end{aligned} \tag{49}$$

This proves the normalizing-constant claim.

□

Proposition 2 ($M \to \infty$ consistency). Fix K, a finite horizon H, and an input (I, x). Before each next-token draw, $\mathcal{F}_{t-1}$ determines the actual proposal $q^{k,m}_t$ used by every particle. At routed steps, the proposal may depend on the particle population observed at the preceding routing checkpoint. Both the base and scout proposals are normalized over the same finite vocabulary $\mathcal{V}$ and assign positive probability wherever the corresponding target increment is positive. Exact score ties favor the larger island index, then the larger within-island particle index, and finally the larger region index. Under this deterministic ordering, the routing decision and hence the actual proposal are determined by $\mathcal{F}_{t-1}$. For the fixed input and finite H, the set of possible particle-local proposal states is finite and independent of M. It comprises active and EOS-padded response states, scout status, routed regions, and the cache states determined by these finite histories. All proposal hyperparameters are also fixed independently of M. Because each realized proposal is positive on the support of its target increment, the minimum positive proposal probability over these particle-local states is strictly positive. The incremental importance weights therefore have uniformly bounded second moments: there exists a finite constant C, independent of M and the realized population history, such that

$$\mathbb{E}\left[\Big(G^{k,m}_t(Y^{k,m}_t)\Big)^2 \mid \mathcal{F}_{t-1}\right] \le C, \qquad t \le H. \tag{50}$$

Conditional on $\mathcal{F}_{t-1}$, the particles use independent random draws. Let

$$Z_t := \sum_{y_{1:t}}\varphi_t(y_{1:t}), \qquad \pi_t := \frac{\varphi_t}{Z_t}.$$

At the terminal step, $\pi_H = \pi_\alpha^F$ and $Z_H = Z_\alpha^F$. Under these conditions, for every bounded f,

$$\sum_{k=1}^{K}\sum_{m=1}^{M} \widetilde{W}_H^{k,m} f(y^{k,m}) \xrightarrow[M\to\infty]{\mathsf{P}} \sum_{y} \pi_H(y) f(y). \tag{51}$$

Here $y^{k,m}$ denotes the completed terminal response of particle (k,m). In addition, the island normalizing-mass estimates satisfy

$$\frac{1}{K}\sum_{k=1}^{K} \widehat{Z}_{k,H} \xrightarrow[M\to\infty]{\mathsf{P}} Z_H = Z_\alpha^F.$$

Proof. We proceed by induction over token steps $t = 0, \dots, H$.

Induction statements. For each fixed island k, after completing token step t and any optional resampling, we prove the following three statements.

Statement 1 (empirical-measure consistency). For every bounded test function f on the prefix space at step t,

$$\sum_{m=1}^{M} \bar{w}_t^{k,m} f(y_{1:t}^{k,m}) \xrightarrow[M\to\infty]{\mathsf{P}} \sum_{y_{1:t}} \pi_t(y_{1:t}) f(y_{1:t}). \tag{52}$$

Statement 2 (weight control). The squared normalized-weight mass satisfies

$$S_{t,M}^k := \sum_{m=1}^{M} (\bar{w}_t^{k,m})^2, \qquad S_{t,M}^k = O_{\mathsf{P}}(M^{-1}). \tag{53}$$

Statement 3 (normalizing-mass consistency). The island estimate satisfies $\widehat{Z}_{k,t} \xrightarrow{\mathsf{P}} Z_t$.

Base case ($t = 0$). Every particle represents the empty prefix and has weight $1/M$, while π_0 is concentrated on that same empty prefix. Therefore the two sides of Statement 1 are identical, $S_{0,M}^k = M(1/M)^2 = 1/M$, and $\widehat{Z}_{k,0} = Z_0 = 1$. All three statements hold at $t = 0$.

Inductive step ($t-1$ to t). Assume all three statements hold at step $t-1$. Let $\bar{w}_{t-1}^{k,m}$ be the normalized weight carried by particle (k,m) into token step t, after any optional resampling at step $t-1$. We first establish the three statements after propagation and importance weighting, before optional resampling at step t.

(a) Shared corrected-propagation result. At step t, particle (k,m) draws $Y_t^{k,m} \sim q_t^{k,m}(\cdot \mid \mathcal{F}_{t-1})$. Write $G_t^{k,m}(v)$ for the incremental weight associated with candidate token v. After sampling, the realized incremental weight is $G_t^{k,m}(Y_t^{k,m})$. For any bounded test function f on the prefix space at step t, define $X_t^{k,m} := G_t^{k,m}(Y_t^{k,m}) f((y_{<t}^{k,m}, Y_t^{k,m}))$. Here $(y_{<t}^{k,m}, v)$ denotes the length-t state obtained by appending token v to an active prefix. Under the absorbing-EOS convention, it denotes the same completed response for a completed state. Conditioning on the complete population history gives

$$\begin{aligned}
&\mathbb{E}\Big[X_t^{k,m} \mid \mathcal{F}_{t-1}\Big] \\
&= \sum_{v\in\mathcal{V}_t(y_{<t}^{k,m})} q_t^{k,m}(v \mid \mathcal{F}_{t-1}) G_t^{k,m}(v) f((y_{<t}^{k,m}, v)) \\
&= \sum_{v\in\mathcal{V}_t(y_{<t}^{k,m})} q_t^{k,m}(v \mid \mathcal{F}_{t-1}) \frac{\varphi_t((y_{<t}^{k,m}, v))}{\varphi_{t-1}(y_{<t}^{k,m}) q_t^{k,m}(v \mid \mathcal{F}_{t-1})} f((y_{<t}^{k,m}, v)) \\
&= \sum_{v\in\mathcal{V}_t(y_{<t}^{k,m})} \frac{\varphi_t((y_{<t}^{k,m}, v))}{\varphi_{t-1}(y_{<t}^{k,m})} f((y_{<t}^{k,m}, v)).
\end{aligned} \tag{54}$$

The population-dependent proposal therefore cancels from the conditional target update. Consequently, the consistency argument does not require the route-selected proposal itself to converge as M grows. Define the propagated weighted sum before normalization

$$U^k_{t,M}(f) := \sum_{m=1}^{M} \bar{w}^{k,m}_{t-1} G^{k,m}_t(Y^{k,m}_t) f((y^{k,m}_{<t}, Y^{k,m}_t)). \tag{55}$$

Conditional mean. By the definition of $U^k_{t,M}(f)$, linearity of conditional expectation, and the proposal cancellation above,

$$\begin{aligned}
&\mathbb{E}\big[U^k_{t,M}(f) \mid \mathcal{F}_{t-1}\big] \\
&= \mathbb{E}\left[\sum_{m=1}^{M} \bar{w}^{k,m}_{t-1} G^{k,m}_t(Y^{k,m}_t) f((y^{k,m}_{<t}, Y^{k,m}_t)) \middle| \mathcal{F}_{t-1}\right] \\
&= \sum_{m=1}^{M} \bar{w}^{k,m}_{t-1} \mathbb{E}\Big[G^{k,m}_t(Y^{k,m}_t) f((y^{k,m}_{<t}, Y^{k,m}_t)) \mid \mathcal{F}_{t-1}\Big] \\
&= \sum_{m=1}^{M} \bar{w}^{k,m}_{t-1} \sum_{v \in \mathcal{V}_t(y^{k,m}_{<t})} q^{k,m}_t(v \mid \mathcal{F}_{t-1}) G^{k,m}_t(v) f((y^{k,m}_{<t}, v)) \\
&= \sum_{m=1}^{M} \bar{w}^{k,m}_{t-1} \sum_{v \in \mathcal{V}_t(y^{k,m}_{<t})} \frac{\varphi_t((y^{k,m}_{<t}, v))}{\varphi_{t-1}(y^{k,m}_{<t})} f((y^{k,m}_{<t}, v)).
\end{aligned} \tag{56}$$

The inner sum in the last line is a bounded function of the incoming prefix: f is bounded, $\mathcal{V}$ is finite, every reachable prefix has $\varphi_{t-1}(y_{<t}) > 0$, and a fixed finite horizon admits only finitely many reachable active and EOS-padded states. Statement 1 at step $t-1$ therefore applies to this inner sum. Substituting $\pi_{t-1}(y_{<t}) = \varphi_{t-1}(y_{<t})/Z_{t-1}$ gives

$$\begin{aligned}
&\mathbb{E}\big[U^k_{t,M}(f) \mid \mathcal{F}_{t-1}\big] \\
&\xrightarrow{\mathsf{P}} \sum_{y_{<t}} \pi_{t-1}(y_{<t}) \sum_{v \in \mathcal{V}_t(y_{<t})} \frac{\varphi_t((y_{<t}, v))}{\varphi_{t-1}(y_{<t})} f((y_{<t}, v)) \\
&= \frac{1}{Z_{t-1}} \sum_{y_{<t}} \sum_{v \in \mathcal{V}_t(y_{<t})} \varphi_t((y_{<t}, v)) f((y_{<t}, v)) \\
&= \frac{\Gamma_t(f)}{Z_{t-1}}.
\end{aligned} \tag{57}$$

The second line substitutes the induction hypothesis. The following equality uses the definition of π_{t-1} and cancels $\varphi_{t-1}(y_{<t})$. The double sum then equals $\Gamma_t(f)$ because $(y_{<t}, v)$ ranges over all length-t prefixes.

Vanishing propagation noise. Conditional independence, the uniform second-moment bound, and Statement 2 at step $t-1$ give

$$\begin{aligned}
&\mathrm{Var}\big[U^k_{t,M}(f) \mid \mathcal{F}_{t-1}\big] \\
&= \mathrm{Var}\left[\sum_{m=1}^{M} \bar{w}^{k,m}_{t-1} X^{k,m}_t \,\middle|\, \mathcal{F}_{t-1}\right] \\
&= \sum_{m=1}^{M} (\bar{w}^{k,m}_{t-1})^2 \,\mathrm{Var}\Big[X^{k,m}_t \mid \mathcal{F}_{t-1}\Big] \\
&\le \sum_{m=1}^{M} (\bar{w}^{k,m}_{t-1})^2 \mathbb{E}\Big[(X^{k,m}_t)^2 \mid \mathcal{F}_{t-1}\Big] \\
&= \sum_{m=1}^{M} (\bar{w}^{k,m}_{t-1})^2 \mathbb{E}\left[\left(G^{k,m}_t(Y^{k,m}_t)\right)^2 f((y^{k,m}_{<t}, Y^{k,m}_t))^2 \mid \mathcal{F}_{t-1}\right] \\
&\le C\|f\|^2_\infty \sum_{m=1}^{M} (\bar{w}^{k,m}_{t-1})^2 = C\|f\|^2_\infty S^k_{t-1,M} = O_{\mathsf{P}}(M^{-1}).
\end{aligned} \tag{58}$$

Here $\|f\|_\infty$ is the largest absolute value of f over length-t prefixes and is finite because f is bounded. For any $\varepsilon > 0$, conditional Chebyshev's inequality gives

$$\begin{aligned}
&\Pr\big(\big|U^k_{t,M}(f) - \mathbb{E}[U^k_{t,M}(f) \mid \mathcal{F}_{t-1}]\big| > \varepsilon \big| \mathcal{F}_{t-1}\big) \\
&\quad \le \frac{\mathrm{Var}[U^k_{t,M}(f) \mid \mathcal{F}_{t-1}]}{\varepsilon^2} \le \frac{C\|f\|^2_\infty}{\varepsilon^2} S^k_{t-1,M} \xrightarrow{\mathsf{P}} 0.
\end{aligned} \tag{59}$$

These conditional probabilities lie in $[0,1]$. Taking expectations and using the law of total probability therefore gives

$$U^k_{t,M}(f) - \mathbb{E}\big[U^k_{t,M}(f) \mid \mathcal{F}_{t-1}\big] \xrightarrow{\mathsf{P}} 0. \tag{60}$$

Thus the conditional mean approaches the correct target update, and the random fluctuation around that mean vanishes.

Combining Eq. (60) and Eq. (57) makes both contributions explicit:

$$\begin{aligned}
U^k_{t,M}(f) &= \big(U^k_{t,M}(f) - \mathbb{E}[U^k_{t,M}(f) \mid \mathcal{F}_{t-1}]\big) + \mathbb{E}[U^k_{t,M}(f) \mid \mathcal{F}_{t-1}] \\
&\xrightarrow{\mathsf{P}} 0 + \frac{\Gamma_t(f)}{Z_{t-1}} = \frac{\Gamma_t(f)}{Z_{t-1}}.
\end{aligned} \tag{61}$$

(b) Proof of Statement 1: empirical-measure consistency. Taking $f = 1$ in the shared propagation limit above gives

$$c_{k,t} := U^k_{t,M}(1) \xrightarrow{\mathsf{P}} \frac{Z_t}{Z_{t-1}} > 0. \tag{62}$$

The propagated normalized weights satisfy $\bar{w}^{k,m}_t = \bar{w}^{k,m}_{t-1} G^{k,m}_t(Y^{k,m}_t)/c_{k,t}$. Therefore,

$$\begin{aligned}
\sum_{m=1}^{M} \bar{w}^{k,m}_t f(y^{k,m}_{1:t}) &= \frac{U^k_{t,M}(f)}{c_{k,t}} \\
&\xrightarrow{\mathsf{P}} \frac{\Gamma_t(f)/Z_{t-1}}{Z_t/Z_{t-1}} = \sum_{y_{1:t}} \pi_t(y_{1:t}) f(y_{1:t}).
\end{aligned} \tag{63}$$

The numerator uses Eq. (61), and the positive denominator uses Eq. (62). This proves Statement 1 at step t before optional resampling.

(c) Proof of Statement 2: weight control. The uniform second-moment bound gives

$$\begin{aligned}
&\mathbb{E}\left[\sum_{m=1}^{M} (\bar{w}^{k,m}_{t-1})^2 \left(G^{k,m}_t(Y^{k,m}_t)\right)^2 \,\middle|\, \mathcal{F}_{t-1}\right] \\
&\quad \le C \sum_{m=1}^{M} (\bar{w}^{k,m}_{t-1})^2 = O_{\mathsf{P}}(M^{-1}).
\end{aligned} \tag{64}$$

Conditional Markov's inequality shows that the sum inside the expectation is itself $O_{\mathsf{P}}(M^{-1})$. Dividing it by $c_{k,t}^2$, whose limit is positive, gives

$$\sum_{m=1}^{M} (\bar{w}_t^{k,m})^2 = O_{\mathsf{P}}(M^{-1}). \tag{65}$$

This proves Statement 2 at step t before optional resampling.

(d) Proof of Statement 3: normalizing-mass consistency. By Statement 3 at step $t-1$, the island-mass recursion, and Eq. (62),

$$\widehat{Z}_{k,t} = \widehat{Z}_{k,t-1} c_{k,t} \xrightarrow{\mathsf{P}} Z_{t-1} \frac{Z_t}{Z_{t-1}} = Z_t. \tag{66}$$

This proves Statement 3 at step t before optional resampling.

(e) Preservation of Statements 1–3 under optional resampling. If resampling is not triggered, the particles and normalized weights are carried forward unchanged, so all three statements remain true. If it is triggered, stratified resampling draws island-local ancestors $A_{k,t}^{1:M}$ and assigns weight $1/M$ to every descendant. Conditional on the pre-resampling population $\mathcal{F}_t^-$,

$$\begin{aligned} \mathbb{E}\left[\frac{1}{M} \sum_{m=1}^{M} f(y_{1:t}^{k,A_{k,t}^m}) \middle| \mathcal{F}_t^- \right] &= \sum_{j=1}^{M} \bar{w}_t^{k,j} f(y_{1:t}^{k,j}), \\ \mathrm{Var}\left[\frac{1}{M} \sum_{m=1}^{M} f(y_{1:t}^{k,A_{k,t}^m}) \middle| \mathcal{F}_t^- \right] &\le \frac{\|f\|_\infty^2}{M}. \end{aligned} \tag{67}$$

Statement 1 is preserved because the first line is conditionally unbiased and the second, together with Chebyshev's inequality, makes the resampling error converge to zero in probability. **Statement 2** is preserved because the equal descendant weights have squared mass exactly $1/M$. **Statement 3** is preserved because resampling leaves $\widehat{Z}_{k,t}$ unchanged. The ESS trigger is determined by $\mathcal{F}_t^-$, so the argument covers both branches without requiring the trigger to stabilize as M grows. This completes the inductive step from $t-1$ to t, and hence the induction over $t = 0, \dots, H$.

Set $t = H$. Statements 1 and 2 give, for each fixed island k,

$$\sum_{m=1}^{M} \bar{w}_H^{k,m} f(y^{k,m}) \xrightarrow{\mathsf{P}} \sum_{y} \pi_H(y) f(y), \qquad \sum_{m=1}^{M} (\bar{w}_H^{k,m})^2 = O_{\mathsf{P}}(M^{-1}). \tag{68}$$

Statement 3 gives $\widehat{Z}_{k,H} \xrightarrow{\mathsf{P}} Z_H$.

Finally, substitute $\widetilde{W}_H^{k,m} = \widehat{Z}_{k,H} \bar{w}_H^{k,m} / \sum_{j=1}^{K} \widehat{Z}_{j,H}$. Because K is fixed, the per-island limits above imply

$$\begin{aligned} \sum_{k=1}^{K} \sum_{m=1}^{M} \widetilde{W}_H^{k,m} f(y^{k,m}) &= \frac{\sum_{k=1}^{K} \widehat{Z}_{k,H} \sum_{m=1}^{M} \bar{w}_H^{k,m} f(y^{k,m})}{\sum_{j=1}^{K} \widehat{Z}_{j,H}} \\ &\xrightarrow{\mathsf{P}} \frac{\sum_{k=1}^{K} Z_H \sum_y \pi_H(y) f(y)}{K Z_H} = \sum_y \pi_H(y) f(y). \end{aligned} \tag{69}$$

This proves Eq. (51). The same normalizer limit gives $K^{-1} \sum_{k=1}^{K} \widehat{Z}_{k,H} \xrightarrow{\mathsf{P}} Z_H = Z_\alpha^F$, which completes the proof.

□

E ANSWER-MARGINAL POWER READOUT

For the population-level analysis, write $\pi_\alpha(y) = \pi_\alpha^F(y \mid I, x)$ and suppress the fixed conditioning pair (I, x). Let $\mathrm{Ans} : \mathcal{Y}_H \to \mathcal{A}$ map a completed response to its canonical answer. The sequence-power

target induces

$$\mu_\alpha(a) = \sum_{y:\mathrm{Ans}(y)=a} \pi_\alpha(y), \tag{70}$$

and the corresponding finite answer power is

$$q_{\alpha,\gamma}(a) = \frac{\mu_\alpha(a)^\gamma}{\sum_b \mu_\alpha(b)^\gamma}, \qquad \gamma \geq 1. \tag{71}$$

These population quantities are approximated by Eqs. (13) and (14).

After drawing $a^\star \sim \widehat{q}_{\alpha,\gamma}$, the implementation returns a complete response by drawing one supporting terminal particle according to

$$\Pr(k, m \mid a^\star) = \frac{\widetilde{W}_H^{k,m} \mathbf{1}[\mathrm{Ans}(y^{k,m}) = a^\star]}{\widehat{\mu}_\alpha(a^\star)}. \tag{72}$$

For a terminal particle (k, m), let $a_{k,m} := \mathrm{Ans}(y^{k,m})$. This particle can be returned only when the first-stage answer draw selects $a_{k,m}$. Therefore,

$$\begin{aligned} \Pr(k, m) &= \Pr(a^\star = a_{k,m}) \Pr(k, m \mid a^\star = a_{k,m}) \\ &= \frac{\widehat{\mu}_\alpha(a_{k,m})^\gamma}{\sum_b \widehat{\mu}_\alpha(b)^\gamma} \frac{\widetilde{W}_H^{k,m}}{\widehat{\mu}_\alpha(a_{k,m})}. \end{aligned} \tag{73}$$

The indicator in Eq. (72) equals one by the definition of $a_{k,m}$. Cancelling one factor of $\widehat{\mu}_\alpha(a_{k,m})$ gives the induced terminal-particle distribution

$$\widehat{\pi}_{\alpha,\gamma}(k, m) = \frac{\widetilde{W}_H^{k,m} \widehat{\mu}_\alpha(\mathrm{Ans}(y^{k,m}))^{\gamma-1}}{\sum_b \widehat{\mu}_\alpha(b)^\gamma}. \tag{74}$$

The population analogue replaces the weighted terminal-particle measure by π_α. For a completed trajectory y, let $a_y := \mathrm{Ans}(y)$. At the population level, this gives

$$\begin{aligned} \pi_{\alpha,\gamma}^{\mathrm{2stage}}(y) &:= q_{\alpha,\gamma}(a_y) \pi_\alpha(y \mid \mathrm{Ans}(Y) = a_y) \\ &= \frac{\mu_\alpha(a_y)^\gamma}{\sum_b \mu_\alpha(b)^\gamma} \frac{\pi_\alpha(y)}{\mu_\alpha(a_y)} \\ &= \frac{\pi_\alpha(y) \mu_\alpha(\mathrm{Ans}(y))^{\gamma-1}}{\sum_b \mu_\alpha(b)^\gamma}. \end{aligned} \tag{75}$$

Empirically, Eq. (13) defines the pushforward of the weighted terminal particle measure through Ans. Both distributions recover their respective first-stage terminal measures when $\gamma = 1$. For $\gamma > 1$, trajectories supporting the same answer receive the same answer-mass multiplier, preserving their relative probabilities within that answer while favoring answers with larger marginal mass. This empirical readout does not alter the unbiased normalizing-constant statement in Eq. (44).

E.1 Split sharpening and cross-trajectory support

For an answer a, enumerate its trajectories as $\mathcal{Y}_a = \{y_{a,1}, y_{a,2}, \ldots\}$ and write $p_{a,i} = p_F(y_{a,i} \mid I, x)$. Common normalizing constants can be omitted when comparing answers. Applying one sequence exponent $\alpha\gamma$ gives the answer mass

$$D_{\alpha\gamma}(a) = \sum_i p_{a,i}^{\alpha\gamma}, \tag{76}$$

whereas splitting the powers across trajectories and answers gives

$$S_{\alpha,\gamma}(a) = \left(\sum_i p_{a,i}^\alpha \right)^\gamma. \tag{77}$$

For $\gamma = 2$,

$$S_{\alpha,2}(a) = D_{2\alpha}(a) + 2 \sum_{i<j} p_{a,i}^\alpha p_{a,j}^\alpha. \tag{78}$$

The excess over direct sequence power is therefore exactly the sum of same-answer cross terms. It is nonnegative and is zero only when at most one trajectory with positive mass supports the answer. Because this excess differs across answers, normalization can change their ranking. In particular, $(\alpha, \gamma) = (2, 2)$ retains the direct fourth-power term while also crediting collective support from multiple trajectories.

A separating two-token example. Following the two-token construction used to illustrate sequence-power lookahead (Karan & Du, 2025), let the first token choose a reasoning branch and the second choose a completion within that branch. Consider six complete trajectories with

$$\begin{aligned} p(a_1) = p(a_2) = p(a_3) &= 3/22, \\ p(b_1) = p(b_2) &= 4/22, \\ p(c_1) &= 5/22, \end{aligned} \tag{79}$$

where branches a, b, and c yield answers A, B, and C, respectively. This is a valid autoregressive model: the first-token probabilities are $(9, 8, 5)/22$, and completions are uniform within each branch. Under sequence-level $\alpha = 4$, the first-token and answer masses are obtained by summing the fourth powers of complete future trajectories:

$$D_4(A) : D_4(B) : D_4(C) = 3 \cdot 3^4 : 2 \cdot 4^4 : 5^4 = 243 : 512 : 625. \tag{80}$$

Thus weak continuations are suppressed before the first token is sampled. With the split construction, the sequence-square marginal is $\mu_2(A) : \mu_2(B) : \mu_2(C) = 27 : 32 : 25$, and the answer-level square gives $q_{2,2}(A) : q_{2,2}(B) : q_{2,2}(C) = 729 : 1024 : 625$.

Table 12: A two-token autoregressive example in which base sampling, split sharpening, and direct fourth-power sequence sampling have different answer modes.

	Base $\alpha = 1$	Split $(2, 2)$	Seq. $\alpha = 4$
A (3 weak)	$9/22$	$729/2378$	$243/1380$
B (2 moderate)	$8/22$	$1024/2378$	$512/1380$
C (1 strong)	$5/22$	$625/2378$	$625/1380$
Mode	A	B	C

Base sampling favors A because its three individually weakest trajectories carry the largest total mass. Direct sequence-level fourth power selects C, whose powered future mass is already largest at the first token. Split sharpening selects B: the trajectory square discounts weak paths while retaining collective support from two moderate paths, and the second square is applied only after their answer mass is combined.

The split therefore remains fourth-order in the base trajectory probabilities while avoiding both unweighted majority behavior and the pathwise concentration of direct fourth-power sequence sampling.

F Finite-Particle Behavior on Direct-Answer Tasks

MMStar-P consists of single-choice questions, and our instruction asks the model to end with `Final answer: ...`. Across four seeds, 1,854 of 2,000 selected outputs under $(\alpha, \gamma) = (2, 2)$ and 1,866 of 2,000 under $(4, 1)$ contain this marker. All 3,060 selected RealWorldQA outputs give a direct answer. These formats motivate a simplified analysis in which one generated token determines the answer.

Direct-answer model. We assume that each admissible answer a corresponds to one complete trajectory. The answer is either the first generated token or the token immediately following the fixed generated prefix `Final answer:`. We assume one fixed format for each question within a dataset, so the two cases never occur together in one particle population. Under the Qwen tokenization used here, `Final answer:` comprises the tokens `Final`, `answer`, and `:`; an option such as `A` is the next token. Thus the answer position is $t_* = 4$ in the fixed-prefix format and $t_* = 1$ in the immediate-answer format. We assume that the remaining suffix and termination are fixed and

contribute the same likelihood factor for every answer. Particle draws are independent, without scouting or resampling.

Let p_a be the base-model probability of selecting answer a, with $\sum_a p_a = 1$. At the answer position, the run with sequence exponent α draws from

$$Q_\alpha(a) = \frac{p_a^{\beta_{t_*}^{(\alpha)}}}{\sum_b p_b^{\beta_{t_*}^{(\alpha)}}}, \qquad \beta_t^{(\alpha)} = 1 + (\alpha - 1)\min\{t/128, 1\}. \tag{81}$$

For $t_* = 1$, the proposal exponents are $\beta_1^{(2)} = 1.0078125$ and $\beta_1^{(4)} = 1.0234375$. For $t_* = 4$, they are $\beta_4^{(2)} = 1.03125$ and $\beta_4^{(4)} = 1.09375$. These exponents determine the proposals, not the terminal targets. Importance correction and the updates after completion bring each trajectory to terminal exponent α. Since the other trajectory factors are common to all answers and cancel on normalization, its terminal importance weight is proportional to

$$W_\alpha(a) = \frac{p_a^\alpha}{Q_\alpha(a)}. \tag{82}$$

Finite-particle readouts. Let $n = KM = 32$ and let $N_a^{(\alpha)}$ be the number of particles selecting a in a run with exponent α. Direct sequence power $(4, 1)$ assigns a the probability

$$\widehat{P}_{4,1}(a) = \frac{N_a^{(4)} W_4(a)}{\sum_b N_b^{(4)} W_4(b)}. \tag{83}$$

The two-stage $(2, 2)$ readout first aggregates particles by answer and then squares each aggregate:

$$\widehat{P}_{2,2}(a) = \frac{\left(N_a^{(2)} W_2(a)\right)^2}{\sum_b \left(N_b^{(2)} W_2(b)\right)^2}. \tag{84}$$

As $n \to \infty$, $N_a^{(\alpha)}/n \to Q_\alpha(a)$. Both readouts therefore converge to $p_a^4 / \sum_b p_b^4$. At finite n, however, the second-stage power acts on an estimated answer mass. Since $N_a^{(2)} \sim \mathrm{Binomial}(n, Q_2(a))$, its mean is $nQ_2(a)$ and its variance is $nQ_2(a)(1 - Q_2(a))$. Applying $\mathbb{E}[N^2] = \mathrm{Var}(N) + \mathbb{E}[N]^2$ gives

$$\begin{aligned} \mathbb{E}\left[\left(N_a^{(2)} W_2(a)\right)^2\right] &= \frac{p_a^4}{Q_2(a)^2}\left[nQ_2(a)(1 - Q_2(a)) + n^2 Q_2(a)^2\right] \\ &= n^2 p_a^4 + np_a^4\left(\frac{1}{Q_2(a)} - 1\right). \end{aligned} \tag{85}$$

The second term comes from finite-particle count variation. To obtain expected pass@1, the answer masses must be normalized within each realized population before averaging over populations.

Exact finite-particle example. Consider four answers with $p = (0.4, 0.3, 0.2, 0.1)$, where the highest-probability answer A is correct. A population is described by nonnegative counts $\mathbf{n} = (n_A, n_B, n_C, n_D)$ summing to 32. There are $\binom{32+3}{3} = \binom{35}{3} = 6{,}545$ such vectors: distributing 32 selections among four answers is equivalent to placing three separators among 35 positions. Under proposal Q_α, the probability of a count vector is

$$\Pr_\alpha(\mathbf{n}) = \frac{32!}{n_A! n_B! n_C! n_D!} \prod_{a \in \{A,B,C,D\}} Q_\alpha(a)^{n_a}. \tag{86}$$

For each vector, we insert its counts into Eqs. (83) and (84) to obtain the conditional probability of selecting A, multiply by Eq. (86), and sum over all 6,545 vectors:

$$\mathbb{E}[\widehat{P}_{\alpha,\gamma}(A)] = \sum_{\substack{\mathbf{n} \geq 0 \\ n_A + n_B + n_C + n_D = 32}} \Pr_\alpha(\mathbf{n})\, \widehat{P}_{\alpha,\gamma}(A \mid \mathbf{n}), \quad (\alpha, \gamma) \in \{(4, 1), (2, 2)\}. \tag{87}$$

The two runs use their respective proposals Q_4 and Q_2. Evaluating this finite sum gives Tab. 13.

As n grows, both readouts converge to $0.4^4/(0.4^4 + 0.3^4 + 0.2^4 + 0.1^4) = 72.32\%$. This idealized direct-answer case illustrates a finite-particle mechanism relevant to MMStar-P and RealWorldQA: when each answer has only one trajectory, aggregation supplies no cross-trajectory support, while the $(2, 2)$ readout squares noisy empirical answer masses. The exact calculation shows how direct $(4, 1)$ can therefore have higher expected pass@1 in this setting.

Table 13: Expected pass@1 in the 32-particle direct-answer example with $p = (0.4, 0.3, 0.2, 0.1)$ and correct answer A.

Answer position	Direct $(4, 1)$	Two-stage $(2, 2)$
First generated token, $t_* = 1$	71.59%	69.12%
After `Final answer:`, $t_* = 4$	71.62%	69.14%

G Inference algorithm

Algorithm 1 summarizes the complete inference procedure, from island-local SMC and visual scouting to answer-marginal sampling.

H Computational cost

Table 14 reports measured end-to-end latency and peak memory for base sampling, Power-SMC, and ReSight-SMC under a common hardware and software setup.

Table 14: End-to-end inference cost with Qwen2.5-VL-7B-Instruct on one RTX 5090. Latency and peak memory are question-weighted over LogicVista, MathVista, and MMStar-R for seed 0. All systems use the Transformers backend.

System	Population	Latency (s)	Peak VRAM (GB)
Base sampling	1	4.11	15.61
Power-SMC	1×32	7.85	17.27
ReSight-SMC	4×8	10.37	17.35

In the worst case, when all KM trajectories remain active for H decoding steps, Island-SMC requires

$$T_F = \mathcal{O}(KMHC_F^{\text{step}}). \tag{88}$$

The visual phase creates at most B temporary scout decoder states. Let $N_I := |\mathcal{V}_I|$ be the number of original-image visual tokens, N_h the number of heads, d_h the per-head key dimension, and C_F^{step} one full-context decoder step. Its additional computation is

$$T_A = \mathcal{O}\big(KMN_hN_Id_h + BKMG + B(L_{\text{vis}} + 1)C_F^{\text{step}}\big)\,. \tag{89}$$

The first term computes prefix-to-image relevance at the routing layer, the second performs greedy routing, and the final term covers one-token replay plus the bounded scout episode. Image keys and particle caches come from the initial prefill, and all temporary states are released after the visual phase.

Additional peak state memory is

$$S_A = \mathcal{O}(BL(P_F + \tau_{\text{vis}} + L_{\text{vis}})d_{\text{kv}} + BN_I)\,, \tag{90}$$

where L is the number of decoder layers, P_F the original multimodal-prefix length, and d_{kv} the per-layer key–value width. If $U \le KM$ distinct canonical answers occur, terminal aggregation and sampling require $\mathcal{O}(KM)$ time and $\mathcal{O}(U)$ memory, with no additional model forward.

I Experimental details

The main configuration uses $H = 1024$, $\alpha = 2$, and

$$\beta_t = 1 + (\alpha - 1)\min\{t/128, 1\}, \qquad \beta_H = \alpha. \tag{91}$$

The base sampler uses temperature 1. The low-temperature baseline draws one autoregressive response with temperature $1/\alpha = 0.5$ and otherwise uses the same prompt, image preprocessing,

Algorithm 1 ReSight-SMC

Require: Image I, prompt x, horizon H, islands K, particles M

Initialize KM base-proposal particles and a multiscale region bank
Cache the image-token keys at the routing layer
Set the active scout set $\mathcal{S} \leftarrow \varnothing$
for $t = 1, \ldots, H$ **do**
  for all unfinished particles (k, m) **do**
    if $(k, m) \in \mathcal{S}$ **then**
      Sample the next token from $q_t^{A,k,m}$ in Eq. (11)
      Update the base-model likelihood and importance weight
      if the sampled token is nonterminal **then**
        Advance the persistent base target state with that token
        if another scout proposal is required **then**
          Advance the temporary attention state with that token
        end if
      end if
    else
      Sample the next token from q_t^F
      Update the base-model likelihood and importance weight
      if the sampled token is nonterminal **then**
        Advance the persistent base target state with that token
      end if
    end if
  end for
  Advance the bridge for completed trajectories
  Normalize weights and update $\widehat{Z}_{k,t}$ within each island
  if $t = \tau_{\text{vis}} + L_{\text{vis}}$ **then**
    Release the temporary attention states and set $\mathcal{S} \leftarrow \varnothing$
  end if
  if t is an SMC checkpoint **then**
    for $k = 1, \ldots, K$ **do**
      Compute $\mathrm{ESS}_{k,t}$
      if $\mathrm{ESS}_{k,t} < \rho_w M$ **then**
        Resample only island k and reset its weights to $1/M$
      end if
    end for
  end if
  if $t = \tau_{\text{vis}}$ **then**
    Compute the per-island quotas B_k and total budget B from Eq. (34)
    Score eligible particle–region pairs using Eq. (8)
    Select B pairs using Eq. (9)
    Fork their decoder states, replay y_t under Eq. (10), and set $\mathcal{S}$ to the selected particles
  end if
end for
Form terminal masses $\widetilde{W}_H^{k,m}$ using Eq. (6)
Aggregate $\widehat{\mu}_\alpha$ by canonical answer using Eq. (13)
Draw $a^\star \sim \widehat{q}_{\alpha,\gamma}$ using Eq. (14)
Draw a supporting particle conditional on $a^\star$ using Eq. (72) and return its response

response horizon, and full-vocabulary sampling rule. Both use $\mathrm{top_p} = 1$ and $\mathrm{top_k} = 0$. The low-temperature baseline does not apply sequence-level importance weights or resampling. Power-SMC uses one population of 32 particles with global resampling. The island-only variant and ReSight-SMC use $K = 4$ islands with $M = 8$ particles, stratified resampling with threshold $\rho_w = 0.5$, and checkpoint interval $L_{\mathrm{SMC}} = 32$. Every triggered resampling operation is confined to one island.

The visual configuration is

$$\rho_V = 0.25, \quad \tau_{\text{vis}} = 40, \quad L_{\text{vis}} = 16, \quad \lambda_I = \log 2, \quad \lambda_R = \log 4. \tag{92}$$

Thus the default population assigns two scouts to each fully active island, for at most eight scouts in total, and retains base-proposal anchors in every island. Visual scouting is invoked once after token 40: tokens 41–56 use the scout proposals, and tokens 57–64 return to the base proposal before the token-64 resampling checkpoint. The region bank contains 2×2 and 3×3 grid cells together with horizontal and vertical thirds. We set the area exponent to $\zeta = 0.75$ and the overlap coefficient to $\mu = 1$. Routing uses the final decoder layer. The attention bias is applied at every decoder layer and shared across heads. With these values, attention odds for every image token are doubled relative to unmodified keys, and tokens inside the routed region receive a further factor of four.

We fix prompting by benchmark category. MathVista, MMStar-R, and LogicVista use CoT prompting. The perception-focused MMStar-P and RealWorldQA use direct answering. All five benchmarks use $\tau_{\text{vis}} = 40$ and retain the same scout fraction, episode length, attention biases, region bank, and particle budget. A response that terminates before token 40 simply skips the scout stage.

The main ReSight-SMC configuration uses answer exponent $\gamma = 2$. Power-SMC, the island-only variant, and the trajectory-stage ReSight ablation use $\gamma = 1$. Terminal answers are grouped by the reference-free canonicalizer detailed in Appendix I.1. The returned reasoning text comes from a terminal particle that supports the sampled canonical answer.

The component ablation compares $K = 1, M = 32$ with $K = 4, M = 8$ under identical base proposals, bridge schedules, ESS thresholds, and total particle counts. It then adds the complete visual proposal stage and the answer-marginal readout in turn.

The backbone suite contains Qwen2.5-VL-3B-Instruct, Qwen2.5-VL-7B-Instruct, Qwen3-VL-4B-Instruct, and Qwen3-VL-8B-Instruct. The Qwen2.5-VL post-training references are `maveryn/trace-qwen2.5-vl-3b` and `maveryn/trace-qwen2.5-vl-7b` (Alam, 2026), trained with GRPO on the 64K-instance Trace corpus. At 7B we additionally evaluate Game-RL-Qwen2.5-VL-7B (Tong et al., 2025), trained with GRPO on GameQA. Each checkpoint uses the same benchmark-specific prompt, response horizon, image preprocessing, and answer evaluator as its corresponding training-free backbone. Ablations and appendix analyses use Qwen2.5-VL-7B-Instruct.

I.1 Prompts and answer evaluation

Prompt protocol. All compared methods use each checkpoint's native chat template without an additional system message. The user turn contains the image, the benchmark-released question, and, where applicable, a fixed generation instruction. MathVista, LogicVista, and MMStar-R append `Think step by step and end with 'Final answer: ...'.`, whereas MMStar-P appends `Answer directly and end with 'Final answer: ...'.`. RealWorldQA already includes a direct-answer instruction in its released question, so no additional suffix is appended. Base sampling, low-temperature sampling, Power-SMC, ReSight-SMC, and the post-trained checkpoints use the same dataset-specific prompt.

Deterministic answer extraction. The shared parser applies the following priority order: the last output line that parses as a JSON object with an `answer` field; the last span introduced by `Final answer:`, `Answer:`, or their Chinese equivalent; the last `\boxed{...}` expression; and finally the last nonempty output line. Textual comparisons apply Unicode NFKC normalization, lowercasing, Unicode-minus normalization, removal of common surrounding punctuation and LaTeX `\text{}` or `\mathrm{}` wrappers, and whitespace normalization. No LLM judge or LLM-based answer extractor is used.

Dataset-specific scoring. We use the 1,500 questions in the `Lin-Chen/MMStar` `val` split and partition them by the released `category` field. MMStar-R contains instance reasoning, logical reasoning, math, and science and technology; MMStar-P contains coarse perception and fine-grained perception (250 questions per category). The subsets are disjoint and cover the full split. Cross-subset comparisons reflect both category and prompting differences.

For MathVista, we use the 1,000-example `testmini` split and answer-type-aware normalization. Multiple-choice predictions are mapped from an option label or generated option text to the listed choice; integer answers are converted numerically, floating-point answers are rounded to the dataset-provided precision, and list and other free-form answers use normalized exact match. This follows

MathVista's answer-type-aware normalization while replacing its optional LLM answer-extraction stage with the deterministic parser above. MMStar-R and MMStar-P use the same option-letter exact-match evaluator, and a generated option string is mapped back to its listed letter.

RealWorldQA (xAI, 2024) uses its released 765-question test set, which combines multiple-choice and short-answer questions over real-world scenes. We retain the released question text and direct-answer instructions. Multiple-choice responses are scored by extracting the option letter; other responses use case-insensitive exact match after removing a final period. The visual ablation uses this protocol for both rows.

LogicVista (Xiao et al., 2024) uses its released 448-question test split and reports accuracy over the five reasoning skills. We extract the final option labels and require exact equality with the reference set; answer order is ignored for multi-select questions.

Reference-free answer grouping. Before answer-marginal aggregation, the same deterministic extraction rules canonicalize every terminal particle. Canonicalization may use the dataset identifier, question type, and listed choices, but it never accesses the reference answer. Terminal masses are grouped by these canonical strings, powered, and sampled as described in Appendix E. Only the sampled answer is then passed to the benchmark-specific evaluator.

Numerical evaluation. Proposal distributions and target factors are formed by casting logits to FP32 before a full-vocabulary log-softmax. Incremental and cumulative weights remain in the log domain. The target-preserving SMC path applies no top-k or nucleus truncation, retaining the support required by the importance correction.

Pass@1 averages four independently seeded complete executions with seeds $\{0, 1, 2, 3\}$. Pass@4 uses the same four executions. Coverage@32 uses the terminal particle set from one execution and is reported separately because those candidates share ancestry. All timing measurements include image processing, visual encoding, decoder prefill, token generation, resampling, visual routing, and output selection.