# Imagination as Metered Computation: Value-of-Information Control of a Band-Structured Latent Simulation Workspace

Andrew Kiruluta
UC Berkeley, School of Information
kiruluta@berkeley.edu
https://github.com/andrew-jeremy/Imagination-vLLM

August 2026

**Abstract**

Language models increasingly spend inference-time compute on deliberation, but that compute is almost entirely spent in one representational format: token sequences. A growing body of work therefore inserts generated images into the reasoning loop. We argue that the interesting scientific question is not *whether* a model can think with generated pictures — several systems already do — but *under what accounting* such a step is ever worth taking, and *what structure* the internal simulation should have so that the compute it consumes buys the most usable evidence.

This paper reformulates internal imagery as a metered information-acquisition problem. We first establish a negative result that we believe should discipline the whole subarea: because an imagined latent is generated by the reasoner itself, it forms a Markov chain with the state that produced it, so it carries *zero* additional Shannon information about the answer. Any measured benefit is therefore purely *computational* — it makes already-encoded information accessible to a bounded decoder. We formalise this as *computational value of information* (cVOI), prove that cVOI is non-negative under free disposal while the value *realised* by a fixed, over-trusting reader can be strictly negative, and derive a budgeted stopping rule in which the compute penalty is exactly the shadow price of the deliberation budget.

This accounting then *derives*, rather than assumes, the architecture. Treating the imaginal pathway as a rate-limited channel and applying task-weighted reverse water-filling yields an optimal allocation of representational precision across radial frequency bands. We show that a phase-preserving, strictly invertible spectral gate implements this allocation over the range its bounded parameterisation can express, that the resulting recurrent workspace is a diagonal linear time-invariant system in the Fourier basis, and that giving each band its own decay constant produces a principled separation between persistent coarse layout and volatile fine detail. We prove invertibility, bi-Lipschitz stability and a bounded condition number for the operator, and, for non-negative critic scores, a $(1 - 1/e)$ budgeted-greedy guarantee for counterfactual branch selection via submodularity of expected-max scores.

We contribute Vivid, an architecture and reference implementation instantiating this account: a frozen pretrained language model, latent diffusion generator and vision encoder, coupled by lightweight bridges to a band-structured recurrent workspace, a cVOI-regressing budgeted gate, a calibration-constrained evidence critic that cannot let imagined evidence inflate confidence beyond its measured reliability, and a provenance-typed audit trace. We further define a verified recursive-improvement protocol — offline self-distillation behind immutable regression gates with sequential testing and rollback — and analyse when it is expected to fail.

Finally, we specify a falsification programme whose central comparison is an *equal-compute* generated-image baseline, and add three tests that follow directly from the theory and that prior

work has not run: whether learned band gains track independently measured per-band task sensitivity, whether the gate's predicted value is calibrated against realised gain, and whether benefit decays as the reader's own budget grows.

We run these tests on a controlled synthetic testbed in which every quantity the theory names is observable by construction, and report eight outcomes: four supported, one weakly, three not. Learned band gains track measured sensitivity on the texture-dominated family — positive slope at every operator range, rising with that range, reaching 35–61% of what the bounded operator can express — but *not* on the layout-dominated one, where the slope is indistinguishable from zero. A gate regressing realised gain learns the value structure cleanly (+0.47 on the family imagery helps, $\approx 0$ on the two it does not) yet fails our own abstention criterion, because thresholding gain at zero is not the priced rule the paper specifies. Alignment before fusion cuts registered reconstruction error 6.5×, and the advantage of imagery decays as the reader's capacity grows, as the computational-benefit account requires. The spectral workspace did *not* separate from an equal-compute generic-imagery condition. These are pilots on a synthetic world with no pretrained backbones: they are evidence about the mechanism, not about any benchmark. **No benchmark numbers are reported anywhere, and the results table for public benchmarks remains an explicitly empty template.**

**Naming and relation to the earlier draft.** This manuscript supersedes an earlier proposal circulated as vLLM-SI. The name is retired for two reasons: it collided with the unrelated `vllm` inference server, and, more importantly, it named a mechanism ("spectral imagination") rather than a claim. VIVID — *Value-of-Information Visual Imagination and Deliberation* — names the claim, which is that internal imagery should be treated as metered computation. The spectral workspace survives, but it is now a consequence of the rate-allocation argument in §3 rather than a stipulated inductive bias.

# 1 Introduction

## 1.1 Deliberation has a format problem

Contemporary language models improve markedly when they are allowed to spend more computation before committing to an answer. Chain-of-thought prompting [46], best-of-$n$ and verifier-guided search [43], and reinforcement-learned long-form reasoning [14] all exploit the same resource: extra forward passes spent restructuring a problem before answering.

Almost all of that computation is spent in a single representational format. The intermediate states are token sequences, and token sequences are a poor medium for some of the operations reasoning actually requires. Predicting what a folded net looks like when assembled, deciding whether two routes cross, tracking which object is occluded after a viewpoint change, or asking what a plotted curve does between two sampled points are all problems where the natural intermediate object is a configuration in a metric space, not a string. Serialising such an object into language costs tokens, loses precision, and — critically — forces the model to *compute* relations that a perceptual encoder would simply *read off*.

Human cognition does not appear to face this restriction. Visual mental imagery recruits early visual cortex, behaves in many respects like weak perception, and supports operations such as mental rotation whose latency scales with the angular disparity of the transformation [41, 35, 23]. Individual differences are large — aphantasic individuals report absent voluntary imagery yet solve many spatial tasks by other routes [51] — which is itself informative: imagery looks like an *optional accelerator* rather than a mandatory stage.

### 1.2 The subarea's real open question

Several systems now place generated images inside the reasoning loop. Visual chain-of-thought introduced multimodal infillings [38]; *Thinking with Generated Images* produces and self-critiques intermediate visual thoughts [3]; *Visual Thoughts* gives a unified account of multimodal chain-of-thought [2]; *Latent Sketchpad* keeps an internal visual scratchpad [53]; *MindJourney* couples a vision–language model to a diffusion world model for spatial test-time scaling [48]; and more recent work optimises interleaved text–image reasoning trajectories directly [31, 20]. Purely latent deliberation — continuous thought vectors [17] and recurrent-depth latent reasoning [9] — attacks the same format problem without images at all.

Against that background, "a model can construct and inspect internal images" is no longer a contribution. Nor, we think, is any particular bundle of engineering ingredients. An earlier version of this work listed six combined components and asserted that the combination was novel. That is a weak claim, and reviewers are right to treat such lists as incremental. The refactor presented here replaces the list with two questions that we believe are genuinely open:

1. **What is the correct accounting?** An imagined image is not evidence in the ordinary sense — nothing new was observed. Under what formal notion, then, can imagining ever be *rational*, and what is the exact price of a diffusion call in units of expected task reward?

2. **What structure should the simulation have?** If internal imagery is a rate-limited channel rather than a display, its representation should be chosen to maximise usable evidence per unit of compute — which is an allocation problem, and allocation problems have solutions.

### 1.3 A negative result that reframes the area

We begin from an observation that is elementary but, as far as we can tell, not stated in this literature. Let $Y$ be the answer and $S_t = (h_t, m_t, \mathcal{P}_t)$ the reasoner's full internal state at step $t$ — the language summary, the imaginal workspace and the provenance ledger. The imagined latent $Z_t$ is sampled from a generator conditioned only on $S_t$ (through a bridge), with sampling noise independent of $Y$. Then $Y \to S_t \to Z_t$ is a Markov chain, so $Z_t$ is conditionally independent of $Y$ and

$$I(Y; Z_t \mid S_t) = 0, \tag{1}$$

with the data-processing inequality extending this to every function of $Z_t$ — the transformed latent, the decoded image, the re-encoded features.

*Imagination adds no information.* Every bit that an imagined scene appears to supply was already present in the state that generated it. We call this the **no-free-evidence** property (Proposition 6), and it has three consequences that shape the rest of the paper.

First, it tells us what a positive result would even mean. Any measured gain must be *computational*: the imagined latent, once re-encoded by a vision encoder, presents already-held information in a form a *bounded* decoder can use in fewer operations. We therefore replace classical value of information [18] with a *computational* value of information defined relative to a bounded decoder class (Definition 2), in the spirit of bounded-optimal agents [39], resource-rational analysis [26], information-theoretic bounded rationality [34] and the expected value of control [40].

Second, it predicts when the effect should vanish. If the benefit is computational, it must shrink as the reader's own budget grows. A method whose advantage over a text-only baseline is *independent* of the baseline's token budget is almost certainly measuring capacity, not mechanism. We turn this into an explicit experiment (§8.3.3).

Third, it explains the field's characteristic failure mode. Classical VOI is non-negative only for a Bayes-rational decision maker. For a bounded, imperfectly calibrated reader, consuming self-generated evidence can *reduce* expected reward (Proposition 8) — which is precisely the observed pathology in which a model hallucinates a scene, re-perceives it, and becomes more confident without becoming more correct. The remedy is not a better image; it is a calibration constraint on how imagined evidence may move the posterior (§3.8).

### 1.4 From accounting to architecture

The second question — what structure the simulation should have — is where the accounting starts paying for itself. Suppose the imaginal pathway is a channel with a finite rate: a diffusion call at a chosen resolution and step count can only convey so much detail before the vision encoder's own bottleneck discards it. Then the design problem is: *across which degrees of freedom should the available precision be spent?*

For natural images the classical answer is spatial frequency. Image statistics are strongly non-white, scale-space theory identifies smooth multi-scale decompositions as the canonical uncommitted front end [27], and rate–distortion theory tells us that under a squared-error criterion the optimal allocation of rate across independent Gaussian components is *reverse water-filling* [4]: components with variance below a threshold are not coded at all, and the rest are coded to a common distortion floor.

Our contribution is to make this allocation *task-weighted* and *learned*. Weighting each band's distortion by the sensitivity of the downstream task loss to that band yields a modified water-filling solution (Proposition 11) in which the optimal per-band precision is proportional to task sensitivity. We then show that a real-valued, strictly positive radial gate applied in the Fourier domain implements exactly this reallocation, is phase-preserving and shift-equivariant, is *invertible* with a bounded condition number, and has provably bounded distortion (Propositions 16–18). The invertibility matters: it is the architectural echo of no-free-evidence. The operator provably cannot destroy or create information; it can only change what is easy to read.

Finally, making the recurrent workspace band-diagonal in the Fourier basis gives every band its own memory constant. A coarse-layout band with decay $\lambda \to 1$ persists across many reasoning steps; a texture band with small $\lambda$ is refreshed each step. This is a single-line change from the scalar-decay memory of the earlier draft, but it is the difference between an unmotivated leaky average and a structured linear dynamical system with per-band time constants (Proposition 20) — and it makes a cognitively meaningful prediction: gist should outlive detail in the workspace, and forcing it not to should selectively damage transformation-heavy tasks.

### 1.5 Contributions

1. **A formal account of self-generated evidence.** The no-free-evidence property (Prop. 6); computational value of information relative to a bounded decoder class (Def. 2); non-negativity under free disposal (Prop. 7) alongside strictly negative *realised* value for an over-trusting reader (Prop. 8); and a budgeted stopping rule whose compute penalty $\beta$ is the shadow price of the deliberation budget (Prop. 10), with a dual-ascent procedure for setting it.

2. **A derivation of the band-structured workspace.** Task-weighted reverse water-filling for the imaginal channel (Prop. 11); the equivalence between its solution and a learned radial spectral gate (Cor. 12); phase preservation, exact shift-equivariance and approximate rotation-equivariance (Props. 16, 17); invertibility, bi-Lipschitz stability and condition number $\leq (1+\rho\bar{a})/(1-\rho\bar{a})$ (Prop. 18).

3. **Band-wise recurrent memory.** The workspace as a diagonal LTI system in the Fourier basis, with per-band decay constants, stability condition, and closed-form effective memory horizon (Prop. 20). This replaces the scalar leaky average of the earlier draft.

4. **A guarantee for counterfactual branching.** Expected-max critic value is non-negative, monotone and submodular, so budgeted greedy branch selection attains a $(1 - 1/e)$ approximation (Prop. 25) — and the same fact makes the selected set's value provably concave in the branch count, a useful sanity check on inference-time scaling claims.

5. **Calibration-constrained evidence fusion.** A provenance-typed reader whose confidence conditional on imagined-only evidence is capped by the measured reliability of imagined evidence, with an *Imaginal Calibration Gap* metric and an admissibility criterion (§3.8).

6. **A verified recursive-improvement protocol** with immutable gates, sequential testing with explicit error control, replay against reference data, diversity monitoring, and rollback — plus an analysis of the conditions under which it degenerates (§6), connected to known results on training with recursively generated data [42].

7. **A falsification programme** centred on equal-compute baselines, plus three theory-driven tests absent from prior work: band-gain/sensitivity correspondence, cVOI calibration, and the budget-sweep decay prediction (§8).

8. **Latent transport before fusion.** Unregistered fusion of jittered views multiplies the spectrum by the offset distribution's characteristic function — a low-pass filter whose cutoff falls with the jitter and compounds over steps (Prop. 23). We estimate a low-parameter alignment in closed form and accept it only when it measurably reduces residual, with the acceptance margin chosen from a measured trade-off curve rather than asserted.

9. **Priced band allocation that does not collapse.** The controller chooses which bands to spend on, against the same shadow price, with an initialisation and an entropy schedule that prevent the default-to-all-bands failure, and with allocation entropy logged so that collapse is observable (§4.6).

10. **Preliminary measurements** on a controlled synthetic testbed where every quantity the theory names is observable by construction (§7), plus measured per-component costs. Of eight testable statements, four are supported, one weakly, and three are not — including the central mechanism claim. The scorecard is in Table 10.

11. **A reference implementation** of the control loop, spectral operator, band-wise memory, transport, cVOI gate with dual ascent, band allocator, calibrated critic, provenance-typed trace, diagnostic task generators and the pilot harness, runnable on CPU (§B, §F).

### 1.6 What this paper does not claim

We state this plainly because the subarea is crowded and easy to overclaim in.

- We report **no benchmark results**. Table 13 is an empty template. Every performance question in §8 is open. The pilots of §7 are run on a synthetic world with no pretrained backbones; they test whether the mechanism behaves as the theory says under conditions where the theory's assumptions hold, and they transfer to no benchmark claim.
- We do not claim that reasoning with generated images is new. It is not; see §12.

- We do not claim that a spectral operator is guaranteed to help. We claim it is the operator implied by a rate-allocation argument, and we specify the ablations that would show the argument is wrong.
- We do not claim the theory settles anything empirical. Propositions 6 and 8 are constraints on *interpretation*; they tell us what a gain could and could not mean, not whether a gain exists.
- "Recursive self-improvement" here means offline, gated, human-supervised self-distillation with rollback. It does not mean autonomous code modification or unrestricted online weight updates, and the protocol is designed to make that distinction enforceable rather than rhetorical.

### 1.7 Roadmap

§2 formalises reasoning as sequential information acquisition under a budget and defines cVOI. §3 contains the theory: no-free-evidence, water-filling, the spectral operator's properties, band-wise memory, submodular branching, and the calibration constraint. §4 specifies VIVID. §5 gives the staged training programme. §6 gives the verified improvement protocol. §8 gives the evaluation and §9 the falsification conditions. §10 treats risks, §11 non-visual modalities, §12 related work. Appendices contain proofs (A), the implementation (B), hyperparameters (C), diagnostic generators (D) and a reporting checklist (E).

## 2 Problem Formulation: Deliberation as Metered Information Acquisition

This section sets up the objects the rest of the paper reasons about. The aim is to make "should the model imagine now?" a well-posed decision problem rather than a heuristic, and to make the price of a diffusion call an explicit, measurable quantity.

### 2.1 Notation

The symbols that carry structure, with their shapes and the operators that act on them. Elementwise product is written $\odot$; all products between a gate and a spectrum are elementwise, broadcast over the batch axis and over channels where the gate has no channel index.

Two conventions to fix before the table. *$\mathcal{F}$ denotes the full unitary DFT*, for which Parseval holds exactly ($\|z\|_2 = \|\mathcal{F}z\|_2$) — this is what every norm bound in §3 is stated through. The implementation uses the real half-spectrum `rfft2`, which is the same operator restricted to the non-redundant half; Parseval on that half requires the conjugate-symmetric counterpart to be counted, so a direct norm comparison against the half-spectrum will not match, and the code's guarantees are asserted in the signal domain for that reason. And a handful of letters are reused across contexts in the standard way ($\eta$ as a step size, as the critic's parameters $Q_\eta$ and as the transport margin; $\nu$ as the encoder's parameters $V_\nu$ and as regulariser weights); these are disambiguated by subscript or by their argument at each use.

### 2.2 The deliberation process

A task instance is a pair $(x, y)$ with input $x \in \mathcal{X}$ (text, optionally accompanied by genuinely observed images or sensor data) and target $y \in \mathcal{Y}$. We treat $y$ as a realisation of a random variable $Y$ whose distribution given $x$ is unknown to the model. A scalar task reward $R(\hat{y}, y) \in [0, 1]$ scores

| Symbol | Shape | Meaning |
|---|---|---|
| $z_t$ | $\mathbb{R}^{B\times C\times H\times W}$ | imagined latent at step $t$; $B$ batch, $C$ latent channels |
| $\tilde{z}_t$ | $\mathbb{R}^{B\times C\times H\times W}$ | gated latent, $\tilde{z}_t = S_\omega(z_t)$ |
| $m_t$ | $\mathbb{R}^{B\times C\times H\times W}$ | imaginal workspace (memory) |
| $\mathcal{F}z$ | $\mathbb{C}^{B\times C\times H\times W}$ | full unitary 2-D DFT; $\|z\|_2 = \|\mathcal{F}z\|_2$. Code uses the half-spectrum `rfft2`, shape $H \times (\lfloor W/2 \rfloor + 1)$ |
| $\xi$ | index | frequency index; $r(\xi) = \|\xi\|_2 / \max \|\xi\|_2 \in [0, 1]$ |
| $M_k(\xi)$ | $\mathbb{R}^{H\times(\lfloor W/2\rfloor+1)}$ | radial mask for band $k$; $M_k \geq 0$, $\sum_{k=1}^{K} M_k \equiv 1$ |
| $\alpha_{ck}$ | $\mathbb{R}^{C\times K}$ | raw gate parameters ($CK$ scalars in total) |
| $a_{ck}$ | $\mathbb{R}^{C\times K}$, $(-1, 1)$ | bounded gains, $a_{ck} = \tanh(\alpha_{ck})$; $\bar{a} = \max_{c,k} \|a_{ck}\|$ |
| $G_{\rho,c}(\xi)$ | $\mathbb{R}^{C\times H\times(\lfloor W/2\rfloor+1)}$ | effective gate, one value *per channel and frequency* |
| $\rho$ | $[0, 1]$ | residual strength; bounds distortion and expressible range |
| $\lambda_k$ | $\mathbb{R}^K$, $[0, 1)$ | per-band memory decay; $\Lambda(\xi) = \sum_k \lambda_k M_k(\xi)$ |
| $\tau_k$ | $\mathbb{R}^K$ | band time constant, $-1/\ln \lambda_k$ |
| $\pi_t^k$ | $[0, 1]^K$ | band-activation probability from the allocator |
| $\mathcal{A}_t$ | $\subseteq \{1..K\}$ | active band set at step $t$ |
| $w_k$ | $\mathbb{R}^K_{\geq 0}$ | task sensitivity spectrum, Eq. (12) |
| $T_t$ | warp | latent transport from incoming latent to memory |
| $h_t$ | $\mathbb{R}^{B\times d_L}$ | pooled language state |
| $v_t$ | $\mathbb{R}^{B\times d_V}$ | re-encoded evidence features, $v_t = V_\nu(\tilde{z}_t)$ |
| $s_t$ | — | full state $(h_t, m_t, \mathcal{P}_t)$ |
| $\mathcal{P}_t$ | typed set | provenance ledger, values in $\{\text{OBS}, \text{RET}, \text{IMG}\}$ |
| $\text{cVOI}_n$ | $\mathbb{R}$ | computational value of information w.r.t. decoder class $\mathcal{A}_n$ |
| $\Delta_t(\hat{a})$ | $\mathbb{R}$ | gain realised by the deployed reader $\hat{a}$ |
| $\beta$ | $\mathbb{R}_{\geq 0}$ | shadow price of compute (reward per normalised FLOP) |
| $B_{\text{br}}, S, \mathcal{B}$ | $\mathbb{N}, \mathbb{N}, \mathbb{R}_+$ | branch count, denoising steps, deliberation budget |
| $K$ | $\mathbb{N}$ | number of radial bands |
| $\sigma_k^2, D_k, R_k$ | $\mathbb{R}^K_{\geq 0}$ | band power, allowed distortion, allocated rate |
| $\theta$ | $\mathbb{R}_{>0}$ | water level in (14) |
| $\hat{v}_k, c_k$ | $\mathbb{R}^K$ | predicted band utility and band cost, (29) |
| $\epsilon_{\text{tol}}$ | $\mathbb{R}_{\geq 0}$ | confidence-drift tolerance (plays the role of $2\epsilon$) |
| $\pi_{\text{IMG}}$ | $[0, 1]$ | measured reliability of the imagination channel |
| ICD, ICG | $\mathbb{R}$ | imaginal confidence drift; imaginal calibration gap |

**Table 1:** Notation. Shapes are given because the channel index on the gate is easy to lose: $G_\rho$ carries *one value per channel and per frequency*, and is applied as $(\mathcal{F}S_\omega(z))_{b,c,\xi} = G_{\rho,c}(\xi)\,(\mathcal{F}z)_{b,c,\xi}$.

a prediction; for multiple-choice benchmarks $R$ is 0/1 accuracy, for open-ended tasks a verifier or exact-match program.

Deliberation proceeds in discrete steps $t = 0, 1, \ldots, T$. The reasoner maintains a state

$$s_t \coloneqq (h_t, m_t, \mathcal{P}_t), \tag{2}$$

where $h_t$ is the language model's hidden summary, $m_t$ is the *imaginal workspace* (a frequency-structured latent tensor, §3.5), and $\mathcal{P}_t$ is a *provenance ledger*: a typed record of which elements of the state derive from observation, retrieval, or self-generated simulation (§4.9). Keeping $\mathcal{P}_t$ in the state — rather than as a logging afterthought — is what makes the calibration constraint of §3.8 expressible.

At each step the controller selects an action $a_t \in \{\text{ANSWER}\} \cup \{\text{THINK}\} \cup \{\text{IMAGINE}(B)\}_{B\geq 1}$: emit an answer and stop; spend a bounded number of additional language tokens; or spend a diffusion call generating $B$ counterfactual latent hypotheses, spectrally transform them, re-encode them, and

fold the winner into $m_t$ and $h_t$. Each action has a cost $c(a_t) \geq 0$ measured in a common unit (§2.5), and the episode is subject to a budget

$$\sum_t c(a_t) \leq \mathcal{B}. \tag{3}$$

### 2.3 Why this is not an ordinary POMDP

It is tempting to describe the above as a partially observed Markov decision process in which imagination is an information-gathering action. That framing is subtly wrong, and the error matters.

In a standard POMDP, an information-gathering action queries the *environment*: it draws a sample from a channel whose output depends on the hidden state $Y$. Consequently the posterior over $Y$ genuinely sharpens, and the classical result that value of information is non-negative [18] applies.

An imagination action queries *the agent's own generative model*. The sampled latent $Z_t \sim D_\psi(\cdot \mid c_t, m_{t-1})$ depends on $Y$ only through $s_t$. No channel to the environment is opened. This is the content of Proposition 6, and it means that the Bayesian posterior over $Y$ given the full history is *unchanged* by imagining. What changes is which functionals of that posterior the agent can actually compute within its remaining budget.

We therefore model the reasoner explicitly as a *bounded* decision maker, and define the value of imagining relative to that boundedness.

### 2.4 Bounded decoders and computational value of information

**Definition 1** (Bounded decoder class)**.** A *decoder class* $\mathcal{A}_n$ is a set of measurable maps from the reasoner's accessible state to a prediction in $\mathcal{Y}$, realisable within a fixed computational allowance $n$ — concretely, the language model run for at most $n$ additional decoding steps with a fixed architecture and weights. We require $\mathcal{A}_n$ to be closed under *free disposal*: for every $a \in \mathcal{A}_n$ acting on $(h, v)$ there exists $a' \in \mathcal{A}_n$ with $a'(h, v) = a(h, \varnothing)$, i.e. the decoder is always permitted to ignore a supplied visual feature.

Free disposal is not a technicality. It is an architectural requirement: the $V \to L$ bridge must be able to contribute nothing, which in practice means residual injection with a learnable gate initialised at zero, so that the imagination-off behaviour is exactly recoverable. Without it, the non-negativity result below fails and the system can be worse than its own backbone.

**Definition 2** (Computational value of information)**.** Let $v_t = V_\nu(\tilde{z}_t)$ be the re-encoded features of an imagined, spectrally transformed latent. The *computational value of information* of the imagination step at state $s_t$, with respect to decoder class $\mathcal{A}_n$, is

$$\mathrm{cVOI}_n(s_t) \coloneqq \max_{a \in \mathcal{A}_n} \mathbb{E}\big[R(a(h_t, v_t), Y) \,\big|\, s_t\big] \;-\; \max_{a \in \mathcal{A}_n} \mathbb{E}\big[R(a(h_t, \varnothing), Y) \,\big|\, s_t\big]. \tag{4}$$

Two remarks. First, $\mathrm{cVOI}_n$ depends on $n$: it is a statement about a *particular* bounded reader, not about information. Second, the expectation is over both $Y$ and the sampling randomness of the generator, so $\mathrm{cVOI}_n$ is a property of the *policy* of imagining at $s_t$, not of one particular hallucinated picture. An individual sample can be actively misleading even when $\mathrm{cVOI}_n(s_t) > 0$.

**Remark 3** (The classical limit)**.** As $n \to \infty$ with $\mathcal{A}_n$ eventually containing the Bayes-optimal map from $h_t$, $\mathrm{cVOI}_n(s_t) \to 0$ by Proposition 6. The benefit of imagination is a *boundedness dividend*. This is the formal content of the budget-sweep prediction in §8.3.3.

### 2.5 The cost model

Costs must be commensurable or the gate is meaningless. We define cost in units of *normalised inference FLOPs*, with wall-clock and energy reported alongside but not used in the objective (they are hardware-dependent and would make the learned policy non-portable).

For a language step of $\ell$ generated tokens with a model of $N_L$ non-embedding parameters we use the standard estimate $c_{\text{lang}} \approx 2N_L\ell$. For an imagination step with $B$ branches, $S$ denoising steps at latent resolution $H \times W$, a denoiser of $N_D$ parameters evaluated with classifier-free guidance factor $g \in \{1, 2\}$, a VAE decode/encode pair of cost $c_{\text{vae}}$ (skipped in latent-only mode), and a vision encoder of $N_V$ parameters:

$$c_{\text{img}}(B, S) \;\approx\; B\Big[g\, S\, \phi_D(N_D, H, W) \;+\; \mathbb{1}[\text{decode}]\, c_{\text{vae}} \;+\; \phi_V(N_V)\Big] \;+\; c_{\text{spec}}, \tag{5}$$

where $\phi_D, \phi_V$ are the per-call FLOP counts of the denoiser and encoder and $c_{\text{spec}}$ is the (negligible) cost of the spectral operator, $O(CHW \log HW)$. Appendix C tabulates the constants for the reference configurations. Two facts follow immediately and drive much of the design:

1. $c_{\text{img}} \gg c_{\text{lang}}$ for realistic settings — a single 20-step SDXL-class call is comparable to thousands of decoded tokens. An imagination step must therefore clear a high bar, and a system that imagines by default will lose an equal-compute comparison even if imagery helps.

2. $c_{\text{img}}$ scales linearly in $B$ and $S$ but only quadratically-with-log in resolution via $\phi_D$; and the marginal value of extra branches is *sublinear* (Prop. 25). Together these imply the operating point should be many-few: low resolution, few denoising steps, modest $B$ — *reasoning fidelity, not display fidelity.*

### 2.6 The budgeted objective

Write $\pi$ for the policy over actions. The constrained problem is

$$\max_{\pi}\; \mathbb{E}_\pi\big[R(\hat{y}, Y)\big] \quad \text{subject to} \quad \mathbb{E}_\pi\Big[\textstyle\sum_t c(a_t)\Big] \le \mathcal{B}. \tag{6}$$

Its Lagrangian relaxation is

$$\mathcal{L}(\pi, \beta) = \mathbb{E}_\pi\big[R(\hat{y}, Y)\big] - \beta\Big(\mathbb{E}_\pi\big[\textstyle\sum_t c(a_t)\big] - \mathcal{B}\Big), \qquad \beta \ge 0. \tag{7}$$

Proposition 10 shows that at the optimum $\beta$ is exactly the marginal task reward per unit of compute — the shadow price of the budget — and that the greedy rule

$$\boxed{\text{IMAGINE} \iff \text{cVOI}_n(s_t) \;>\; \beta \cdot c_{\text{img}}(B, S)} \tag{8}$$

is optimal among myopic policies. This is the gate VIVID implements. It replaces the earlier draft's heuristic "$p_t = \sigma(C_\phi(h_t, u_t))$ thresholded at $\tau$" with a rule in which every quantity is separately measurable: cVOI has a regression target (§5.4), $c_{\text{img}}$ is computed from (5), and $\beta$ is obtained by dual ascent to hit a target budget rather than tuned by hand.

**Remark 4** (Myopia and its cost)**.** (8) is myopic: it ignores that imagining now may enable a more valuable imagination later. Non-myopic corrections are possible (a one-step-lookahead or a learned $Q$-function over {THINK, IMAGINE, ANSWER}), and §5.5 describes both. We default to the myopic rule because its target is directly estimable from paired rollouts, whereas the non-myopic target requires credit assignment through the whole trajectory and is the first thing to break at small data scale. Reporting both is part of the ablation set.

### 2.7 Uncertainty is an input, not the objective

The earlier draft used uncertainty $u_t$ as a proxy for when to imagine. Uncertainty is a useful *feature* — it correlates with headroom — but it is the wrong *target*, for a reason worth stating: a model can be maximally uncertain on a problem that no amount of imagined imagery would help (an unanswerable question, a missing fact), and confidently wrong on a problem where a single mental rotation would fix it. The gate must predict *gain*, not *doubt*. We therefore feed the controller a vector of uncertainty signals

$$u_t = \big( \underbrace{\mathcal{H}[p(\hat{y} \mid h_t)]}_{\text{answer entropy}}, \underbrace{\mathrm{Var}_k \hat{y}^{(k)}}_{\text{self-consistency}}, \underbrace{\mathrm{sd}\, Q_\eta}_{\text{critic disagreement}}, \underbrace{\sigma_{\text{conf}}(h_t)}_{\text{learned confidence}}, \underbrace{\mathbb{1}[\text{visual schema}]}_{\text{task-type prior}} \big), \tag{9}$$

and train $C_\phi(h_t, u_t)$ against the realised cVOI target. Whether the learned gate beats a pure-uncertainty gate is an ablation, not an assumption (§8.4).

## 3 Theory: What Imagination Can and Cannot Buy

This section contains the paper's analytical core. §3.1 establishes that self-generated evidence carries no information and works out what follows. §3.3 derives the band structure of the workspace from a rate-allocation argument. §3.4 gives the operator's exact properties, §3.5 the dynamics of the recurrent workspace, §3.7 the guarantee for counterfactual branching, and §3.8 the constraint that keeps imagined evidence from inflating confidence. Proofs are in Appendix A.

Throughout, $\mathcal{F}$ denotes the unitary (orthonormal) two-dimensional discrete Fourier transform on the $H \times W$ latent grid, so Parseval's identity $\|z\|_2 = \|\mathcal{F}z\|_2$ holds exactly. Frequencies are indexed by $\xi = (\xi_1, \xi_2)$ and $r(\xi) = \|\xi\|_2 / \max \|\xi\|_2 \in [0, 1]$ is the normalised radial coordinate.

### 3.1 No free evidence

**Assumption 5** (Closed generative loop)**.** The imagined latent is drawn as $Z_t \sim D_\psi(\cdot \mid c_t, m_{t-1})$ with $c_t = B_{L \to D}(h_t)$, where $D_\psi, B_{L \to D}$ have fixed parameters during inference and the sampling noise is independent of $Y$ given $s_t$. No external query (retrieval, tool call, sensor read) occurs during the imagination step.

**Proposition 6** (No free evidence)**.** *Under Assumption 5, $Y \to S_t \to Z_t$ is a Markov chain and therefore*

$$I(Y; Z_t \mid S_t) = 0, \qquad \textit{and consequently} \qquad p(Y \mid s_t, z_t) = p(Y \mid s_t) \ \ \textit{a.s.} \tag{10}$$

*The same holds for any measurable function of $Z_t$, including the spectrally transformed latent $\tilde{Z}_t = S_\omega(Z_t)$, the decoded image, and the re-encoded features $V_\nu(\tilde{Z}_t)$.*

The proposition is elementary. We state it as a proposition because the literature it applies to routinely describes generated images as "visual evidence," "visual thoughts that ground the answer," or "information gathered by the model," and those descriptions are strictly speaking false. It is worth being precise about what survives:

- **What is false.** That imagining sharpens the Bayesian posterior over the answer; that an imagined scene can confirm a hypothesis; that self-consistency across imagined branches is evidence of correctness in the epistemic sense.

- **What is true and still interesting.** That a bounded reader may be unable to extract from $h_t$ a functional of the posterior that it can extract easily from $V_\nu(\tilde{Z}_t)$; that the generator's learned prior over scenes constitutes *compiled* knowledge which re-perception makes locally accessible; and that this is a real, measurable, engineering-relevant effect.
- **What changes about evaluation.** A gain must be attributed to computation, so the control condition is an equal-compute reader, and the theory predicts the gain shrinks as that reader's budget grows (§8.3.3).

**Proposition 7** (Non-negativity under free disposal)**.** *If $\mathcal{A}_n$ satisfies free disposal (Definition 1) then* $\mathrm{cVOI}_n(s_t) \geq 0$ *for all $s_t$.*

**Proposition 8** (Realised value can be negative)**.** *Let $\hat{a} \in \mathcal{A}_n$ be the* actual *reader used at inference (not the argmax), and let*

$$\Delta_t(\hat{a}) \coloneqq \mathbb{E}\big[R(\hat{a}(h_t, v_t), Y) \mid s_t\big] - \mathbb{E}\big[R(\hat{a}(h_t, \varnothing), Y) \mid s_t\big].$$

*There exist tasks, generators and readers for which $\Delta_t(\hat{a}) < 0$. A two-state construction in which a reader that defers to a self-generated cue of error rate $\gamma$ achieves $\Delta_t(\hat{a}) = -\gamma$ — and, for a reader that defers only on disagreement with probability $p$, $\Delta_t(\hat{a}) = -p\gamma$ — is given in Appendix A. The gap* $\mathrm{cVOI}_n(s_t) - \Delta_t(\hat{a})$ *is the* difference *of the reader's excess risks on the two branches, and equals its excess risk from treating self-generated features as observations whenever $\hat{a}$ is optimal on the no-imagination branch.*

**Remark 9** (Two different quantities)**.** Propositions 7 and 8 are not in tension, and it is worth being explicit about why: $\mathrm{cVOI}_n$ is defined by a *maximum* over the decoder class, whereas $\Delta_t(\hat{a})$ is the gain of one *fixed* reader. The opportunity to imagine is never harmful; a particular reader's use of what it imagines can be. Everything downstream — what the controller can be trained to predict (§5.4), what the calibration constraint protects against — turns on that distinction.

The practical message is that the difference between the two is entirely a property of the reader, which in turn motivates two architectural commitments that are not optional in VIVID: zero-initialised gated injection on the $V \to L$ bridge (so free disposal actually holds), and the calibration constraint of §3.8 (so the reader's use of imagined features is bounded by their measured reliability).

## 3.2 The budgeted stopping rule

**Proposition 10** (Shadow price and the greedy gate)**.** *Consider the constrained problem (6) with Lagrangian (7). Assume $R$ is bounded and the per-step values* $\mathrm{cVOI}_n(s_t)$ *are conditionally independent of future costs given $s_t$. Then (i) strong duality holds for the randomised-policy relaxation; (ii) at the optimal multiplier $\beta^\star$, $\beta^\star = \partial\mathbb{E}[R]/\partial\mathcal{B}$, i.e. the marginal task reward per unit compute; and (iii) among policies that condition only on $s_t$ and act myopically, the optimal rule is (8): imagine iff* $\mathrm{cVOI}_n(s_t) > \beta^\star c_{img}$.

The multiplier need not be tuned. Running dual ascent

$$\beta \leftarrow \big[\beta + \eta\,(\bar{c} - \mathcal{B})\big]_+, \tag{11}$$

with $\bar{c}$ the running average episode cost, drives the realised budget to target and yields a *portable* controller: the same trained gate can be deployed at any budget by re-solving for $\beta$, without retraining. This is a concrete advantage over a fixed decision threshold, and it is directly testable — §8.5 asks for the accuracy–compute frontier traced by sweeping $\beta$, which a fixed-threshold controller cannot produce.

### 3.3 Why frequency bands: task-weighted reverse water-filling

We now derive the workspace's structure instead of positing it.

**The imaginal channel.** An imagination step must convey a scene from the generator to the reader through a pipeline with a finite effective rate: a limited number of denoising steps, a limited latent resolution, a VAE and a vision encoder each with their own bottleneck, and finally a small number of injected tokens. Model the composition as a channel that reproduces the ideal latent $z$ with per-component squared error, subject to a total rate budget $R$ nats.

**Band decomposition.** Partition the spectrum with the smooth radial masks $\{M_k\}_{k=1}^{K}$, $M_k \geq 0$, $\sum_k M_k \equiv 1$. Let $\sigma_k^2$ be the expected power of the latent in band $k$ (for natural-image latents these decay roughly as a power law in $r$), and let $D_k$ be the mean squared error incurred in band $k$.

**Task weighting.** The quantity we care about is not reconstruction error but downstream loss. Expand $\mathcal{L}_{\text{task}}$ about the clean latent under a zero-mean band-limited distortion $\delta$ with per-band power $D_k$. The first-order term $\mathbb{E}\langle \partial\mathcal{L}/\partial z,\ \delta\rangle$ *vanishes*, so the coefficient that multiplies distortion is curvature, not gradient:

$$\mathbb{E}\big[\mathcal{L}_{\text{task}}(z+\delta) - \mathcal{L}_{\text{task}}(z)\big] \ \approx\ \sum_{k=1}^{K} w_k D_k, \qquad w_k \coloneqq \tfrac{1}{2}\,\overline{\text{tr}}_k\left(\frac{\partial^2 \mathcal{L}_{\text{task}}}{\partial z^2}\right), \tag{12}$$

where $\overline{\text{tr}}_k$ is the per-coefficient average of the Hessian's trace restricted to band $k$, taken through the frozen vision encoder and reader. The weights $w_k$ are the *task sensitivity spectrum*, and $w_k D_k$ has the units of the loss.

We estimate $w_k$ without forming a Hessian, by symmetric finite differences along band-restricted random probes $v_k$ (a band-limited Hutchinson estimator):

$$\hat{w}_k = \frac{1}{2}\,\mathbb{E}_{v_k}\left[\frac{\mathcal{L}(z+\varepsilon v_k) + \mathcal{L}(z-\varepsilon v_k) - 2\mathcal{L}(z)}{\varepsilon^2\,\|v_k\|^2/n}\right], \tag{13}$$

which needs only forward passes and is therefore usable with a black-box reader. A cheaper gradient-power surrogate, $w_k^{\text{grad}} = \mathbb{E}\|M_k \odot \mathcal{F}(\partial\mathcal{L}/\partial z)\|_2^2$, is available in one backward pass; it typically *ranks* bands identically but is not the coefficient in (12), so we use it for the regulariser and report (13) for the headline measurement.

The weights differ across task families, and that is the point: mental rotation should be low-band dominated, chart reading and fine-grained counting high-band dominated.

**Proposition 11** (Task-weighted reverse water-filling)**.** *For independent Gaussian band coefficients with variances $\sigma_k^2$ and weights $w_k > 0$, the allocation minimising $\sum_k w_k D_k$ subject to $\sum_k \frac{1}{2}\log^+(\sigma_k^2/D_k) \leq R$ is*

$$D_k^\star = \min\Big(\frac{\theta}{w_k},\ \sigma_k^2\Big), \qquad R_k^\star = \tfrac{1}{2}\log^+\Big(\frac{w_k\sigma_k^2}{\theta}\Big), \tag{14}$$

*with $\theta > 0$ chosen so that $\sum_k R_k^\star = R$. Bands with $w_k\sigma_k^2 \leq \theta$ receive zero rate.*

Two structural facts fall out. Bands whose *task-weighted* power lies below the water level are not worth representing at all — which is a principled criterion for choosing the imaginal resolution, and formalises the intuition that reasoning imagery should be low-fidelity. And among coded bands, precision is allocated inversely to sensitivity-scaled distortion, not uniformly.

**Corollary 12** (A spectral gate implements the allocation)**.** *Suppose the pipeline injects noise of band-independent power* $\epsilon^2$ after *the gate (a reasonable model of quantisation, few-step denoising error, and encoder bottleneck noise, all of which act approximately whitely in the transformed domain). Applying a real gain* $G_k$ *to band* $k$ *before the noise yields effective distortion* $D_k = \epsilon^2/G_k^2$*. Matching this to* (14) *gives*

$$\boxed{G_k^\star \;\propto\; \sqrt{w_k}} \tag{15}$$

*for all coded bands, up to a single global constant.*

Two honest qualifications. The corollary *identifies* the gate that realises the optimal allocation under a fixed post-gate noise floor; it does not by itself determine the operating point, since the gain model carries no rate constraint of its own — $\theta$ enters only through the water-filling solution it is matched to. And the operator family is bounded: by Lemma 15, $G_\rho \in [1-\rho\bar{a},\, 1+\rho\bar{a}]$, so

$$\frac{\max_k G_k}{\min_k G_k} \;\le\; \frac{1+\rho}{1-\rho} \qquad\Longrightarrow\qquad \frac{\max_k w_k}{\min_k w_k} \;\le\; \left(\frac{1+\rho}{1-\rho}\right)^2. \tag{16}$$

At the reference $\rho = 0.2$ that ceiling is 2.25. A task whose measured sensitivity spectrum spans an order of magnitude therefore *cannot* be matched by this operator, and the global constant does not rescue it — the gate is pinned at unity plus a bounded perturbation, so there is no free scale to absorb. This is a real limitation, not a tuning detail: $\rho$ must be chosen large enough for the measured dynamic range of the task family, at the cost of the distortion bound in Proposition 18, and when it is not, Prediction 1 can only be tested on the ordering of gains rather than on their magnitudes. We report the measured range of $w_k$ alongside $\rho$ for exactly this reason.

This is the paper's sharpest empirical prediction, and it is cheap to test. Corollary 12 says that band gains *learned end-to-end from task loss* should, up to a positive scale, be proportional to the square root of *independently measured* band sensitivity (12). §8.3.1 specifies the test (a rank correlation across bands, tasks and seeds, with a permutation null). If the correlation is absent, the rate-allocation story is wrong even if the architecture happens to score well — and we would rather know that than report the score.

**Remark 13** (Relation to classical pre-emphasis)**.** Equation (15) is the reasoning-workspace analogue of pre-emphasis/de-emphasis in analogue communications and of perceptual weighting in transform codecs. We claim no novelty for the principle. The contribution is the identification of the imaginal pathway as such a channel, the task-weighted form of the weights, and the resulting falsifiable prediction — none of which, to our knowledge, has been stated for latent reasoning workspaces.

### 3.4 The spectral operator and its guarantees

**Definition 14** (Spectral band gate)**.** Let $a_{ck} = \tanh(\alpha_{ck}) \in (-1, 1)$ be learned per-channel, per-band gains, let $\rho \in [0, 1]$ be a residual strength, and let $\mathcal{A}_t \subseteq \{1, \dots, K\}$ be the set of bands the allocator has activated (§4.6; $\mathcal{A}_t = \{1..K\}$ when allocation is disabled). Define the gate, *per channel* $c$ *and frequency* $\xi$,

$$G_c(\xi) \;=\; 1 + \sum_{k \in \mathcal{A}_t} a_{ck}\, M_k(\xi), \tag{17}$$

and the operator, acting elementwise in the transform domain,

$$\big(\mathcal{F}S_\omega(z)\big)_{b,c,\xi} \;=\; \big[(1-\rho)+\rho\, G_c(\xi)\big]\,\big(\mathcal{F}z\big)_{b,c,\xi}, \qquad \text{equivalently} \qquad S_\omega(z) = z+\rho\big[\mathcal{F}^{-1}(G\odot\mathcal{F}z)-z\big]. \tag{18}$$

The batch index $b$ is untouched and the channel index is *not* summed over: each channel has its own $K$ coefficients and therefore its own gate. Bands outside $\mathcal{A}_t$ contribute no gain, which is what makes allocation a saving rather than a label.

**Lemma 15** (Residual form is itself a gate)**.** *$S_\omega(z) = \mathcal{F}^{-1}(G_\rho \odot \mathcal{F}z)$ with $G_{\rho,c}(\xi) = 1 + \rho \sum_k a_{ck} M_k(\xi)$. Writing $\bar{a} \coloneqq \max_{c,k} |a_{ck}| < 1$, we have $G_{\rho,c}(\xi) \in [1 - \rho\bar{a},\, 1 + \rho\bar{a}]$ for all $c, \xi$; in particular $G_\rho$ is real and strictly positive, since $\rho\bar{a} < 1$.*

This small lemma is what makes the remaining properties immediate — and it corrects a presentational weakness of the earlier draft, which treated the residual mixing as a separate device rather than recognising that it merely reparameterises the gate into a strictly positive range.

**Proposition 16** (Exact phase preservation)**.** *For every channel $c$ and frequency $\xi$, $\arg\big(\mathcal{F}S_\omega(z)\big)_c(\xi) = \arg\big(\mathcal{F}z\big)_c(\xi)$, and $\big|\mathcal{F}S_\omega(z)\big|_c(\xi) = G_{\rho,c}(\xi)\,\big|\mathcal{F}z\big|_c(\xi)$.*

Phase carries the geometry: the classical demonstration is that swapping the phase spectra of two images swaps their perceived content [33]. An operator that modulates amplitude while provably fixing phase can re-weight *how strongly* structure at each scale is expressed without moving anything. This is the property that makes the workspace safe to apply repeatedly — and it yields a clean ablation: destroying phase while preserving amplitudes should be catastrophic for spatial tasks and comparatively benign for texture-statistics tasks (§8.4).

**Proposition 17** (Equivariance)**.** *$S_\omega$ commutes exactly with cyclic translations of the latent grid. If the masks $M_k$ depend on $\xi$ only through $r(\xi)$, then $S_\omega$ commutes with rotations of the continuous plane; on a discrete $H \times W$ grid the commutation error is $O(\Delta r)$ where $\Delta r$ is the radial quantisation of the mask, and is exact for the four axis-aligned right-angle rotations when $H = W$.*

**Proposition 18** (Invertibility, bi-Lipschitz stability, bounded distortion)**.** *$S_\omega$ is a bijection on $\mathbb{R}^{C\times H\times W}$ with inverse $S_\omega^{-1}(z) = \mathcal{F}^{-1}(G_\rho^{-1} \odot \mathcal{F}z)$, and for all $z$:*

$$(1 - \rho\bar{a})\,\|z\|_2 \;\le\; \|S_\omega(z)\|_2 \;\le\; (1 + \rho\bar{a})\,\|z\|_2, \tag{19}$$

$$\|S_\omega(z) - z\|_2 \;\le\; \rho\,\bar{a}\,\|z\|_2, \tag{20}$$

$$\kappa(S_\omega) \;\le\; \frac{1 + \rho\bar{a}}{1 - \rho\bar{a}}. \tag{21}$$

*With the reference setting $\rho = 0.2$, $\bar{a} < 1$, the operator changes the latent by at most 20% in norm and has condition number below 1.5.*

Proposition 18 is the architectural counterpart of no-free-evidence: because $S_\omega$ is invertible, it *cannot* add or destroy information about the scene. Whatever it contributes is a change in accessibility. It also gives training stability for free — an invertible, well-conditioned, phase-preserving operator inserted at arbitrary denoising steps cannot make the latent trajectory diverge, which is the failure mode that typically kills naive latent interventions inside a sampler.

### 3.5 The workspace as a band-diagonal linear system

The earlier draft's memory was a scalar leaky average, $m_t = \lambda m_{t-1} + (1 - \lambda)\tilde{z}_t$. Nothing in that choice reflects the band structure the rest of the design is built on. We replace it with per-band decay.

**Definition 19** (Band-wise imaginal memory)**.** Let $\lambda_1, \dots, \lambda_K \in [0, 1)$, let $\Lambda(\xi) \coloneqq \sum_k \lambda_k M_k(\xi)$, and write $\tilde{z}_t = S_\omega(z_t)$ for the gated latent and $T_t$ for the transport alignment of §3.6 ($T_t = \mathrm{Id}$ when transport is disabled or its estimate is rejected). The workspace update is

$$\mathcal{F}m_t \;=\; \Lambda \odot \mathcal{F}m_{t-1} \;+\; (1 - \Lambda) \odot \mathcal{F}\big(T_t\, \tilde{z}_t\big). \tag{22}$$

The gate is applied *before* the memory, not inside its update. This matters: it keeps the workspace's steady-state gain at unity, so the memory neither amplifies nor attenuates a persistent hypothesis, and all spectral shaping remains localised in $S_\omega$ where the guarantees of §3.4 apply.

**Proposition 20** (Diagonal LTI dynamics)**.** (22) *is a diagonal linear time-invariant system in the Fourier basis. Its solution is*

$$\mathcal{F}m_t = \Lambda^t \odot \mathcal{F}m_0 \;+\; \sum_{j=0}^{t-1} \Lambda^j \odot (1 - \Lambda) \odot \mathcal{F}\big(T_{t-j} \tilde{z}_{t-j}\big), \tag{23}$$

*it is BIBO stable iff* $\sup_\xi \Lambda(\xi) < 1$*, its steady-state gain from the aligned input to* $m$ *is unity at every frequency, and its impulse response at frequency* $\xi$ *decays geometrically with time constant* $\tau(\xi) = -1/\ln \Lambda(\xi)$ *and mean lag* $\Lambda(\xi)/(1 - \Lambda(\xi))$*.*

**Remark 21** ($K$ constants, not $K$ filters)**.** Because the masks are smooth and overlap, $\Lambda(\xi)$ takes a continuum of values and the system is strictly a bank of one scalar filter per frequency, whose decay profile is *parameterised* by $K$ constants. It reduces to exactly $K$ independent filters only for a hard indicator partition. We keep the smooth masks — they are what make the operator's equivariance well behaved (Prop. 17) — and report $\tau_k$ as the mask-weighted time constant of band $k$, which is the quantity the code computes and the one that is comparable across runs.

Two consequences.

*Architecturally*, this places the workspace in the same family as diagonal structured state-space models [12, 11], but in a *fixed, interpretable* basis rather than a learned one. The per-band time constants $\tau_k$ are directly readable, which turns an opaque memory into a measurable object: with the reference decays $\lambda = 0.95 \to 0.30$ one can report "layout persists for $\sim 19$ steps, texture for $\sim 0.8$" and check whether that matches what the task needs.

*Cognitively*, it lets the model do something the scalar version cannot: keep the gist and refresh the detail. Setting $\lambda_{\text{low}} \to 0.95$ and $\lambda_{\text{high}} \to 0.3$ means a scene's coarse spatial arrangement survives many transformation steps while its surface appearance is re-sampled each time. This matches the coarse-to-fine, low-detail character reported for voluntary visual imagery [35], and it makes a prediction: *forcing* $\lambda$ *to be uniform across bands should selectively damage multi-step transformation tasks while leaving single-shot recognition tasks unaffected* (§8.4).

**Remark 22** (Why not learn the basis?)**.** A learned mixing matrix would subsume the radial masks. We keep the basis fixed for three reasons: it makes $w_k$, $G_k$ and $\tau_k$ comparable quantities that can be plotted against each other (which is the whole point of §8.3.1); it keeps the operator's guarantees exact; and it removes the most obvious route by which the workspace could silently become a generic extra parameter block, which is precisely the confound the equal-compute baseline is meant to exclude. A learned-basis variant is an ablation, and if it wins without preserving the gain–sensitivity correspondence, that is evidence *against* our account, not for it.

### 3.6 Latent transport: why fusion needs registration first

The workspace update (22) fuses an incoming latent into memory. That is sound only if the two are *registered*. They typically are not: two samples of the same scene from a generator conditioned on the same state differ by a small viewpoint offset, and averaging unregistered views does not accumulate geometry — it cancels it.

**Proposition 23** (Unregistered fusion attenuates structure)**.** *Let $\tilde{z}^{(j)} = T_{\delta_j} z$ be cyclic translations of a common scene $z$ by i.i.d. offsets $\delta_j$ on the $H \times W$ grid, with $\varphi(\xi) := \mathbb{E}[e^{-2\pi i\langle \delta, \xi\rangle}]$. Fix weights $c_j \geq 0$ and let $\Sigma$ be the offset covariance. Then:*

*(i)* Mean spectrum. $\mathbb{E}\Big[\mathcal{F}\big(\sum_j c_j \tilde{z}^{(j)}\big)\Big](\xi) = \big(\sum_j c_j\big)\, \varphi(\xi)\, \mathcal{F}z(\xi)$.

*(ii)* Attenuation. $|\varphi(\xi)| \leq 1$, *with* $1 - |\varphi(\xi)| \approx 2\pi^2\, \xi^\top \Sigma\, \xi$ *near the origin, so attenuation grows with* $\|\xi\|$ for small $\|\xi\|$ *at a rate set by the jitter. Strict attenuation at every* $\xi \neq 0$ *requires the offset law to be aperiodic on the grid (its support must generate the full group); a law supported on a proper coset — $\delta_1 \in \{0, H/2\}$, say — leaves $|\varphi| = 1$ at every even $\xi_1$.*

*(iii)* Retained energy. *With* $\sum_j c_j = 1$, $\mathbb{E}\big|\sum_j c_j e^{-2\pi i\langle \delta_j, \xi\rangle}\big|^2 = |\varphi(\xi)|^2 + \big(1 - |\varphi(\xi)|^2\big)\sum_j c_j^2$, *which lies strictly between $|\varphi|^2$ and 1 and decreases toward $|\varphi|^2$ as the weights spread out.*

Three things are worth stating precisely, because the loose version of this proposition is easy to state and wrong in three separate ways.

*Monotone decay in $\|\xi\|$ is not general.* It holds for Gaussian and symmetric stable offsets. For the bounded uniform jitter our pilots use, $|\varphi|$ is a Dirichlet kernel with genuine zeros and side-lobes, and is not monotone. What is general is the near-origin expansion in (ii): coarse bands are attenuated least, which is the direction the design relies on, without any claim about the far tail.

*Attenuation does not compound over the recursion as stated.* Unrolling (22) with offsets drawn independently about a *common* anchor gives $\mathbb{E}[\mathcal{F}m_t] = \Lambda^t \odot \mathcal{F}m_0 + (1 - \Lambda^t) \odot \varphi \odot \mathcal{F}z$ — a single factor of $\varphi$, however many steps are fused. Compounding requires the offsets to *accumulate*: each latent registered against a memory that has itself drifted, so that the effective offset performs a random walk and $\varphi$ enters once per step. That is the realistic regime for a workspace carried across many reasoning steps, and it is a stronger assumption than (i)–(iii) require, so we flag it as an assumption rather than deriving it.

*Part (iii) is the part the experiments measure.* Retained high-frequency *energy*, not the mean spectrum, is what a downstream encoder reads, and the two differ: an unbiased-but-scattered ensemble has mean spectrum $\varphi\mathcal{F}z$ while retaining energy strictly above $|\varphi|^2$. Stating only (i) would predict a quantity nobody measures.

Even in the conservative single-factor reading, the damage is *band-selective*: the attenuation at band $k$ scales with $\bar{r}_k^2$ through (ii), so high bands are hit first. A workspace that fuses without registering therefore silently specialises toward coarse tasks — which is separately testable, and is a sharper prediction than "averaging blurs".

The remedy is to estimate an alignment $T_t$ from the incoming latent to the memory and apply it before fusing. Two design constraints keep this from becoming a new source of drift.

*The warp family is deliberately tiny.* We use translation (estimated in closed form by normalised phase correlation, which costs two FFTs and is exact for a pure cyclic shift), optionally extended to a similarity transform by a coarse search over rotation and scale. A low-parameter warp cannot invent structure; it can only move what is already present. This matters because the operator is fitted against a noisy objective, and an expressive warp family fitted against a noisy objective

will eventually find a pathological warp that reduces residual without corresponding to any real transformation.

*Alignment must earn its acceptance.* The estimated warp is applied only if it reduces residual energy against the memory by a margin $\eta$:

$$T_t \;=\; \begin{cases} \hat{T}, & \text{if } \big(\|m_{t-1} - z_t\|^2 - \|m_{t-1} - \hat{T} z_t\|^2\big) / \|m_{t-1} - z_t\|^2 \;>\; \eta, \\ \mathrm{Id}, & \text{otherwise,} \end{cases} \tag{24}$$

and rejected outright if the implied displacement exceeds a plausible fraction of the grid — a large estimated shift usually means these are not two views of one scene, in which case fusing them at all is the wrong move. The threshold $\eta$ is not a free parameter to be asserted: it trades acceptance of genuine translations against acceptance of unrelated scenes, and §7.6 measures that trade-off and picks $\eta$ from it.

**Remark 24** (Phase correlation is the natural estimator here)**.** Registration by normalised phase correlation uses only the *phase* of the cross-power spectrum. That pairs precisely with Proposition 16: the gate modulates amplitude and provably fixes phase, so applying the gate cannot corrupt the signal the aligner reads. The two components are orthogonal by construction rather than by luck, and either can be ablated without disturbing the other.

### 3.7 Counterfactual branching: a guarantee and a warning

At an imagination step the model samples $B$ hypotheses and keeps the best by critic score. Let $\mathcal{U}$ be a large candidate pool and $q_b$ the (random) critic score of branch $b$.

**Proposition 25** (Submodularity and budgeted greedy)**.** *Let $q_b \geq 0$ (the calibrated critic emits reliabilities, so this holds by construction) and set $f(\mathcal{S}) \coloneqq \mathbb{E}\big[\max_{b\in\mathcal{S}} q_b\big]$ with $f(\emptyset) = 0$. Then $f$ is non-negative, monotone and submodular for an arbitrary joint distribution of $(q_b)_{b\in\mathcal{U}}$. Consequently greedy selection under a cardinality constraint attains $(1-1/e)\, f(\mathcal{S}^\star)$, and under a knapsack constraint on $\sum_b c_b$ the greedy-with-enumeration algorithm of [44] attains the same factor. If the critic is an $\epsilon$-approximate value oracle, the guarantee degrades gracefully to $(1-1/e)f(\mathcal{S}^\star) - |\mathcal{S}|\epsilon$.*

Non-negativity and $f(\emptyset) = 0$ are not decoration: the Nemhauser–Wolsey–Fisher bound is $f(\mathcal{S}_{\text{greedy}}) - f(\emptyset) \geq (1-1/e)[f(\mathcal{S}^\star) - f(\emptyset)]$, which collapses to the familiar form only for a normalised, non-negative $f$. This is why §3.8's requirement that the critic emit a calibrated reliability rather than an unbounded preference has a second payoff beyond honesty about confidence.

**Corollary 26** (Selected-branch value must saturate)**.** *Submodularity implies diminishing returns in $B$: along the greedy sequence, the marginal value $f(\mathcal{S}_{B+1}) - f(\mathcal{S}_B)$ is non-increasing, so $B \mapsto f(\mathcal{S}_B)$ is concave. If downstream accuracy is a non-decreasing* concave *transform of $f$, accuracy is concave in $B$ as well.*

The caveat in the second sentence is real and we flag it rather than bury it: monotonicity of the reader's success probability in branch utility does *not* imply concavity, and a bounded accuracy curve can plausibly be sigmoidal — convex at small $B$ before turning over. What Corollary 26 licenses without extra assumptions is therefore the weaker statement that the *critic value* of the selected set must be concave in $B$, which is directly measurable and which we report alongside accuracy. An accuracy curve that keeps rising linearly while the critic value has plainly saturated is the signature worth auditing — most commonly a critic correlated with the answer key, or test-set leakage. We keep this as a stop-and-audit trigger (§9) rather than as a falsifier, because of exactly that gap.

### 3.8 Calibration: what stops imagined evidence from inflating confidence

Proposition 6 has a diagnostic corollary that we have not seen used and that costs nothing to compute.

**Corollary 27** (Confidence must be paid for in accuracy)**.** *For a Bayes-consistent reader, the posterior over $Y$ is invariant across imagination steps (Prop. 6); hence its confidence in the eventual answer is a martingale with respect to the imagination filtration. For a bounded reader the computed posterior does change, but any systematic increase in confidence must be matched by a corresponding increase in accuracy. Define, over imagination steps,*

$$\mathrm{ICD} \coloneqq \mathbb{E}\big[\hat{p}_{t+1} - \hat{p}_t\big] \qquad \textit{(imaginal confidence drift)}, \qquad \Delta\mathrm{Acc} \coloneqq \mathbb{E}\big[\mathbb{1}\{\hat{y}_{t+1} = y\} - \mathbb{1}\{\hat{y}_t = y\}\big]. \tag{25}$$

*If the reader is calibrated at level $\epsilon$ at both steps then $\mathrm{ICD} \le \Delta\mathrm{Acc} + 2\epsilon$ (one $\epsilon$ per step). Contrapositively, $\mathrm{ICD} > \Delta\mathrm{Acc} + 2\epsilon$ certifies miscalibration: a regime with $\mathrm{ICD} \gg \Delta\mathrm{Acc} \approx 0$ is hallucination amplification, detectable without any external judge.*

We write the admissibility test with a single tolerance $\epsilon_{\mathrm{tol}}$ that plays the role of $2\epsilon$; the implementation exposes it as one number, and the factor of two is absorbed there rather than carried through every use.

**Definition 28** (Provenance-calibrated reader)**.** Stratify predictions by provenance class $P \in \{\textsc{obs}, \textsc{ret}, \textsc{img}\}$ — whether the decisive evidence was observed, retrieved, or imagined — as recorded in the ledger $\mathcal{P}_t$. A reader is *provenance-calibrated at level $\epsilon$* if for every class $P$ and every confidence bin $\mathcal{I}$, $\big|\mathbb{P}\big(Y = \hat{y} \,\big|\, \hat{p} \in \mathcal{I},\, P\big) - \mathbb{E}[\hat{p} \mid \hat{p} \in \mathcal{I}, P]\big| \le \epsilon$.

Vivid enforces this in three ways, all cheap and all testable.

1. **Stratified recalibration.** Temperature (or vector) scaling parameters are fit *separately* per provenance class on a held-out split, rather than a single global temperature. Post-hoc scaling is the standard remedy for systematic over-confidence [13]; we stratify it because a new evidence channel plausibly carries its own bias, and we report the fitted temperatures per class so that whether it did can be seen rather than assumed.

2. **A reliability cap.** Let $\pi_{\textsc{img}}$ be the measured held-out accuracy of decisions determined by imagined evidence. The reader's reported confidence on such decisions is capped at $\pi_{\textsc{img}}$. This is a hard ceiling, not a penalty: no amount of internal agreement across imagined branches can push confidence above the empirical reliability of the imagination channel.

3. **An admissibility gate.** Define the *Imaginal Calibration Gap* $\mathrm{ICG} \coloneqq \mathrm{ECE}_{\text{imagine-on}} - \mathrm{ECE}_{\text{imagine-off}}$. A checkpoint is inadmissible for promotion if $\mathrm{ICG} > \epsilon_{\max}$ or if $\mathrm{ICD} > \Delta\mathrm{Acc} + \epsilon_{\mathrm{tol}}$, *regardless of accuracy gains* (§6.3).

**Remark 29** (Why a cap rather than a loss term)**.** A calibration penalty added to the training loss is traded off against accuracy and will be sacrificed when the accuracy gradient is large. A cap and a promotion gate are not tradeable. Given Proposition 8 — realised value from imagined evidence can be negative precisely when the reader over-trusts it — we think the non-tradeable version is the correct design, and we report both so the cost of the constraint is visible.

| Result | Statement | Design consequence |
|---|---|---|
| No free evidence (P6) | $I(Y; Z_t \mid S_t) = 0$ | Gains are computational; equal-compute baseline is mandatory |
| Free disposal (P7) | $\text{cVOI}_n \geq 0$ | Zero-initialised gated $V \to L$ injection |
| Negative realised value (P8) | $\Delta_t(\hat{a})$ can be $< 0$ | Calibration cap on imagined evidence |
| Shadow price (P10) | $\beta^\star = \partial \mathbb{E}[R]/\partial \mathcal{B}$ | Dual ascent replaces threshold tuning; budget-portable gate |
| Water-filling (P11, C12) | $G_k^\star \propto \sqrt{w_k}$ | Radial gate parameterisation; low imaginal resolution; a falsifiable prediction |
| Operator guarantees (P16–18) | phase-exact, invertible, $\kappa < 1.5$ | Safe to insert inside the sampler; phase-destruction ablation |
| Band-diagonal LTI (P20) | bank of $K$ IIR filters, readable $\tau_k$ | Per-band decay; gist-outlives-detail prediction |
| Submodularity (P25, C26) | $(1 - 1/e)$ greedy; returns saturate | Small $B$; non-saturating curves are a red flag |
| Confidence drift (C27) | $\text{ICD} \leq \Delta\text{Acc} + \epsilon_{\text{tol}}$ | Judge-free hallucination-amplification detector |

**Table 2:** The theory, and what each result forces in the architecture or the evaluation.

### 3.9 Summary of the theoretical picture

Table 2 collects the results. The through-line is that once internal imagery is treated as metered computation rather than as evidence, most of the architecture stops being a design choice: the band structure, the gate's parameterisation, the compute penalty, the branch count, and the calibration constraint are all consequences, and each comes with something that could falsify it.

## 4 Architecture

### 4.1 System overview

VIVID couples four pretrained components through five small trainable ones. The pretrained components are a language model $L_\theta$, a latent diffusion generator $D_\psi$ with its autoencoder $(\mathcal{E}, \mathcal{D})$, and a visual encoder $V_\nu$. The trainable components are the semantic→diffusion bridge $B_{L\to D}$, the spectral operator $S_\omega$ with its band-wise memory, the vision→language bridge $B_{V\to L}$, the budgeted controller $C_\phi$ that regresses cVOI, and the calibrated evidence critic $Q_\eta$. Figure 1 shows the loop.

The control flow is deliberately *asymmetric*: the default path is the frozen backbone answering directly, and the imaginal path is an exception that must justify itself. This is the opposite of most interleaved multimodal reasoners, which imagine on a fixed schedule, and it is what makes the equal-compute comparison winnable in principle.

### 4.2 Pretrained bootstrapping and the parameter budget

Stage-1 training freezes all backbones:

$$\theta, \psi, \nu \leftarrow \text{pretrained}, \qquad \text{trainable} = \{\phi,\ \omega,\ \lambda_{1:K},\ B_{L\to D},\ B_{V\to L},\ \eta\}. \tag{26}$$

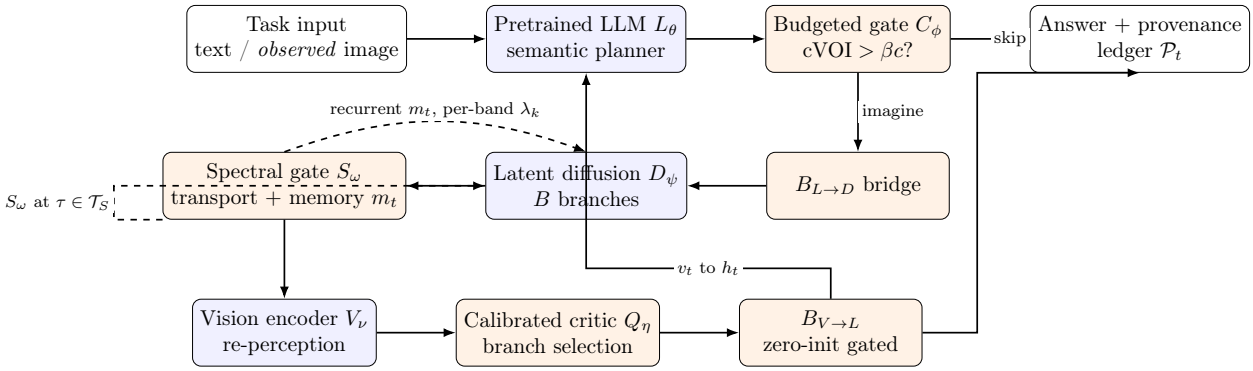


**Figure 1:** VIVID. Blue components are frozen pretrained backbones; orange are trained. The workspace is entered only when the controller's predicted computational value of information exceeds the shadow price of compute. The dashed paths carry the band-structured recurrent state back into the sampler, so the workspace shapes generation rather than only post-processing it.

The trainable set is small by construction. For $C = 4$ latent channels, $K = 6$ bands, $d_L = 4096$, $d_{\text{cond}} = 2048$, $n_{\text{tok}} = 8$ and a $d_V = 1152$ vision encoder, the parameter counts are approximately: spectral gains $CK = 24$; band decays $K = 6$; $B_{L\to D} \approx d_L \cdot n_{\text{tok}} d_{\text{cond}} \approx 67$M; $B_{V\to L} \approx d_V \cdot n_{\text{tok}} d_L \approx$ 38M; controller and critic $< 5$M each. The spectral workspace itself — the component the paper is about — has **thirty parameters**.

We draw attention to that number because it is methodologically important. A mechanism with 30 parameters cannot plausibly be absorbing capacity, which removes the most common confound in "architecture X helps" claims. If the spectral workspace helps, it is not because it added a model. Conversely, if a 30-parameter operator produces a large gain, that gain deserves unusual scrutiny, and §8.4 supplies it (randomised gains, frozen-at-init gains, phase destruction, and a learned-basis control).

### 4.3 Three integration depths

The operator can be attached to the generator at three depths, in increasing order of coupling and of training risk. All three are implemented; the middle one is the default.

1. **Post-latent (D1).** Generate a latent normally, apply $S_\omega$ once, then encode. The weakest coupling: the gate cannot influence what is generated, only how it is read. Cheap, trivially stable, and useful as a lower bound on what the mechanism can do.

2. **Scheduled spectral injection (D2, default).** Apply $S_\omega(\cdot\,; m_{t-1})$ at a chosen subset $\mathcal{T}_S$ of denoising steps via the sampler's step-end callback:

$$z_{\tau-1} = D_\psi(z_\tau, c, \tau), \qquad z_{\tau-1} \leftarrow S_\omega\big(z_{\tau-1};\, m_{t-1}\big) \quad \text{for } \tau \in \mathcal{T}_S. \tag{27}$$

   Because $S_\omega$ is invertible and well conditioned (Prop. 18), this cannot destabilise the trajectory — the practical objection to mid-sampler edits. *Where* in the schedule matters: early steps determine layout, late steps determine texture, so we default to $\mathcal{T}_S$ concentrated in the first third of the schedule and treat the placement as a swept hyperparameter (§C).

3. **Spectral-conditioned denoiser (D3).** Feed per-band summaries of $m_{t-1}$ into the denoiser through FiLM or cross-attention adapters trained with LoRA [19]. Strongest coupling, highest training risk, and the only depth that requires touching $\psi$ at all. Reserved for after D2 has shown an effect.

Reporting all three is part of the mechanism argument: a monotone D1 < D2 < D3 ordering is what the account predicts, and a flat ordering would suggest the operator is acting as a generic regulariser rather than as a workspace.

### 4.4 The bridges

$B_{L\to D}$ maps the pooled language state to $n_{\text{tok}}$ conditioning tokens in the generator's text-embedding space, trained with a combination of a contrastive objective against the generator's own text encoder on paired captions and the downstream task loss. Conditioning on *hidden states* rather than on a decoded prompt string is a deliberate choice: it avoids a lossy text bottleneck, keeps the pathway differentiable, and — relevant to Prop. 6 — makes explicit that the generator sees nothing the reasoner did not already have.

$B_{V\to L}$ maps re-encoded visual features back into $n_{\text{tok}}$ language-space tokens. It is injected residually with a learnable scalar gate initialised at zero:

$$h \leftarrow h + g_V \cdot B_{V\to L}(v), \qquad g_V \leftarrow 0 \text{ at init.} \tag{28}$$

This is what makes free disposal (Definition 1) hold exactly at initialisation and approximately thereafter, and is therefore a prerequisite for Proposition 7. It also means the system provably starts as its own backbone, so any measured change is attributable.

### 4.5 The budgeted controller

$C_\phi$ takes $(h_t, u_t)$ and outputs a scalar estimate $\widehat{\text{cVOI}}(s_t)$ — *not* a probability. The decision is (8). Because $\widehat{\text{cVOI}}$ is on the scale of task reward, three things become possible that a thresholded logit does not allow: the gate transfers across budgets by re-solving for $\beta$; the controller's predictions can be *calibrated* against realised gain and that calibration reported (§8.5); and the compute penalty has units, so (5) enters the decision directly rather than through a tuned constant.

The controller also chooses the shape of the step, not just whether to take it. Given the water-filling solution (14), bands below the water level need no rate, which determines an adequate latent resolution; the marginal value of branches saturates (Cor. 26), which bounds $B$. We therefore expose $(B, S, H\times W)$ as controller outputs selected from a small discrete menu, with the same cVOI $> \beta c$ test applied to each configuration — a knapsack choice rather than a binary one. It is worth noting a trap here: the menu is only meaningful if the deliberation budget exceeds the cost of its widest entry, or the expensive configurations are pruned before they are ever scored. Appendix C sets $\mathcal{B}$ accordingly, and the implementation reports the scored value of every menu entry — including unaffordable ones — so a degenerate menu is visible in the trace rather than silent.

### 4.6 Band allocation, and how it is kept from collapsing

The controller decides not only *whether* to imagine but *which bands to spend on*. A band allocator $A_\chi$ emits per-band probabilities $\pi_t^k = \sigma(A_\chi(h_t, u_t))_k$ and an active set

$$\mathcal{A}_t = \{\, k : \pi_t^k \cdot \hat{v}_k > \beta\, c_k \,\}, \tag{29}$$

where $c_k$ is the band's cost (higher radial bands carry more coefficients and are genuinely more expensive to represent and denoise for) and $\hat{v}_k$ is a running estimate of the band's marginal utility. This is the step-level knapsack of §4.5 applied one level down, with the *same* shadow price, so bands and configurations are priced in the same units.

**The initialisation problem, and the fix.** A logit head initialised at zero gives $\pi^k = 0.5$ for every band. Under any threshold rule that activates roughly half the spectrum at random, and the moment task loss starts flowing, the policy collapses to all-bands-active — because early in training the cheapest way to reduce loss is to stop discarding information, and the compute term has not yet accumulated enough weight to oppose it. Three measures together prevent this:

1. **Start sparse.** The output bias is initialised negative ($-1.5$ by default, giving $\pi \approx 0.18$), so the policy begins in the abstaining regime and each band must earn activation against its price rather than being switched off after the fact.
2. **Buy exploration early and cheaply.** An entropy bonus on $\pi$, decayed linearly to zero over training, keeps the allocator from committing before it has evidence — and decays so that the final policy is not paying an exploration tax.
3. **Make collapse visible.** Allocation entropy and the mean active-band count are logged every step and written to the trace. Collapse is then something one *observes* rather than infers from a disappointing ablation, which is the difference between a diagnosable failure and a mysterious one. §7.4 reports these curves.

The allocator's target is measured, not assumed: $\hat{v}_k$ is estimated by band ablation — drop band $k$ and see how much accuracy falls — on a subsample, at a fixed interval, since the measurement costs one extra forward pass per band.

### 4.7 Latent transport

Before an imagined latent is folded into memory it is aligned to the memory (§3.6). The implementation is a closed-form translation estimate by normalised phase correlation, optionally extended to a similarity transform; the estimate is applied only if it reduces residual energy by a margin (24) and implies a plausible displacement.

Three properties are worth stating explicitly because they bear on the feasibility concerns a reader will reasonably have. The operator has *no learned parameters* in its default mode, so it cannot be the source of a capacity confound. Its cost is two FFTs, an argmax and a roll — the same order as the gate itself, and negligible against a denoiser call (§4.11). And its decisions are recorded: the estimated shift, the residual before and after, and whether it was accepted all go into the trace, so a run in which alignment is silently rejecting every step looks different from one in which it is working.

### 4.8 The calibrated evidence critic

$Q_\eta(h_t, v_t^{(b)})$ scores branches for *task utility*, not image quality. Two design points distinguish it from a CLIP-similarity scorer (which is the standard baseline and is included as an ablation):

*It is trained against realised outcomes.* The target is whether injecting branch $b$ changed the answer from wrong to right, estimated from paired rollouts (§5.4). A critic trained on prompt-image agreement scores prettiness and prompt adherence; those are not the same quantity and, for counterfactual reasoning, can be anti-correlated with usefulness — the most useful imagined scene for testing a hypothesis is often the one that violates the prompt.

*It emits a calibrated reliability, not a preference.* $Q_\eta$'s output is passed through a provenance-stratified temperature (§3.8) so that the scalar it hands the reader is interpretable as "probability that acting on this branch is correct," capped at the measured channel reliability $\pi_{\text{IMG}}$.

**Algorithm 1** VIVID inference: budgeted spectral imagination

**Require:** task $x$; frozen $L_\theta, D_\psi, V_\nu$; trained $C_\phi, S_\omega, Q_\eta, B_{L\to D}, B_{V\to L}$; shadow price $\beta$; budget $\mathcal{B}$; configuration menu $\mathcal{C}$

$h \leftarrow L_\theta(x)$; $m \leftarrow \varnothing$; $\mathcal{P} \leftarrow \{\text{OBS}\}$; $c_{\text{used}} \leftarrow 0$
**for** $t = 1, \dots, T_{\max}$ **do**
  $u_t \leftarrow$ uncertainty features$(h)$ ▷ entropy, self-consistency, critic spread
  $(\hat{V}, \text{cfg}) \leftarrow \max_{\text{cfg}\in\mathcal{C}} \left[C_\phi(h, u_t; \text{cfg}) - \beta\, c_{\text{img}}(\text{cfg})\right]$
  **if** $\hat{V} \le 0$ **or** $c_{\text{used}} + c_{\text{img}}(\text{cfg}) > \mathcal{B}$ **then**
    **break** ▷ answer from language state alone
  **end if**
  $c_t \leftarrow B_{L\to D}(h)$
  **for** $b = 1, \dots, B_{\text{cfg}}$ **do** ▷ counterfactual branches
    $z^{(b)} \sim D_\psi(\cdot \mid c_t, m)$ ▷ $S_\omega$ applied at $\tau \in \mathcal{T}_S$ if depth D2/D3
    $\tilde{z}^{(b)} \leftarrow S_\omega(z^{(b)}; m)$; $v^{(b)} \leftarrow V_\nu(\tilde{z}^{(b)})$
    $q^{(b)} \leftarrow Q_\eta(h, v^{(b)})$ ▷ calibrated, capped at $\pi_{\text{IMG}}$
  **end for**
  $b^\star \leftarrow \arg\max_b q^{(b)}$ ▷ exact over the $B$ sampled branches
  $\mathcal{F}m \leftarrow \Lambda \odot \mathcal{F}m + (1-\Lambda) \odot \mathcal{F}\tilde{z}^{(b^\star)}$ ▷ band-wise decay
  $h \leftarrow h + g_V \cdot B_{V\to L}(v^{(b^\star)})$; tag $\mathcal{P} \leftarrow \mathcal{P} \cup \{\text{IMG}\}$
  $c_{\text{used}} \leftarrow c_{\text{used}} + c_{\text{img}}(\text{cfg})$; append trace step
**end for**
$\hat{y}, \hat{p} \leftarrow L_\theta(h)$; recalibrate $\hat{p}$ by provenance class; cap at $\pi_{\text{IMG}}$ if IMG $\in \mathcal{P}$
**return** $\hat{y}$, $\hat{p}$, audit trace

### 4.9 Provenance typing and the audit trace

Every element that can influence the answer carries a provenance tag in $\{\text{OBS}, \text{RET}, \text{IMG}\}$. The tags are not decoration: they select which calibration parameters apply (Definition 28), they determine the confidence cap, and they are what a verifier consults to refuse to treat imagined pixels as observations.

The trace records, per step: whether imagination was invoked and in which configuration; a short natural-language hypothesis description; predicted cVOI and realised gain when available; critic score and its calibrated reliability; the uncertainty vector; the per-band gains and memory state norms; the incurred cost; and the verification status. This is compact, structured, and scientifically useful, and it deliberately does *not* require storing or exposing unrestricted hidden chain-of-thought — the trace is a record of *decisions and costs*, not of the model's private deliberation.

### 4.10 Inference algorithm

Algorithm 1 differs from the earlier draft's loop in four substantive ways: the gate compares a reward-scaled value against a priced cost rather than thresholding a logit; the configuration (branches, steps, resolution) is chosen, not fixed; the memory update is band-wise; and the final confidence is recalibrated and capped by provenance. Each of those is a consequence of §3.

### 4.11 Complexity, memory and measured cost

A matched-compute protocol is only credible if the components it matches on are actually accounted for. The workspace's asymptotic costs, for a latent of $C$ channels and $H \times W$ resolution with $K$

bands:

| Component | Complexity | Note |
|---|---|---|
| Spectral gate $S_\omega$ | $O(CHW \log HW)$ | two FFTs, one elementwise multiply |
| Radial masks | $O(KHW)$, cached | computed once per $(H, W, K)$ |
| Band-wise memory | $O(CHW \log HW)$ | two FFTs, a blend, one inverse |
| Transport (translation) | $O(CHW \log HW)$ | two FFTs, argmax, roll; no parameters |
| Transport (similarity) | $O(RS\,CHW \log HW)$ | $R$ rotations $\times$ $S$ scales searched |
| Vision encoder | $O(\phi_V)$ | fixed per call |
| Denoiser step | $O(\phi_D)$, attention quadratic in tokens | dominates everything above |

**Table 3:** Every workspace component is quasi-linear in the number of latent coefficients. The denoiser is not.

Measured on a single CPU core, batch 8, $C = 4$, $K = 6$ (Appendix F): at a $32 \times 32$ latent the gate takes 0.63 ms, the memory update 0.75 ms and transport 0.38 ms, for a workspace total of 1.76 ms and an analytic working set of 0.42 MB. Growing the latent to $128 \times 128$ — 64× the coefficients — raises the total to 30.4 ms, a 21.7× increase. That is *sub*linear rather than the $\approx 112\times$ quasi-linear scaling predicts, because the smallest latents are dominated by fixed overhead; the measurement therefore bounds the cost from above and does not confirm the exponent. Against the cost model of (5), the spectral term's share of a single imagination step is $1.7 \times 10^{-6}$ at the default configuration and $4.9 \times 10^{-5}$ at the cheapest.

The practical consequence for §8.2 is concrete: **the workspace components do not need to be equalised in the matched-compute protocol, because they are four to six orders of magnitude below the diffusion call they accompany.** What must be equalised is the number and shape of generator calls, which is what the protocol matches on. We state this as a measured claim rather than an assumption because the reverse — a workspace whose bookkeeping rivalled its generator — would invalidate every comparison in the paper.

Memory, as opposed to compute, is dominated by the backbones rather than the workspace: the imaginal state is a single latent tensor per branch, and the analytic working set above is under 7 MB even at $128^2$. The binding constraint is holding the denoiser and the language model resident, which is why §10 treats accelerator memory as a deployment risk of the *generator*, not of the mechanism this paper proposes.

### 4.12 What to build first

The system has many interacting parts — controller, allocator, bridges, gate, memory, transport, critic, calibrator, verifier — and a reader is right to worry that a result from the whole assembly would be hard to attribute. Two things address that.

First, the components are separately ablatable *by construction* rather than by configuration flag: setting $a = 0$ makes the gate the exact identity, setting $\lambda$ uniform reduces the workspace to a scalar leaky average, and disabling transport leaves the alignment path an identity. The correctness of each reduction is asserted in the test suite, so an ablation cannot silently fail to ablate.

Second, we recommend building in this order, and stopping at the first stage that does not pay:

1. **Gate only.** Frozen backbones, a fixed single imagination configuration, no spectral operator, no memory. Establishes whether imagery helps at all against B1, and whether a priced gate can abstain. If this fails, nothing downstream matters.

2. + **workspace memory.** Adds recurrence across steps. Establishes whether multi-step accumulation helps, before any spectral structure is introduced.
3. + **transport.** Cheapest remaining component, no parameters, and the one with the clearest predicted failure mode (Prop. 23).
4. + **spectral gate.** $CK$ parameters. This is the paper's mechanism claim.
5. + **band allocation.** Only worth adding once the gate demonstrably matters, since allocation is a refinement of a mechanism that must first exist.
6. + **deeper integration (D3), branching, recursive improvement.**

Stages 1–3 involve no learned spectral parameters at all, so a positive result there would be evidence for the metering account and *not* for the spectral one — which is the sort of partial outcome we want to be able to report cleanly.

### 4.13 Memory, precision and deployment

A frozen-backbone bootstrap fits comfortably on a single high-memory accelerator: 4-bit or 8-bit quantisation of $L_\theta$, bf16 for $D_\psi$ with attention slicing and optional CPU offload of the autoencoder, and bf16 for $V_\nu$. Only the bridges, controller, critic and the 30-parameter spectral block hold optimiser state.

Three deployment-oriented consequences of the theory are worth making explicit. *Latent-only mode*: because the vision encoder can be adapted to consume latents directly, the VAE decode/encode round trip can be skipped entirely, removing $c_{\text{vae}}$ from (5); the model then reasons over imagined latents it never renders. *Low resolution*: water-filling says bands above the encoder's effective cutoff get zero rate, so $256^2$ pixels (or $32^2$ latents) is the starting point, not a compromise. *Distillation*: since the workspace is a bank of first-order filters over a generator's output, the entire imaginal step is a candidate for distillation into a cheap direct latent predictor — which, if it preserves the gain, would be the strongest possible confirmation that the benefit was computational rather than informational (§8.3.3).

## 5 Training

Training is staged so that failures localise. Each stage has an entry criterion, an objective, and an exit criterion; a stage that fails its exit criterion is diagnosed rather than trained through.

### 5.1 Stage 0: backbone qualification

Before any coupling is trained, each frozen backbone is measured alone. This stage exists because the most likely explanation for a null result in a system of this kind is that one component was never adequate for the task, and the second most likely is that the backbone had already seen the benchmark.

- **Language model.** Zero-shot and few-shot accuracy on the text-only splits; token budget sweep (needed as the baseline for §8.3.3); calibration (ECE, reliability diagram) before any modification.
- **Generator.** Prompt adherence and spatial-relation fidelity on a small compositional probe — can it place "a red cube *behind* a blue sphere"? A generator that cannot render the relations a task asks about cannot supply useful evidence about them, and this is measurable in an hour.

- **Vision encoder.** Linear-probe accuracy on the relations of interest from *rendered* images, which upper-bounds what re-perception can recover.
- **Provenance.** Exact checkpoint hashes, revision IDs, licences, and tokenizer versions recorded to the manifest (Appendix E).

**Exit criterion.** All three probes above chance by a preregistered margin, and the composed ceiling — generator fidelity × encoder recoverability — above the text-only baseline on at least one diagnostic family. Otherwise the configuration is rejected and replaced, and this is reported.

**Is a frozen diffusion latent space even suitable?** This is the load-bearing feasibility question for the whole frozen-backbone design, and Stage 0 exists to answer it before any coupling is trained rather than to assume it away. A latent space optimised for photorealistic synthesis is not obviously organised for structured spatial reasoning: it may encode "a red cube behind a blue sphere" in a way that renders correctly while making the *relation* no easier to read off than it was in text. We therefore add an explicit probe — linear decodability of the target relations from the frozen latent, before decoding — and treat it as a gate, with a stated fallback ladder if it fails:

1. **Change the intervention point.** Read from an intermediate denoiser activation rather than the final latent; mid-network features are often better organised for relations than the output latent. Costs nothing extra and preserves the frozen design.
2. **Adapt the encoder, not the generator.** Train the vision encoder's projection (or a small adapter on it) on the relation probes. Cheap, and leaves $\psi$ untouched.
3. **LoRA on the denoiser (depth D3).** Touches $\psi$, and is the point at which the frozen-backbone claim becomes a frozen-*plus-adapters* claim. We would say so.
4. **Declare the configuration unsuitable** and report it. A generator whose latent space resists relational reading is a legitimate negative result about that generator, and is far more useful published than silently worked around with expensive fine-tuning.

The honest position is that we do not know in advance which rung is needed, that rungs 3 and 4 would each materially weaken a headline claim of cheap bootstrapping, and that the cost of finding out is one afternoon of probing rather than a training run.

### 5.2 Stage 1: cross-modal alignment with frozen backbones

All backbones frozen; train $\{\phi, \omega, \lambda_{1:K}, B_{L\to D}, B_{V\to L}, \eta\}$.

$$\mathcal{L}_1 = \underbrace{\lambda_{\mathrm{a}}\mathcal{L}_{\mathrm{align}}}_{\text{bridges}} + \underbrace{\lambda_{\mathrm{t}}\mathcal{L}_{\mathrm{task}}}_{\text{end task}} + \underbrace{\lambda_{\mathrm{s}}\mathcal{L}_{\mathrm{spec}}}_{\text{operator}} + \underbrace{\lambda_{\mathrm{q}}\mathcal{L}_{\mathrm{critic}}}_{\text{critic}} + \underbrace{\lambda_{\mathrm{c}}\mathcal{L}_{\mathrm{ctrl}}}_{\text{gate}} + \underbrace{\lambda_{\mathrm{k}}\mathcal{L}_{\mathrm{cal}}}_{\text{calibration}} . \tag{30}$$

**$\mathcal{L}_{\mathrm{align}}$.** A symmetric InfoNCE between $B_{L\to D}(h)$ and the generator's own text-encoder embedding of the paired caption, plus a reconstruction term requiring $B_{V\to L}(V_\nu(\mathcal{E}(I)))$ to predict caption tokens for image $I$. The first anchors the forward bridge in the generator's conditioning geometry; the second ensures the return path carries decodable content rather than an arbitrary code.

**$\mathcal{L}_{\text{spec}}$: the water-filling regulariser.** This term is new relative to the earlier draft, where $\mathcal{L}_{\text{spec}}$ was a vague "consistency under semantic-preserving perturbation." Corollary 12 tells us what the gains should look like, so we can regularise toward it while still allowing the data to overrule:

$$\mathcal{L}_{\text{spec}} = \underbrace{\Big\| \log |G_{\rho,\cdot k}| - \tfrac{1}{2}\log \hat{w}_k + \text{const} \Big\|_2^2}_{\text{water-filling prior}} + \underbrace{\nu \sum_k \big(\lambda_{k+1} - \lambda_k\big)_+^2}_{\text{monotone time constants}} + \underbrace{\zeta \big\| S_\omega(z) - z \big\|_2^2 / \|z\|_2^2}_{\text{distortion budget}}, \tag{31}$$

where $G_{\rho,\cdot k}$ is the *realised* band gain — the mask-weighted average of the gate the operator actually applies, not the nominal coefficient $1 + \rho a_{ck}$, since the masks overlap — and $\hat{w}_k$ is the running estimate of band sensitivity from the gradient-power surrogate. Bands are ordered coarse-to-fine, so the monotonicity term fires on $\lambda_{k+1} > \lambda_k$, encoding the prior that coarse bands persist at least as long as fine ones. The constant absorbs the global scale that Corollary 12 leaves free.

The regulariser's weight $\lambda_{\text{s}}$ must be swept down to zero as an ablation. If the gain–sensitivity correspondence only appears when we regularise for it, we have assumed our conclusion; the prediction of §8.3.1 is about *unregularised* learned gains.

**$\mathcal{L}_{\text{critic}}$ and $\mathcal{L}_{\text{cal}}$.** The critic is trained with a ranking loss over branches whose targets come from paired rollouts (below), plus a proper scoring rule (Brier or log loss) on its calibrated output. $\mathcal{L}_{\text{cal}}$ is a binned calibration penalty computed per provenance stratum; the hard cap of §3.8 is applied at inference regardless.

### 5.3 Every loss term, written out

The terms above are named in (30); here is each one as a formula, so that the pipeline is reproducible without guessing. $\hat{y}$ is the model's answer distribution, $y$ the target, $q_b$ the critic score of branch $b$, $o_b \in \{0, 1\}$ whether acting on branch $b$ was correct, $f_b$ the branch's feature vector, and $\mathbb{1}$ the indicator.

Not all of these are implemented in the reference code, and the reader should know which: the repository provides $\mathcal{L}_{\text{spec}}$, $\mathcal{L}_{\text{ctrl}}$, $\mathcal{L}_{\text{div}}$, the ranking half of $\mathcal{L}_{\text{critic}}$ and an unstratified calibration penalty. $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{recurrent}}$ are specified here but not implemented, because both require pretrained backbones that the CPU-only reference path does not have. They are marked as such in `losses.py`.

$$\mathcal{L}_{\text{task}} = -\log p_{\hat{y}}(y \mid h_T) \;+\; \underbrace{\gamma_{\text{aux}} \sum_{t<T} \omega_t \big(-\log p_{\hat{y}}(y \mid h_t)\big)}_{\text{optional per-step supervision, } \omega_t \propto t} \tag{32}$$

$$\mathcal{L}_{\text{recurrent}} = \underbrace{\big\| \mathcal{F} m_t - \Lambda \odot \mathcal{F} m_{t-1} - (1-\Lambda) \odot \mathcal{F}(T_t \tilde{z}_t) \big\|_2^2}_{\text{workspace consistency; identically zero at depth D1, a drift diagnostic at D2/D3}} + \underbrace{\nu_1 \sum_t \big\| T_t \tilde{z}_t - \tilde{z}_t^{(\text{anchor})} \big\|_2^2}_{\text{registration residual, not a pull toward the in}} \tag{33}$$

$$\mathcal{L}_{\text{align}} = -\tfrac{1}{2} \sum_i \left[ \log \frac{e^{\langle B_{L\to D}(h_i), e_i \rangle / \kappa}}{\sum_j e^{\langle B_{L\to D}(h_i), e_j \rangle / \kappa}} + \log \frac{e^{\langle B_{L\to D}(h_i), e_i \rangle / \kappa}}{\sum_j e^{\langle B_{L\to D}(h_j), e_i \rangle / \kappa}} \right] \;-\; \textstyle\sum_i \log p_{\text{cap}}\big(\text{caption}_i \mid B_{V\to L}(v_i)\big) \tag{34}$$

$$\mathcal{L}_{\text{critic}} = \underbrace{\frac{1}{|\mathcal{P}|} \sum_{(b,b') \in \mathcal{P}} \big(\varepsilon - (q_b - q_{b'})\big)_+}_{\text{pairwise ranking, } \mathcal{P} = \{(b,b') : o_b > o_{b'}\}} + \underbrace{\tfrac{1}{B} \textstyle\sum_b (q_b - o_b)^2}_{\text{Brier}} \tag{35}$$

$$\mathcal{L}_{\text{cal}} = \sum_{P \in \{\text{OBS,RET,IMG}\}} \sum_{\text{bins } I} \frac{|I_P|}{N} \Big| \overline{\mathbb{1}[\hat{y} = y]}_{I_P} - \bar{\hat{p}}_{I_P} \Big| \tag{36}$$

$$\mathcal{L}_{\text{div}} = \frac{1}{B(B-1)} \sum_{b \neq b'} \left( \frac{\langle f_b, f_{b'} \rangle}{\|f_b\| \|f_{b'}\|} \right)_+ \tag{37}$$

$\mathcal{L}_{\text{ctrl}}$ and $\mathcal{L}_{\text{spec}}$ are given in §5.4 and (31). Three notes on choices that are easy to get wrong:

*$\mathcal{L}_{task}$'s auxiliary term is optional and off by default.* Supervising intermediate steps stabilises training but also teaches the model to answer early, which fights the gate's job of deciding when to stop. If it is used, $\omega_t$ should increase with $t$ so that later states carry more weight.

*$\mathcal{L}_{recurrent}$'s terms need care, and an earlier draft of this section got both wrong.* The first term is identically zero at depth D1, where the update is executed exactly as written; it earns its place only at D2/D3, where the workspace is realised inside the sampler and the residual measures how far the implementation has drifted from the object the theory analyses. It must be written against the *aligned* input $T_t \tilde{z}_t$: written against $\tilde{z}_t$, as we first had it, its gradient pushes $T_t \to \mathrm{Id}$ and it actively fights §3.6.

The second term is a *registration* residual — how well the aligned latent matches the anchor it was registered to — and not, as we first wrote it, $\|m_t - T_t z_t\|^2$. That earlier form is minimised at $m_t = T_t z_t$, i.e. at $\Lambda \to 0$: a term that instructs the workspace to forget everything, in direct contradiction of the per-band decay design it sits next to. Both errors are the same kind of error — a plausible-looking regulariser whose minimiser deletes the mechanism — and both are why every loss term is now written out rather than named.

*$\mathcal{L}_{div}$ uses a rectified cosine, not a determinantal term.* A determinantal penalty is better behaved in principle and considerably more expensive; at the branch counts Corollary 26 recommends ($B \leq 8$) the difference is not worth the cost.

**Exit criterion.** $B_{V\to L}$ gate $g_V$ has moved away from zero; imagination-on accuracy is not *below* imagination-off accuracy on the diagnostic set; ICG within tolerance.

### 5.4 Estimating the controller's target

The controller regresses cVOI, so it needs a target with the units of task reward. This is the part of the pipeline most likely to be done badly, so we specify it carefully.

**Paired rollouts.** For a training instance and state $s_t$, run two continuations with shared randomness wherever possible: one that imagines, one that does not. The difference in realised reward,

$$\widetilde{\Delta}_t = R(\hat{y}^{\text{img}}, y) - R(\hat{y}^{\text{no-img}}, y), \tag{38}$$

is an unbiased but high-variance estimate of $\Delta_t(\hat{a})$ — the gain *realised* by the deployed reader — not of $\text{cVOI}_n(s_t)$, which is a maximum over the decoder class (Remark 9). We regress the realised quantity deliberately: it is what the system will actually obtain, it is directly estimable, and by Proposition 8 it can be negative, so the controller must be able to represent negative values. The gate rule (8) is therefore implemented with $\Delta_t$ in place of $\text{cVOI}_n$, which is exactly optimal when the deployed reader is the argmax reader and an approximation otherwise; the gap is the reader's excess risk, and shrinks as calibration improves.

Variance reduction matters more than sample count here: common random numbers across the pair, antithetic sampling of the diffusion noise, and control variates using the no-imagination confidence are the standard tools, and we report the achieved standard error rather than asserting a reduction factor.

**Doubly robust correction.** Because the behaviour policy that generated the rollouts is not the target policy, we use a doubly robust estimator combining the direct method (the critic's predicted gain) with importance weighting on the gate's propensity:

$$\hat{\Delta}_t^{\text{DR}} = \hat{g}(s_t) + \frac{\mathbb{1}[a_t = \text{IMAGINE}]}{\pi_{\text{beh}}(\text{IMAGINE} \mid s_t)} \Big( \widetilde{\Delta}_t - \hat{g}(s_t) \Big), \tag{39}$$

with $\hat{g}$ the direct-method regressor. Propensities are bounded away from zero by forcing $\epsilon$-exploration in the gate during data collection — without which the controller trains only on states it already chose to imagine in, and its estimates on the skip branch are unidentified. This is, in our judgement, the single most common way a learned gate silently fails.

**Controller objective.** With targets in hand,

$$\mathcal{L}_{\text{ctrl}} = \mathbb{E}\Big[ \big( C_\phi(h_t, u_t) - \hat{\Delta}_t^{\text{DR}} \big)^2 \Big] \; + \; \varsigma \, \mathbb{E}\big[ \text{pinball}_q\big( C_\phi, \hat{\Delta}^{\text{DR}} \big) \big], \tag{40}$$

where the optional quantile (pinball) term yields a conservative lower estimate at level $q$, so the gate can be made risk-averse about spending compute without changing $\beta$.

### 5.5 Stage 2: interleaved trajectory learning

Stage 2 trains over whole trajectories $\tau = (a_1^L, a_1^V, a_2^L, \dots)$ mixing language and visual actions. Supervision comes from three sources, in decreasing order of trustworthiness: programmatically solvable spatial tasks with exact verifiers (§D); teacher-forced traces from a stronger multimodal model; and successful self-rollouts. We keep the proportions fixed and reported, because a silent drift toward self-generated data is how Stage 3's failure modes leak into Stage 2.

$$\mathcal{L}_2 = \mathcal{L}_{\text{ans}} \; + \; \gamma \, \mathcal{L}_{\text{rank}} \; + \; \beta \sum_t \mathbb{1}[a_t = \text{IMAGINE}] \; c(a_t) \; + \; \mu \, \mathcal{L}_{\text{div}}. \tag{41}$$

The compute term uses the *same* $\beta$ as the inference gate, updated by dual ascent (11) against a target budget — so the policy is trained at the operating point it will be deployed at. $\mathcal{L}_{\text{div}}$ is a branch-diversity term (a determinantal or pairwise repulsion penalty on branch features) that counteracts mode collapse across counterfactuals; without it, $B$ branches degenerate into $B$ copies and Proposition 25's guarantee, while still formally true, becomes vacuous because $f$ is nearly modular in the degenerate regime.

**Non-myopic variant.** Replacing the myopic gate with a learned $Q$-function over the three actions and training it by TD or by policy gradient with a shared task reward is a natural extension. We treat it as an ablation rather than the default: it requires credit assignment through the full trajectory, it is the component most sensitive to reward sparsity, and if it wins we want to know how much of the win came from non-myopia rather than from extra capacity. Both variants share the cost model and the budget, so the comparison is clean.

### 5.6 Curriculum

Ordering matters more than usual here because the controller can only learn to skip if it has seen tasks where skipping is right. The curriculum is: (i) purely text tasks, where the correct policy is always to skip — a controller that cannot learn this fails immediately and cheaply; (ii) synthetic spatial tasks with exact verifiers and controllable difficulty, where imagining usually helps and the ground-truth cVOI can be computed by exhaustive paired rollout; (iii) mixed batches at a fixed ratio, which is where the gate's discrimination is actually trained; (iv) real benchmark distributions.

Stages (i) and (ii) are cheap and are where most diagnostic information lives; we recommend spending a disproportionate share of the compute budget there.

### 5.7 What we expect to go wrong

Naming the likely failures in advance is part of making the programme falsifiable rather than elastic.

- **The gate collapses to always-skip.** Most likely outcome at small scale, because $c_{\text{img}}$ is large and early-training cVOI is near zero. Diagnosis: check whether the oracle gate (imagine iff it would have helped, computed post hoc) beats always-skip. If the oracle gate is flat too, the problem is the imaginal pathway, not the controller.
- **The gate collapses to always-imagine.** Happens when $\beta$ is too small or when the critic is correlated with the label. Diagnosis: the accuracy–compute frontier from sweeping $\beta$ should be monotone; if it is flat, the gate is not discriminating.
- **Bridges learn a degenerate code.** $B_{L\to D}$ can learn a constant conditioning that makes the generator emit a fixed easy scene. Diagnosis: mutual information between $h$ and $c_t$; variance of generated latents across instances.
- **Critic learns prettiness.** Diagnosis: correlation between $Q_\eta$ and CLIP similarity should be *low*; if it is high, the critic has not learned utility.
- **Gains go to the boundary.** $|a_{ck}| \to 1$ means the distortion budget is binding; either $\rho$ is too small or the water-filling prior is fighting the task loss.

**Algorithm 2** One verified improvement cycle

**Require:** checkpoint $\Theta_j$; immutable suites $\mathcal{V}_{\text{cap}}, \mathcal{V}_{\text{cal}}, \mathcal{V}_{\text{saf}}$; reference corpus $\mathcal{R}$; gates $\mathcal{G}$

**Generate**: sample multimodal trajectories from $\Theta_j$ on the *training* pool with $\epsilon$-exploration in the gate

**Verify**: score each trajectory with an external verifier — exact task reward, deterministic simulator, unit tests, or trusted labels. Reject any trajectory whose correctness cannot be established without consulting $\Theta_j$

**Filter**: retain trajectories above a confidence threshold; deduplicate by task and by scene hash; cap per-task contribution to preserve coverage

**Balance**: form $\mathcal{B}_j$ with a fixed, reported ratio of self-generated to replayed reference data from $\mathcal{R}$

**Train candidate**: $\Theta'_j \leftarrow \text{train}(\text{copy of } \Theta_j,\ \mathcal{B}_j)$ ▷ never in place

**Evaluate**: score $\Theta'_j$ and $\Theta_j$ on $\mathcal{V}_{\text{cap}}, \mathcal{V}_{\text{cal}}, \mathcal{V}_{\text{saf}}$ with fixed seeds

**Gate**: evaluate every condition in $\mathcal{G}$ (§6.3)

**if** all gates pass **then**

    $\Theta_{j+1} \leftarrow \Theta'_j$ ▷ promote

**else**

    $\Theta_{j+1} \leftarrow \Theta_j$; record the failing gate and stop the cycle chain if two consecutive failures ▷ roll back

**end if**

# 6 Verified Recursive Improvement

## 6.1 A deliberately narrow reading

"Recursive self-improvement" is used in the literature to mean anything from prompt self-refinement to autonomous code modification. We use it in one narrow, operational sense: *repeated offline self-distillation behind immutable gates.* The model generates candidate trajectories; a verifier external to the model decides which are correct; a *separate* candidate checkpoint is trained on the survivors plus replayed reference data; the candidate is evaluated on suites it cannot see or influence; and it is promoted only if prespecified gates pass. Nothing modifies weights online, nothing modifies the training code, and nothing modifies the gates.

The reason for the narrowness is not rhetorical caution. It is that the loop has a well-documented failure mode — training on recursively generated data degrades tail coverage and can collapse distributions [42] — and that the failure is invisible to the metrics the loop itself optimises. The protocol below is designed so the failure is visible.

## 6.2 Protocol

Three details in Algorithm 2 do real work and are easy to omit.

*Verification must be model-external.* If the verifier is the model, or a sibling of the model, the loop optimises agreement rather than correctness. Where no external verifier exists for a task family, that family contributes evaluation data only, never training data.

*Training is never in place.* $\Theta_j$ is retained until the candidate has passed. This is what makes rollback meaningful, and it is why the protocol is compatible with a running system.

*Coverage is capped, not just filtered.* Retaining "high-confidence successes" is exactly the selection rule that shrinks the tails. Per-task and per-scene caps are the cheapest available counterweight, and diversity is monitored as a first-class gate rather than inferred from accuracy.

### 6.3 The gates

Gates are fixed before the first cycle and versioned. A change to a gate resets the cycle counter and is reported as such. We specify six.

| Gate | Condition | Rationale |
|---|---|---|
| Capability | $\text{Score}(\Theta'_j) \geq \text{Score}(\Theta_j) + \delta_{\min}$ on $\mathcal{V}_{\text{cap}}$, with a sequential test controlling family-wise error across cycles | improvement must be real, not noise |
| Non-regression | No individual suite in $\mathcal{V}_{\text{cap}}$ drops by more than $\delta_{\text{reg}}$ | prevents averaging a large loss into a larger gain |
| Calibration | ECE does not worsen; $\text{ICG} \leq \epsilon_{\max}$; $\text{ICD} \leq \Delta\text{Acc} + \epsilon_{\text{tol}}$ | self-distillation reliably inflates confidence |
| Diversity | Entropy of generated scene descriptors and of answer distributions does not fall below cycle-0 minus $\delta_{\text{div}}$ | direct check against collapse [42] |
| Compute | Mean $c_{\text{img}}$ per example does not increase; the accuracy–compute frontier weakly dominates the previous cycle | stops the loop from buying accuracy with compute |
| Safety | Refusal, provenance-integrity and hallucination probes do not regress | imagined content is a new surface for confident fabrication |

**Table 4:** Promotion gates. All are necessary; the capability gate alone is not sufficient.

**Statistical discipline.** Evaluating the same suites every cycle is multiple testing, and naive per-cycle significance guarantees eventual spurious promotion. We fix a total error budget across the planned cycle count and spend it with an alpha-spending rule, or equivalently use a sequential test with anytime-valid confidence sequences. The cycle count is preregistered. Reporting a cycle chain that was extended because it was still improving invalidates the guarantee, and we say so in the checklist (Appendix E).

**Held-out immutability.** $\mathcal{V}$ suites are never used for model selection within a cycle, never used to choose $\beta$, and are stored with hashes. Any tuning happens on a separate development split. This is standard and routinely violated.

### 6.4 The verifier is a measuring instrument, so measure it

Everything in Algorithm 2 rests on the verifier's judgement, and the protocol as stated so far treats that judgement as ground truth. It is not: a verifier has a false-positive rate $\alpha$ (accepting an incorrect trajectory) and a false-negative rate $\beta_v$ (rejecting a correct one), and the two have very different consequences.

**False positives poison; false negatives merely narrow.** A false positive injects a wrong trajectory into $\mathcal{B}_j$, where it is trained on as though correct. If the retained set has purity $1 - \alpha'$ (with $\alpha'$ the post-filter false-positive rate), the candidate is fitting a mixture of the target distribution and an error distribution, and the errors are not random — they are concentrated exactly where the verifier is weakest, which is where the model is most likely to be systematically wrong. A false negative discards a correct trajectory: it costs sample efficiency and narrows coverage, but it does

not teach anything false. The asymmetry says to run the verifier *conservatively*: prefer $\beta_v$ to $\alpha$, and set the retention threshold accordingly.

**What we measure, every cycle.** Verifier accuracy is estimated on a human-audited sample drawn from the retained set (not from the full generated set — the retained set is the one whose purity matters), of at least $n_{\text{audit}}$ trajectories. We report $\hat{\alpha}$ with a confidence interval, and the estimated purity of $\mathcal{B}_j$. Where a task family has two independent verifier implementations, their disagreement rate is a free lower-bound estimate of $\alpha + \beta_v$ that needs no human labour.

**A gate on the instrument, not just on the model.** If $\hat{\alpha}$ exceeds a preregistered ceiling for a task family, that family contributes evaluation data only and no training data, for that cycle. This is a gate on the *verifier*, and it is the one most likely to be quietly skipped, because it blocks progress for a reason that feels external to the model.

**How the error propagates.** A useful back-of-the-envelope: if a fraction $p$ of the retained set is wrong and the candidate fits it, the expected capability change is roughly $\Delta \approx (1-p)\,g - p\,\ell$, where $g$ is the gain from a correct trajectory and $\ell$ the damage from a wrong one. Since $\ell > g$ in general — unlearning a confidently-fitted error costs more than learning a fact — even a modest $p$ can flip the sign of a cycle while leaving training metrics (which are computed against the same poisoned set) looking healthy. That is precisely the regime the immutable held-out suites exist to catch, and it is why the capability gate is evaluated only on data the loop cannot touch.

### 6.5 Why the loop might still fail

We think it is more useful to describe the ways this protocol fails than to assert that it is safe.

- **Verifier exploitation.** The model finds trajectories the verifier accepts and humans would not. Mitigation: rotate verifier implementations, hold out a human-audited sample each cycle, and gate on the audited subset — not just on the automated one.
- **Distributional narrowing under a passing diversity gate.** Entropy over descriptors is a coarse statistic; a model can keep it constant while losing the specific tails that matter. Mitigation: track per-cluster coverage, and include rare-slice suites in $\mathcal{V}_{\text{cap}}$.
- **Gate erosion.** The most likely real-world failure is social, not technical: a gate that keeps blocking promotion gets loosened. Mitigation: version the gates, report every change with the cycle at which it occurred, and treat a loosened gate as a reset.
- **Improvement concentrated in the imagination path.** The loop may improve trajectories that use imagery while the text-only path stagnates, producing a system that is better only when it spends compute. The compute gate catches this; the accuracy–compute frontier is the object to report, not the accuracy.
- **Self-confirmation through re-perception.** Cycles can teach the model to generate scenes its own critic likes. Mitigation: the critic is retrained on *externally verified* outcomes only, and critic–CLIP correlation is monitored as a drift indicator.

### 6.6 Relation to the earlier draft

The earlier draft specified the six-step cycle and the promote-or-rollback rule, which is the correct skeleton. What it lacked, and what we regard as necessary for the claim to mean anything, is: statistical error control across cycles; explicit diversity and compute gates; calibration gates tied to the theory of §3.8; a requirement that verification be model-external; per-task coverage caps; and a stated account of how the protocol is expected to fail. The reference implementation now exposes all gates as data rather than as a single scalar comparison (§B).

## 7 Preliminary Evidence from a Controlled Testbed

### 7.1 What these results are, and what they are not

Everything up to this point is theory and specification. This section reports the first *measurements*: five pilot experiments run on a synthetic world, on a single CPU, with no pretrained backbones.

We are explicit about their standing, because the temptation to over-read a pilot is exactly the failure mode §9 was written against.

- **These are not benchmark results.** No MMMU, MathVista, ScienceQA or Winoground number appears anywhere in this paper. Table 13 remains empty.
- **They are not evidence about pretrained models.** The generator, encoder and reader here are small and purpose-built. Nothing measured transfers to a claim about a diffusion model or an LLM.
- **They are evidence about the mechanism.** The paper's claims are mechanistic — band gains track sensitivity, a priced gate abstains, alignment beats naive fusion, the benefit is computational and so decays with reader capacity. On a real benchmark *none of those quantities is observable*: one cannot measure the true band sensitivity of MMMU, and one cannot vary a frozen backbone's boundedness. Here every one of them is known by construction.
- **Passing here is necessary, not sufficient.** If the mechanism fails in a world built to satisfy the theory's assumptions, it will not work where they only approximately hold.

### 7.2 SpectralWorld

A scene is a latent built from a *layout* component (a few Gaussian blobs, energy concentrated at low radial frequency) and a *texture* component (band-limited noise on one side of the image, energy at high radial frequency). Three task families have exact ground truth: LAYOUT (is blob A left of blob B?), TEXTURE (which half carries the fine detail?), and TEXT_ONLY (parity of a number given in the prompt).

The reader receives the full scene code through a fixed high-bandwidth random Fourier map. This is the construction that makes the setting faithful to Proposition 6: no information is withheld, so imagery genuinely adds none — but recovering a *relational* predicate from an oscillatory encoding is expensive for a small reader and trivial from a rendered scene. The reader's width is the bounded-decoder dial. Appendix F gives every constant.

### 7.3 Pilot A: do learned band gains track measured sensitivity?

This tests Prediction 1, the paper's sharpest claim. The protocol is designed so the answer is not built in: encoder and reader are pre-trained on clean latents and **frozen**; a band-independent noise

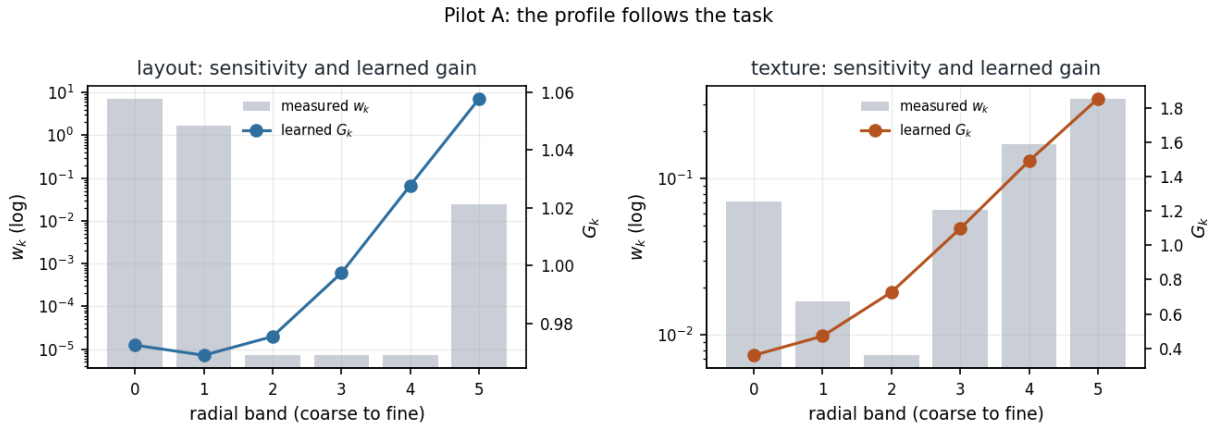


**Figure 2:** Measured sensitivity $w_k$ (bars, log scale) and the gains a task-loss-only gate learned (line), at $\rho = 0.9$, seed 0. For TEXTURE the gate rises monotonically across the six bands and correlates with sensitivity (Spearman 0.60 for this run); note that $w_k$ itself is not monotone in $k$ — the coarsest band has the third-highest sensitivity and the lowest gain. For LAYOUT the sensitivity is concentrated almost entirely in the two coarsest bands, with three bands at the estimator's floor, and the learned gain is nearly flat: there is little profile to match, and the gate does not match it.

floor is inserted *after* the gate; the gated latent is renormalised so only the allocation, not a global scale, can help; and then *only the* $CK = 24$ *gate coefficients* are trained, by task loss alone, with the water-filling regulariser off. Sensitivity $w_k$ is measured afterwards against the frozen pipeline by finite-difference curvature — nothing from gate training enters it.

The two families behave differently, and only one supports the prediction.

| Family | $\rho$ | Fitted slope (95% CI) | | Ceiling | Slope/ceiling | Bands fit | Direction |
|---|---|---|---|---|---|---|---|
| layout | 0.2 | $-0.008$ | $[-0.014, -0.002]$ | 0.078 | $-0.12$ | 4.6 | 0/5 |
| layout | 0.5 | $-0.002$ | $[-0.016, +0.012]$ | 0.214 | $-0.02$ | 4.6 | 1/5 |
| layout | 0.9 | $-0.001$ | $[-0.013, +0.016]$ | 0.566 | $-0.00$ | 4.6 | 1/5 |
| texture | 0.2 | $+0.033$ | $[+0.004, +0.053]$ | 0.064 | $+0.51$ | 4.8 | 4/5 |
| texture | 0.5 | $+0.078$ | $[+0.041, +0.116]$ | 0.156 | $+0.61$ | 5.0 | 4/5 |
| texture | 0.9 | $+0.151$ | $[+0.031, +0.272]$ | 0.394 | $+0.35$ | 5.0 | 4/5 |

**Table 5:** Pilot A, five seeds per cell. Slope of $\log G_k$ on $\log w_k$; the prediction is 0.5. "Ceiling" is log(realisable $G$ range)/ log($w$ range) — the largest slope the bounded operator could produce given that run's measured sensitivity range — averaged over seeds like every other column. "Slope/ceiling" is the mean of the per-seed ratios and so is not the ratio of the two columns beside it. "Bands fit" counts bands whose sensitivity was measured rather than floored by the estimator; floored bands are excluded, because including them injects a fabricated abscissa spanning six decades.

**Texture: supported.** The slope is positive at every $\rho$, with a bootstrap CI excluding zero in all three cells, and it rises monotonically with the operator's range ($+0.033 \to +0.078 \to +0.151$) — which is what (16) predicts, since a wider operator can express more of the profile. The gain-carrying band matched the sensitivity-carrying band in 4 of 5 seeds throughout. The slope reaches 35–61% of what the operator could express, so the allocation is real but partial; no cell's CI contains the nominal 0.5, and the ceilings themselves are below 0.5 at $\rho \leq 0.5$, so the nominal value was never attainable there.

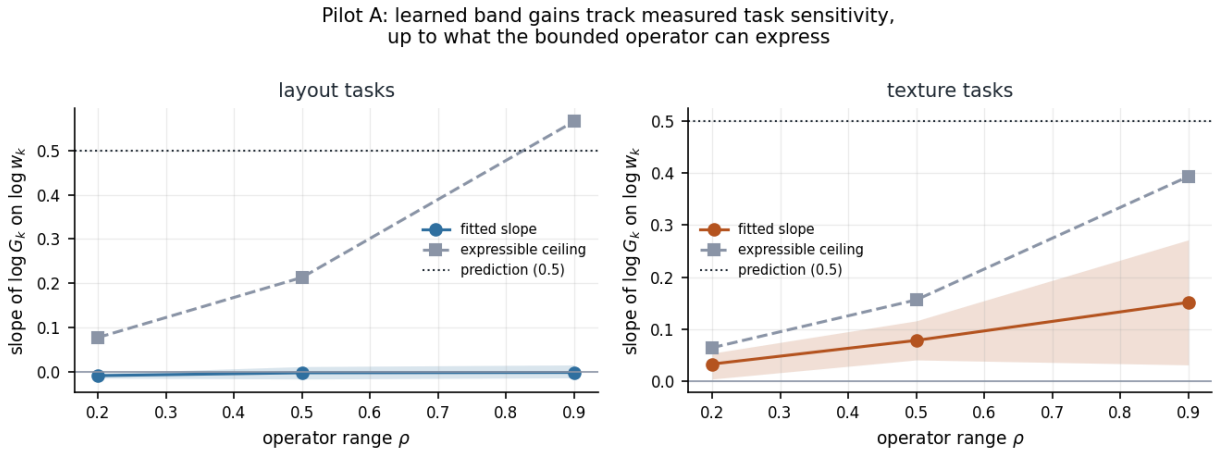


**Figure 3:** Pilot A across the $\rho$ sweep, five seeds, shaded bootstrap CIs. Texture (right) rises with the operator's expressible ceiling; layout (left) sits on zero. Reporting the ceiling alongside the slope matters: on the right, a slope of +0.03 at $\rho = 0.2$ is half of what the operator can express, not a failure to allocate.

**Layout: not supported.** The slope is indistinguishable from zero at every $\rho$ (and significantly *negative* at $\rho = 0.2$), and the direction match is 0–1 of 5. The explanation visible in the data is that layout's sensitivity is concentrated in two bands with three more at the estimator's floor (Figure 2), leaving a profile that is close to a step function; there is little for a smooth radial gate to track, and it does not track it. Whether that is a limitation of the task, of the sensitivity estimator at a near-flat optimum, or of the prediction itself, this pilot cannot separate.

**An earlier version of this analysis was wrong, in our favour.** Before the fix described in Appendix F, floored sensitivity values entered the log-log regression and the expressible ceiling was taken from a single seed rather than averaged. That version reported slopes roughly three times larger, a layout cell whose CI contained 0.5, and direction matches of 4/5 everywhere. Every one of those was an artefact. We record it because the difference between the two analyses is much larger than the difference between the theory and a null, and because a reader has no way to detect this class of error from a results table alone.

### 7.4 Pilot B: does the gate abstain, and does band allocation collapse?

The reviewer's question was specifically about initialisation. Here it is answered by training the gate and the allocator against *measured* quantities — realised paired gain for the gate (both continuations actually run), band-ablation marginal utility for the allocator — with neither told which task family it is looking at.

Three things, one of them uncomfortable.

*The gate learned the true utility structure, not the intended one.* It converges on TEXTURE at rate 1.00 on every seed, with a predicted gain of +0.47, and predicts approximately zero gain for LAYOUT (+0.001) and TEXT_ONLY (−0.003). The LAYOUT outcome was not what we hoped for and is correct: in this pilot the small convolutional encoder cannot recover the left-of relation, so imagery genuinely does not help layout (Pilot C confirms it: pooled across widths, layout accuracy is 0.565 without imagery and 0.558 with it, while texture goes from 0.572 to 0.928). A gate trained on realised gain tracked reality rather than our intentions, which is the behaviour one wants and the reason to regress a measured quantity instead of a heuristic.

| | LAYOUT | TEXTURE | TEXT_ONLY |
|---|---|---|---|
| Imagination rate, initial | 0.73 | 0.74 | 0.65 |
| Imagination rate, trained | 0.50 [0.31, 0.76] | **1.00** [1.00, 1.00] | 0.48 [0.39, 0.54] |
| Predicted gain, trained | +0.001 [−0.051, +0.069] | +**0.474** [+0.451, +0.506] | −0.003 [−0.021, +0.014] |
| Mean active bands (of 6) | 0.14 [0.01, 0.26] | 0.18 [0.01, 0.41] | 0.12 [0.00, 0.25] |

**Table 6:** Pilot B, three seeds; brackets give the min–max spread across seeds, which is wide enough that a single-seed table would misrepresent the result. Allocation entropy ends at 0.385 [0.271, 0.457] against a maximum of 0.693.

*But the gate does not abstain, and by our own criteria that is a failure.* §9 states falsifier C5 as: the gate must suppress imagery on the text-only control, with the imagination rate approaching zero. It does not. The trained rate on TEXT_ONLY is 0.48 — statistically indistinguishable from the 0.50 it assigns to LAYOUT, another family where imagery is worthless. What the gate learned is the *value*, not the *policy*: its predicted gains separate the families cleanly (+0.47 versus ≈ 0), but the decision rule used here fires whenever predicted gain exceeds zero, and near-zero predictions scatter around that threshold. Two things follow. The pilot's rule is not the manuscript's rule — Eq. (8) compares against $\beta c_{\text{img}} > 0$, and with any positive shadow price these near-zero predictions fall below it — so **the pilot did not test the priced gate at all**, and we should not have described it as one. And the correct reading is that the value estimator passed and the thresholding did not, which is a narrower claim than "the gate learns to abstain" and the one the data supports.

*Allocation does not collapse, but the evidence is weaker than it looks.* Mean active bands ends between 0.12 and 0.18 of six with entropy 0.385 of a maximum 0.693, so the anticipated default-to-all-bands failure did not occur — and it did occur, visibly, before we added the negative initialisation bias. Two caveats. The pilot counts active bands as $\pi_k > 0.5$ rather than by the priced rule (29), so it exercises the allocator's probabilities and not its knapsack. And the band-ablation utility it regresses against was identically zero in 17 of 33 logged steps, so for half the run the target was the zero vector; non-collapse under a zero target is weak evidence about a collapse mechanism driven by task loss. The initialisation fix is supported; the claim that it would hold up under strong task pressure is not yet.

### 7.5 Pilots C and D: matched compute, and the budget-sweep decay test

Four conditions across reader widths from 8 to 256, three seeds each. The three imagery conditions use an equal number of generator calls (two per example); the prompt-only condition uses none, and is the capacity baseline rather than a compute-matched one.

| Reader width | 8 | 16 | 32 | 64 | 128 | 256 |
|---|---|---|---|---|---|---|
| Prompt only | 0.691 | 0.694 | 0.708 | 0.707 | 0.739 | 0.757 |
| Generic imagery | 0.833 | 0.835 | 0.830 | 0.823 | 0.825 | 0.840 |
| Spectral, no transport | 0.835 | 0.830 | 0.831 | 0.823 | 0.826 | 0.843 |
| Spectral workspace | 0.833 | 0.835 | 0.834 | 0.825 | 0.820 | 0.845 |
| Advantage over prompt only | +0.141 | +0.141 | +0.126 | +0.118 | +0.082 | +0.088 |
| Advantage over generic imagery | −0.001 | +0.000 | +0.004 | +0.003 | −0.004 | +0.005 |

**Table 7:** Pilots C and D. The three imagery conditions use an equal number of generator calls; prompt only uses none. The benefit of imagery decays as the reader's capacity grows; the benefit of the *spectral* workspace over generic imagery does not separate from zero in this setting.

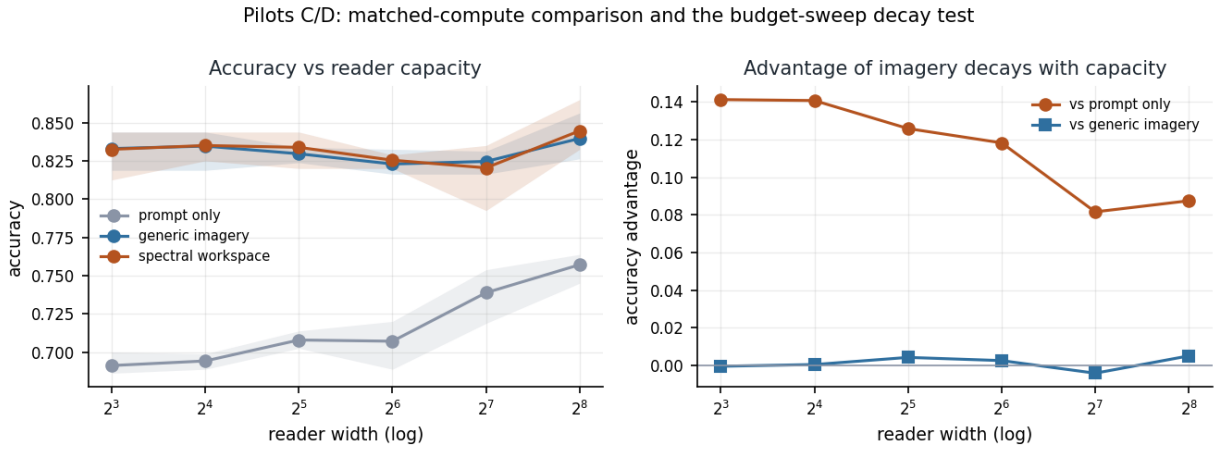


**Figure 4:** Pilots C and D. Left: accuracy against reader capacity; the prompt-only reader improves with capacity while the imagery conditions are flat. Right: the advantage of imagery decays, as Prediction 3 requires, while the advantage of the spectral workspace over generic imagery sits on zero throughout. Shaded bands are bootstrap 95% CIs over three seeds.

**Prediction 3 is supported.** The advantage of imagery over the prompt-only reader falls from +0.141 at width 8 to +0.088 at width 256, while the prompt-only reader climbs from 0.691 to 0.757 and the imagery conditions stay flat (Figure 4). This is the signature the theory demands: the benefit is computational, so it erodes as the reader becomes able to do the computation itself. It is also, to our knowledge, the first direct test of that prediction in this subarea, and it was cheap.

**The spectral mechanism does not separate from generic imagery here.** The advantage of the full workspace over an equal-compute generic-imagery condition ranges from −0.004 to +0.005 across widths — indistinguishable from zero, with three seeds. Transport contributes similarly little (−0.006 to +0.005).

We report this prominently rather than in a footnote, because it is the outcome §9 listed as refuting claim C2, and because it is genuinely informative about *where* the mechanism could matter. Two features of the pilot bear on how far it generalises. The task that imagery helps with here (TEXTURE) already reaches 0.93 pooled with generic imagery — 0.97 at the narrowest reader — leaving little headroom for an allocation mechanism to recover; and the pilot's fusion runs for only two steps under ±3-pixel jitter, which is a mild version of the multi-step, large-transformation regime where Proposition 23 says unregistered fusion should break down. A fair statement is therefore: *in this pilot, at this headroom and this number of steps, the spectral workspace bought nothing over generic imagery.* Whether it does so in a regime with headroom and depth is exactly what §8 is designed to find out — and if the answer there is also null, C2 fails and we will say so.

### 7.6 Pilot E: latent transport under viewpoint variation

Alignment retains substantially more high-frequency energy and reduces registered reconstruction error by a factor of 6.5, with disjoint confidence intervals. This is Proposition 23 measured: unregistered fusion low-passes the memory, and it does so progressively.

**Registration accuracy.** The known cyclic shift was recovered *exactly* in 100% of trials at observation noise 0, 0.3, 1.0 and 3.0 — phase correlation is exact for a pure translation and reads

| Quantity | Naive fusion | | Aligned fusion | |
|---|---|---|---|---|
| Retained high-frequency energy | 0.580 | $[0.563, 0.594]$ | **0.686** | $[0.678, 0.693]$ |
| Reconstruction MSE (registered) | 0.160 | $[0.150, 0.169]$ | **0.0247** | $[0.0246, 0.0249]$ |

**Table 8:** Pilot E, 30 episodes of 6 fusion steps under ±3-pixel jitter and observation noise 0.3, with the acceptance guard open ($\eta = 0$) so that alignment is exercised on every step. Reconstruction is measured after registering the final memory to the truth, so that sharpness is not confounded with a constant offset. *Unregistered* reconstruction MSE runs the other way — naive 0.382 $[0.353, 0.406]$ versus aligned 0.616 $[0.593, 0.643]$ — because the aligned memory is sharp but sits at the first view's offset, and an unregistered comparison rewards a blurred, centred memory. Both are in the released JSON. Bootstrap 95% CIs.

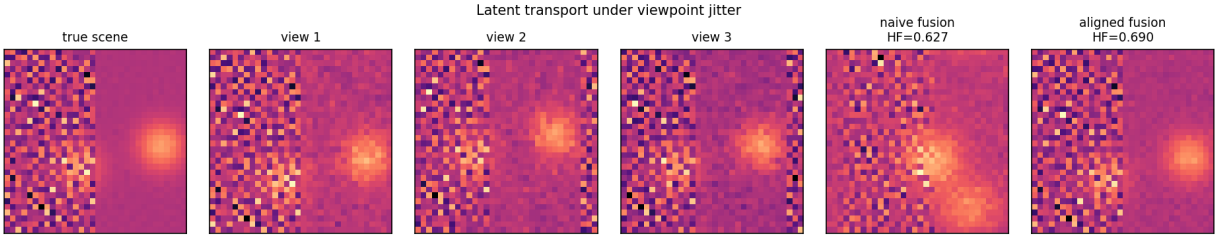


**Figure 5:** Pilot E, qualitative. The true scene, three jittered views, and the fused memory without and with alignment. Naive fusion smears the blobs; aligned fusion accumulates them. "HF" is retained high-frequency energy.

only phase, which is the part of the spectrum the gate provably leaves alone (Prop. 16).

**Failure modes, and the guard.** The anticipated failure is a pathological warp: the operator aligning two scenes that are not views of one another. Sweeping the acceptance margin $\eta$ measures the trade-off directly:

| $\eta$ | 0.00 | 0.02 | 0.05 | 0.10 | 0.20 |
|---|---|---|---|---|---|
| Accept rate, genuine translation | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| Accept rate, unrelated scenes | 0.667 | 0.550 | 0.350 | **0.200** | 0.083 |

Genuine translations are accepted at every threshold, so the margin costs nothing on the cases it should accept; unrelated pairs fall from 67% to 8%. **We changed the default from $\eta = 0.02$ to $\eta = 0.10$ on the strength of this curve** — the original value was chosen by intuition and was, on measurement, too permissive. Displacements beyond the plausible range are rejected outright: with the shift bound tightened to 5% of the grid, a $(13, 11)$ displacement on a $32 \times 32$ latent is accepted at rate 0.000. At the default bound of 50% that same displacement is inside the plausible range and would be accepted — the bound is a statement about what counts as the same scene, and it has to be set for the expected magnitude of viewpoint change.

### 7.7 Measured component costs

Against the cost model, the spectral operator's share of one imagination step is $4.9 \times 10^{-5}$ (cheapest configuration) to $5.8 \times 10^{-7}$ (widest). The workspace is between four and six orders of magnitude cheaper (4.3 at the cheapest configuration, 6.2 at the widest) than the generator it accompanies — which is what licenses §8.2 to match compute on generator calls alone.

| Latent | Gate (ms) | Memory (ms) | Transport (ms) | Workspace total (ms) | Working set (MB) |
|---|---|---|---|---|---|
| $16^2$ | 0.51 | 0.57 | 0.32 | 1.40 | 0.11 |
| $32^2$ | 0.63 | 0.75 | 0.38 | 1.76 | 0.42 |
| $64^2$ | 2.06 | 2.77 | 1.86 | 6.70 | 1.67 |
| $128^2$ | 8.13 | 11.04 | 11.24 | 30.41 | 6.66 |

**Table 9:** Measured on one CPU core, batch 8, $C = 4$, $K = 6$, best-of-seven rounds. A 64× increase in coefficients costs 21.7× the time — *sub*linear, because the smallest latents are dominated by fixed overhead rather than by the transform. Quasi-linear scaling (Table 3) predicts ≈ 112×; the measurement is an upper bound on cost, not a confirmation of the exponent.

### 7.8 Scorecard, and what the pilots changed

| Prediction | Outcome | Detail |
|---|---|---|
| $G_k \propto \sqrt{w_k}$, texture family | **supported** | slope $> 0$ at every $\rho$, CI excludes zero, rises with the operator's range; 35–61% of the expressible ceiling |
| $G_k \propto \sqrt{w_k}$, layout family | **not supported** | slope ≈ 0; direction matches 0–1 of 5 seeds |
| Gate learns the value structure | **supported** | predicted gain +0.47 on the helped family, ≈ 0 on the other two |
| Gate abstains on the control (C5) | **not supported** | text-only rate 0.48, not near zero; the pilot's rule was unpriced |
| Band allocation does not collapse | **weakly supported** | sparse and non-degenerate, but under a zero target for half the run |
| Alignment beats naive fusion | **supported** | registered reconstruction MSE 6.5× lower, disjoint CIs |
| Benefit decays with reader capacity | **supported** | advantage +0.141 → +0.088 over reader widths 8 → 256 |
| Spectral beats generic at matched compute | **not supported** | −0.004 to +0.005 across six widths |

**Table 10:** Eight testable statements, four supported, one weakly, three not. This is the summary we would want a reader to take away, and it is less flattering than the one an earlier draft of this section reported.

Pilots are only worth running if they can change something. These did:

1. $\rho = 0.2$ **is too small** whenever measured sensitivity spans more than ≈ 1.78 — the *realisable* range, which is tighter than the theoretical 2.25; $\rho$ should be chosen from the measured range (§C).
2. **The transport acceptance margin moved from** 0.02 **to** 0.10, on the measured trade-off curve rather than on intuition.
3. **The allocator needs a negative initialisation bias**, confirmed by the collapse we observed before adding one.
4. **Sensitivity must be measured as curvature, not gradient power**, and with an adaptive probe magnitude — our first estimator returned zero for every band on a well-fitted model, which reads as "no sensitivity" and actually means "probe too small".

5. **The spectral mechanism has not yet earned its place.** It did not separate from generic imagery, and the gain–sensitivity correspondence held in only one of two task families. The build order in §4.12 is written so that this component is added only after the cheaper ones have demonstrably paid.
6. **Floored sensitivity values must be excluded from the log-log fit**, and the expressible ceiling must be averaged across seeds like everything else. Getting either wrong inflated the reported slopes roughly threefold in our own first analysis.
7. **The pilot's gate rule was not the manuscript's gate rule.** Thresholding predicted gain at zero is not Eq. (8), and the difference is exactly why the text-only abstention criterion failed. The next iteration must price the decision.

# 8 Evaluation Programme

The evaluation must answer four questions, in order, and stop at the first one that fails:

**Q1** Does internal imagery help *at all*, against an equal-compute text-only reader?

**Q2** Does the *spectral, band-structured* workspace add anything beyond generic generated imagery at matched compute?

**Q3** Does the controller invoke imagery *selectively*, and are its cVOI predictions calibrated?

**Q4** Does the mechanism behave as the theory says — gains tracking sensitivity, returns saturating, benefit decaying with reader budget?

Q4 is the part that distinguishes this programme from the standard "our system scores higher" protocol, and it is the part we would run first if forced to choose, because it is cheap and because a mechanism claim that fails Q4 is not rescued by a good score on Q1.

## 8.1 Benchmarks

The text-only control suite deserves comment. It is the only benchmark in the table where the target behaviour is *not* accuracy: the model should score the same as its backbone while calling the generator approximately zero times. A system that imagines on text-only tasks has not learned the policy the paper is about, however well it scores elsewhere.

## 8.2 Baselines

Comparisons are only interpretable at matched compute, measured by (5) and reported alongside wall-clock and energy.

1. **B0** — backbone, no generated imagery, default token budget.
2. **B1** — backbone with token budget increased until its cost equals VIVID's mean cost. *This is the critical baseline for Q1*: it asks whether the compute would have been better spent on more thinking.
3. **B2** — backbone + ordinary generated images, same number of diffusion calls, no spectral operator, no workspace memory. *The critical baseline for Q2.*

| Benchmark | Capability probed | Primary reporting |
|---|---|---|
| MMMU / MMMU-Pro [49, 50] | expert multi-discipline visual understanding and deliberate reasoning | accuracy, cost, ECE |
| MathVista [30] | mathematical reasoning grounded in charts, figures, visual context | accuracy, cost |
| ScienceQA-IMG [29] | diagram-conditioned science reasoning | accuracy |
| Winoground [45] | compositional image–text relation understanding | text / image / group score |
| BLINK [8] | core visual perception abilities that VLMs still fail | accuracy per sub-task |
| Spatial/Maze suite (ours) | route planning, mental rotation, viewpoint change | exact match, steps-to-solve |
| CounterfactualScene (ours) | imagined edits, object permanence, relational transformation | accuracy, consistency |
| Text-only control (ours) | tasks where imagery is provably useless | **imagination rate** (should $\to 0$) |

**Table 11:** Evaluation suite. Third-party data is obtained under its own licence; the four "(ours)" suites are programmatically generated with exact verifiers (Appendix D) and are regenerable with new seeds, which makes them contamination-proof by construction.

4. **B3** — B2 + scalar-decay latent memory (the earlier draft's mechanism), isolating band-wise decay.
5. **B4** — VIVID with a learned dense mixing matrix replacing the radial band basis, isolating the value of the fixed interpretable basis.
6. **B5** — VIVID with the controller forced always-on.
7. **B6** — VIVID with an oracle controller (imagine iff it would have helped, computed post hoc). Not deployable; it upper-bounds what any gate could achieve and tells us whether a disappointing result is the gate's fault or the workspace's.
8. **B7** — full VIVID.

B6 is the diagnostic we would most regret omitting. Without it, a null result is uninterpretable: it could mean imagery does not help, or that the gate cannot tell when it does.

## 8.3 Theory-driven tests

These four tests exist only because of §3. To our knowledge none has been reported in this subarea.

### 8.3.1 Gain–sensitivity correspondence

**Prediction 1.** Realised band gains satisfy $G_k \propto \sqrt{w_k}$ up to a global positive constant, where $w_k$ is the independently measured task sensitivity (12).

*Procedure.* Train with $\lambda_s = 0$ (no water-filling prior). For each task family, measure $w_k$ on held-out instances with the finite-difference curvature estimator (13), and separately with the gradient-power surrogate as a cheap cross-check. Take $G_k$ to be the *realised* gain — the mask-weighted average of the applied gate — not the nominal coefficient, since the masks overlap and no frequency sees one coefficient alone.

The prediction is about proportionality, so a rank correlation is not sufficient to test it: Spearman is invariant to any monotone transform and cannot distinguish $\sqrt{w_k}$ from $w_k$ or $w_k^3$. The primary statistic is therefore the **fitted slope** $\hat{\eta}$ in $\log G_k = \eta \log w_k + \text{const}$, tested against $\eta = 1/2$, with its confidence interval; the coefficient of determination is reported alongside, and Spearman only as a weaker companion for the ordering. Pool across seeds and use a permutation null over band labels.

Report per task family, since the prediction is that the *ordering changes with the task* — low bands dominant for rotation and route planning, high bands for chart reading and counting. Report the measured dynamic range $\max_k w_k / \min_k w_k$ next to $\rho$: by (16) the operator cannot express a ratio above $((1+\rho)/(1-\rho))^2$, so if the measured range exceeds the achievable one, the slope test is uninformative and only the ordering can be assessed.

*Interpretation.* A significant positive correlation supports the rate-allocation account. A null result falsifies the derivation in §3.3 even if VIVID scores well, and we would report it as such: the architecture would then be working for a reason we do not understand, which is a materially weaker claim than the one this paper makes.

### 8.3.2 cVOI calibration

**Prediction 2.** The controller's predicted $\widehat{\text{cVOI}}$ is calibrated against realised gain: binning examples by $\widehat{\text{cVOI}}$, the mean realised paired gain (38) within each bin should lie on the identity line.

Report the reliability diagram, its slope (a slope $< 1$ indicates over-prediction of value, which will overspend compute), and the induced accuracy–compute frontier from sweeping $\beta$. A gate that is merely *ranked* correctly but poorly calibrated will not transfer across budgets, which is the main practical claim we make for the cVOI formulation.

### 8.3.3 The budget-sweep decay test

**Prediction 3.** Since imagined evidence carries no information (Prop. 6), VIVID's advantage over the text-only reader must approach zero as that reader's own budget $n$ grows toward the regime in which it can compute the same functional itself.

The limit is what the theory gives; monotone decay along the way is an expectation, not a theorem, since a bounded reader's accuracy need not improve smoothly in $n$. We test the trend and report it as such.

*Procedure.* Sweep B1's token budget over at least a decade and plot $\text{Score}_{\text{VIVID}} - \text{Score}_{\text{B1}}(n)$ against $n$ at matched total cost. A decaying curve supports the computational-benefit account. A curve that is *flat or rising* across the whole sweep is the interesting negative result: it would indicate the gain comes from something other than easing the reader's computation — capacity, ensembling, or an evaluation artefact — and should trigger a leakage audit before anything is claimed.

We regard this test as the strongest available guard against the failure mode where a multimodal reasoning system is really just a more expensive sampler.

### 8.3.4 Branch saturation

Plot both the selected-branch critic value $f(\mathcal{S}_B)$ and accuracy against $B \in \{1, 2, 4, 8, 16\}$ at fixed everything else. Corollary 26 requires the *critic value* to be concave; accuracy inherits concavity only under an additional assumption we do not prove, so a mildly sigmoidal accuracy curve is not by itself alarming. The signature to audit is accuracy still climbing linearly after the critic value has visibly saturated. We preregister that conjunction as a stop-and-audit condition rather than as a result.

### 8.4 Mechanistic ablations

Table 12 lists the ablations with the discriminating prediction stated in advance for each. Writing down both columns before running anything is what turns these from demonstrations into tests: for most of them, the outcome that would be reported as "the component matters" and the outcome that would indicate the component is inert are different *patterns*, not different magnitudes, so they cannot be reconciled after the fact.

| Manipulation | Expected effect if the account is right | Expected if it is wrong |
|---|---|---|
| Suppress low bands (zero $M_{\text{low}}$ in the gate) | selective damage to layout/rotation/route tasks | uniform degradation |
| Suppress high bands | selective damage to counting, chart reading, fine relations | uniform degradation |
| Randomise learned gains | loss of the task-specific pattern, near-B2 performance | no change (operator was inert) |
| Freeze gains at initialisation ($a = 0$) | exactly B3 by construction — a correctness check | any difference indicates a bug |
| Destroy phase, keep amplitude | catastrophic on spatial tasks, mild on texture-statistics tasks | uniform |
| Uniform $\lambda$ across bands | selective damage to multi-step transformation tasks | no change |
| $\lambda_k \to 0$ (memory reset each step) | damage grows with required reasoning depth | no change |
| Sweep $\rho \in [0, 0.5]$ | inverted-U: too little is inert, too much distorts | monotone or flat |
| Disable re-perception (use $z$ directly) | loss proportional to encoder's added abstraction | no change |
| Replace $Q_\eta$ with CLIP similarity | drop on counterfactual tasks, little change on descriptive | no change |
| Learned dense basis (B4) | comparable score, *loses* gain–sensitivity correspondence | — |

**Table 12:** Ablations, with the discriminating prediction stated in advance. The right-hand column is what makes these tests rather than demonstrations.

Two mechanical notes, because both are easy to get wrong. *Band suppression is not $a_{ck} = -1$*: at the reference $\rho = 0.2$ that yields $G_\rho = 0.8$, a 20% attenuation rather than removal. To actually remove a band one zeroes its mask contribution in the gate (or runs the ablation at $\rho = 1$, which then voids the distortion bound of Prop. 18 and must be reported as such). *Freezing gains at $a = 0$ yields B3, not B2*: the operator becomes the exact identity, but the band-wise workspace memory keeps running, and B2 is defined without memory. Both distinctions matter because these two rows are advertised as bug detectors, and a detector that fires on a definitional mismatch is worse than none.

### 8.5 Metrics

Beyond task accuracy we report, for every configuration:

- **Cost**: imagination calls per example, diffusion steps, branches, generated tokens, normalised FLOPs (5), wall-clock, peak accelerator memory, and an energy proxy.

- **Gate behaviour**: imagination rate overall and per task family; accuracy conditional on imagining and on skipping; oracle-gate gap (B6 − B7); accepted-branch fraction.
- **Calibration**: ECE overall and stratified by provenance; ICG; ICD against ΔAcc (Cor. 27); reliability of the imagination channel $\pi_{\text{IMG}}$.
- **Workspace**: learned $|G_k|$ and $\lambda_k$ per band; effective time constants $\tau_k$; measured $w_k$; the correlation of Prediction 1.
- **Efficiency**: the accuracy–compute frontier, and the scalar summary

$$\text{ImaginalEfficiency} = \frac{\text{Score}_{\text{VIVID}} - \text{Score}_{\text{B1}}}{\text{mean imagination FLOPs per example}}, \tag{42}$$

  where the numerator uses the *equal-compute* baseline B1, not the naive backbone B0. (The earlier draft used the no-image baseline, which flatters the method by crediting it with the value of the extra compute itself.)

### 8.6 Statistical protocol

Five seeds minimum for every trained configuration; results reported as mean with a bootstrap confidence interval, never as a single run. Paired comparisons use the same instances and, where possible, common random numbers. Multiple-comparison correction across the ablation table is applied and stated. A power analysis is run *before* the experiments to determine the instance count needed to detect the smallest effect we would consider meaningful (we suggest 2 accuracy points on the diagnostics, 1 point on the standard benchmarks); if the available data cannot support that power, the comparison is reported as underpowered rather than as null.

### 8.7 Contamination control

Pretrained backbones may have seen benchmark content, and for a system whose generator was trained on web images this risk extends to the imagery. Controls: report exact checkpoints and revisions; prefer MMMU-Pro's contamination-resistant settings; include the programmatically generated suites, regenerated with fresh seeds at evaluation time; separate zero-shot from fine-tuned results; and run a memorisation probe on the diagnostic generators (accuracy on instances whose surface form is perturbed but whose answer is unchanged should not drop).

### 8.8 Reporting template

| Method | MMMU-Pro | MathVista | ScienceQA | Winoground | Spatial | Img calls | nFLOPs |
|---|---|---|---|---|---|---|---|
| B0 backbone only | – | – | – | – | – | 0 | – |
| B1 backbone, equal compute | – | – | – | – | – | 0 | – |
| B2 generated images, no spectral | – | – | – | – | – | – | – |
| B3 + scalar memory | – | – | – | – | – | – | – |
| B4 + learned dense basis | – | – | – | – | – | – | – |
| B5 always-imagine | – | – | – | – | – | – | – |
| B6 oracle gate (not deployable) | – | – | – | – | – | – | – |
| **B7 full VIVID** | – | – | – | – | – | – | – |

**Table 13:** Reporting template. **Dashes denote unmeasured results.** No benchmark values appear anywhere in this manuscript; none have been run.

# 9 Falsification Conditions

A proposal that cannot be wrong is not worth running. We state the conditions under which we would regard the central claims as refuted, with a decision rule for each rather than a qualitative description. "Refuted" here means the specific claim fails; some failures leave the system useful while removing the paper's explanation for it, and we distinguish those cases explicitly because the temptation to keep the score and quietly drop the theory is exactly what this section exists to resist.

## 9.1 Claim-level falsifiers

| Claim | Refuted if | Consequence |
|---|---|---|
| C1: imagery helps | B7 does not beat B1 (equal compute) by the preregistered margin on any benchmark family, with adequate power | The whole programme fails. Report the null; do not fall back to B0 comparisons. |
| C2: the *spectral workspace* is the mechanism | B2 (equal diffusion calls, no operator) matches B7 within the CI | The imagery may still help, but this paper's mechanism does not exist. Retitle honestly. |
| C3: band structure is task-relevant | Prediction 1 null: no correlation between learned $\lvert G_k \rvert$ and measured $\sqrt{w_k}$, or band ablations degrade uniformly | The rate-allocation derivation (§3.3) is wrong. C2 may survive by another route, which must then be identified. |
| C4: recurrence matters | B3 (scalar memory) matches B7; uniform-$\lambda$ ablation shows no selective damage | Band-wise memory is unmotivated; revert to the simpler operator. |
| C5: gating is learnable | The gate cannot suppress imagery on the text-only control suite (imagination rate does not approach 0), or the accuracy–compute frontier from sweeping $\beta$ is flat | The bounded-rational framing fails operationally even if the theory is sound. |
| C6: benefit is computational | Prediction 3 null: the advantage over B1 does not decay with the reader's budget | Contradicts Prop. 6 applied to this system, which means either an information leak (contamination, an external call, label leakage through the critic) or an evaluation artefact. Stop and audit. |
| C7: imagined evidence is used safely | ICD $\gg \Delta$Acc; ICG exceeds tolerance; confidence rises without accuracy | Hallucination amplification. The calibration machinery of §3.8 has failed and the system should not be deployed regardless of accuracy. |
| C8: recursive improvement is safe | Self-distillation improves training/replay metrics while any immutable gate regresses across two consecutive cycles | The loop is unsound as specified; report the cycle at which it broke. |

**Table 14:** Falsification conditions and what each failure costs.

### 9.2 Results we would treat as suspicious rather than good

Some outcomes are formally positive but violate a prediction of the theory, and we commit in advance to treating them as evidence of a problem.

- **Accuracy still climbing after the critic value has saturated.** Corollary 26 requires the selected set's critic value to be concave in $B$; accuracy rising linearly past that point suggests the critic is correlated with the answer key, or leakage.
- **Gains without band structure.** A large improvement whose ablation profile is uniform means the operator is acting as a generic regulariser or noise source. That may be a real effect, but it is not this paper's effect, and the cheaper way to obtain it should then be found and reported.
- **Improvement that survives randomising the gains.** By Lemma 15 a randomised bounded gate is still an invertible, phase-preserving, mildly distorting filter; if it matches the learned one, the mechanism is stochastic regularisation, not allocation.
- **Improvement concentrated on benchmarks with known contamination risk** and absent on the freshly generated diagnostics.
- **A gate that fires on the text-only control suite yet improves accuracy there.** Imagery cannot help on tasks constructed so that it cannot; such a result means the control suite is not what we think it is.
- **Perfect agreement between $Q_\eta$ and CLIP similarity** alongside a large gain, which would mean the critic learned prettiness and the gain came from elsewhere.

### 9.3 What would count as strong confirmation

Symmetrically, we state in advance the conjunction we would find convincing, so that a partial result cannot be presented as the full one:

1. B7 $>$ B1 and B7 $>$ B2 at matched compute, on at least three benchmark families including one contamination-proof diagnostic, with five seeds and adequate power;
2. band ablations producing *task-specific*, predicted-in-advance degradation patterns;
3. Prediction 1 confirmed with $\lambda_{\mathrm{s}} = 0$;
4. Prediction 2 confirmed, with a monotone accuracy–compute frontier under $\beta$-sweep;
5. Prediction 3 confirmed — the advantage decays with the reader's budget;
6. imagination rate approaching zero on the text-only control suite;
7. ICD $\leq \Delta\mathrm{Acc} + \epsilon_{\mathrm{tol}}$ throughout, with ECE no worse than the backbone;
8. at least two recursive-improvement cycles promoted under unchanged gates.

Items 3, 5 and 7 are the ones we consider load-bearing, because they are the ones that could not be produced by simply spending more compute.

# 10 Risks, Limitations and Mitigations

## 10.1 Epistemic risks

**Imagined evidence is not observed evidence.** This is the central risk and it now has a formal statement: Proposition 6 says an imagined scene contains zero information about the answer *beyond what the state that generated it already held*, and Proposition 8 says a reader that treats it otherwise can be made worse than its own backbone. The mitigations are structural rather than exhortative: provenance typing carried in the state (§4.9); provenance-stratified calibration and a hard confidence cap at the measured channel reliability $\pi_{\text{IMG}}$; and a verifier that is architecturally forbidden from consuming IMG-tagged content as though it were OBS. The residual risk is that provenance tags are lost or laundered through several hops of processing; we therefore propagate tags through the workspace rather than attaching them only at the boundary, and audit tag integrity as a safety gate (Table 4).

**Self-reinforcing error.** Recursive re-perception can make an early wrong scene progressively more convincing — the model imagines, re-perceives, becomes more confident, and conditions the next imagination on the strengthened belief. Mitigations: counterfactual branches with an explicit diversity penalty; per-band memory decay, which bounds how long any single hypothesis can dominate the workspace (Prop. 20 gives the horizon in closed form); periodic workspace resets; and the ICD diagnostic (Cor. 27), which detects the pathology without needing an external judge. We note that *no* mitigation here is theoretically complete: a closed generative loop with a fallible reader can always in principle converge to a confident error, which is why the calibration cap is a cap and not a penalty.

**Anthropomorphic overreading.** Calling this "imagination" invites the inference that the system has something like mental imagery. It does not, and the analogy does no work in the theory — §3 would be unchanged if the workspace were called a cache. We use the cognitive framing for motivation and for the specific, testable prediction that gist should outlive detail; we do not use it as evidence for anything.

## 10.2 Computational risks

**Compute explosion.** A diffusion call costs orders of magnitude more than a decoded token (5), so a system that imagines by default will lose every equal-compute comparison. This is the risk the entire gating apparatus exists to manage, and it is why $\beta$ is a shadow price rather than a tuned constant. Additional mitigations: low imaginal resolution justified by water-filling; latent-only operation skipping the VAE round trip; few-step samplers; caching of imaginal states across sub-questions; and distillation of the imagination step into a direct latent predictor.

**Latency asymmetry.** Even when FLOPs are matched, a diffusion call has worse latency characteristics than token decoding (less amenable to batching across users, larger memory working set). Wall-clock is therefore reported separately and never used to justify a compute claim in either direction.

**Training instability at depth D3.** Feeding workspace summaries into the denoiser via LoRA adapters touches $\psi$ and can degrade the generator. Mitigation: D3 is attempted only after D2 shows

an effect; generator quality is monitored on a fixed probe set each epoch; and the adapters are rolled back on regression.

**The generator fails to produce a coherent scene.** Diffusion models mode-collapse, ignore parts of a prompt, and occasionally emit structureless output — and a workspace that folds such a sample into memory carries the damage forward. Four measures, in the order they fire:

1. **Detect it cheaply.** Two signals cost nothing: the branch-diversity statistic (§5.5) collapses when the generator produces $B$ near-identical samples, and the transport operator's rejection rate spikes when successive samples are not views of a common scene (§4.7). Both are already logged per step.
2. **Refuse the fusion, not the step.** A branch whose critic score falls below the imagination channel's measured reliability $\pi_{\text{IMG}}$ is not folded into memory. The workspace keeps its previous contents rather than accepting a degraded update, which is the conservative direction: a stale memory is recoverable, a poisoned one is not.
3. **Bound the damage structurally.** Even if a bad sample is fused, the per-band decay constants give it a finite lifetime — $\tau_k$ steps, in closed form (Prop. 20) — rather than an indefinite one. This is a concrete benefit of the band-diagonal design that a learned recurrent memory does not offer.
4. **Fall back to abstention.** If the critic scores every branch below threshold, the controller records a failed imagination step, charges its cost, and continues on the language path. The step is wasted, which is correct: the honest outcome of a generator failure is a wasted call, not a confident answer built on noise.

We expect the residual failure to be the *subtle* case — a scene that is coherent, plausible and wrong in the one relation the question turns on. No detector above catches that, and the only defence is the calibration cap, which bounds how much such a scene can move the answer.

**Latency and token cost of the recurrence step.** Re-perception adds a vision-encoder call and $n_{\text{tok}}$ injected tokens per step. Against a text-only reasoner the accounting is: the encoder call is a fixed $\phi_V$ (tens of GFLOPs for a SigLIP-class model, roughly the cost of a few hundred decoded tokens), the injected tokens lengthen the language context by $n_{\text{tok}}$ per imagination step, and the workspace bookkeeping is negligible (§4.11). With $n_{\text{tok}} = 8$ and at most three imagination steps, the context grows by 24 tokens — immaterial against the thousands of tokens an equal-compute reasoning baseline would spend. The dominant term remains the denoiser, by three to four orders of magnitude, which is why the gate is priced against it and nothing else.

### 10.3 Evaluation risks

**Benchmark contamination.** Backbones may have seen benchmark content, and the generator may have seen benchmark imagery — a route to contamination specific to this class of system and absent from text-only work. Mitigations are in §8; the load-bearing one is the programmatically generated diagnostic suite regenerated at evaluation time.

**The capacity confound.** Any added module can improve results by adding parameters. The spectral workspace's 30 parameters make this implausible *for the operator*, but the bridges add $\sim$ 100M and could absorb capacity. Controls: B2 and B4 include the same bridges; the zero-initialised

injection gate means the system starts exactly at its backbone; and B4 in particular holds parameter count roughly fixed while removing the fixed basis.

**Cherry-picked qualitative examples.** Imagined scenes are visually compelling and make persuasive figures. We commit to publishing randomly sampled traces — including failures — at a fixed sampling rate, and to reporting the accepted-branch fraction so that readers can see how many hypotheses were discarded to produce a shown example.

### 10.4 Deployment and misuse risks

**Confident fabrication.** A system that constructs internal scenes and then describes them has a new surface for producing detailed, fluent, wrong descriptions of things that were never observed — particularly dangerous if such output is presented as analysis of a real image the user supplied. Mitigation: outputs derived from IMG-provenance content are marked as such in the user-facing surface, not only in the internal trace, and the confidence cap applies to what is shown.

**Content generation inside the reasoning loop.** The generator is a general image model invoked without an explicit user request for an image. Standard content filtering applies to the imaginal path as it would to an output path; the imaginal path is not a bypass. Because imagined latents may never be rendered, filtering must operate in latent or feature space, which is weaker — we regard this as an open problem and recommend rendering-and-filtering at a sampled rate for monitoring even in latent-only mode.

**Compute cost as an access barrier.** A method whose gains require large inference budgets widens the gap between well- and poorly-resourced users. The accuracy–compute frontier is the honest way to report this, and the distillation path (§4) is the honest way to address it.

### 10.5 Bias in imagined content

A component of this system samples scenes from a generative image prior and then *reasons over them*. That is a different exposure from generating a picture for a user to look at, and it deserves its own treatment.

**The specific mechanism of concern.** Image generators carry well-documented demographic and contextual skews. In an ordinary generation pipeline a skewed sample is visible: the user sees it. In this architecture the imagined scene may never be rendered, is consumed by the model itself, and influences an answer that is presented as reasoning rather than as imagery. A question about "the nurse and the surgeon" that triggers an imagination step could have its answer shaped by whom the generator's prior places in each role — and nothing in the output would reveal that a picture was involved.

**What follows from the theory.** Proposition 6 is unexpectedly useful here. Because the imagined scene carries no information about the answer, *any* systematic influence it has on the answer is attributable to the generator's prior rather than to evidence. That is exactly the definition of a bias channel, and it means the question is not "is the imagery biased" — of course it is, priors are — but "is the reader letting the prior move answers it should not move".

**What we recommend measuring, in the evaluation programme.**

- **Counterfactual prompt invariance.** For question templates that differ only in a demographic term, compare answers with imagination forced on and forced off. A gap that appears only in the imagination-on condition localises the effect to the imaginal path.
- **Imagination-rate parity.** Report the gate's firing rate stratified by the same templates. A controller that imagines more often for some groups than others is spending compute unequally, which is a fairness issue distinct from accuracy.
- **Stratified calibration.** The provenance-stratified reliability of §3.8 should also be reported per group where the evaluation supports it: a uniform $\pi_{\mathrm{IMG}}$ that masks divergent per-group reliabilities is a cap that protects some users and not others.

**What we do not claim.** We have not run these measurements — the diagnostics in §7 are geometric and carry no demographic content — and we do not claim the architecture mitigates generator bias. Provenance typing makes the imaginal path *auditable*, which is a precondition for measuring the problem and not a solution to it. A system deployed on content where these skews matter needs the measurements above run first, and we would regard shipping without them as unjustified.

### 10.6 Limitations we have not solved

- The theory of §3.3 assumes independent Gaussian band coefficients and a locally quadratic task loss. Latent statistics are neither Gaussian nor independent across bands, and the loss is not quadratic. The water-filling result should therefore be read as a design principle with a testable signature, not as an exact optimality claim — which is precisely why Prediction 1 is stated as an empirical test rather than a theorem.
- cVOI is defined relative to a decoder class that we can only approximate in practice (a fixed model at a fixed budget). Comparisons across model families are therefore not directly meaningful.
- The myopic gate is provably optimal only among myopic policies, and only for a reader that attains the maximum over its decoder class. What the controller actually regresses is the gain realised by the deployed reader, which is the estimable and deployable quantity; the two agree as the reader's excess risk goes to zero.
- The band-gain prediction can only be tested on magnitudes when the measured sensitivity range falls inside what the bounded gate can express, (16). Outside that range the prediction degrades to a statement about ordering.
- The value of non-myopia is an empirical question we have not answered.
- Radial masks assume approximate isotropy. Strongly anisotropic tasks (text in images, ruled diagrams) may need oriented filters; a steerable-pyramid variant is a natural extension and is not implemented here.
- Nothing in this work addresses whether the generator's learned scene prior is *correct* about the physical world. A workspace that faithfully simulates a wrong world model will reason confidently and wrongly, and no amount of spectral structure fixes that.

# 11 Beyond Vision: A Generalised Simulation Workspace

The account in §3 is not visual. Nothing in the no-free-evidence property, the cVOI definition, the shadow-price gate, or the rate-allocation argument mentions images. What is visual is only the *instantiation*: images are where mature pretrained generators and encoders exist. This section states the general form and what would be required to instantiate it elsewhere.

## 11.1 The general form

A *simulation modality* $m$ supplies a triple $(D^{(m)}, V^{(m)}, \Phi^{(m)})$: a conditional generator over latents, an encoder mapping latents to features the reader can consume, and a transform basis $\Phi^{(m)}$ in which the workspace's memory is diagonal. The controller becomes a selection problem over modalities:

$$m_t^\star = \operatorname*{arg\,max}_{m \in \mathcal{M} \cup \{\varnothing\}} \Big[ \mathrm{cVOI}_n^{(m)}(s_t) - \beta\, c^{(m)}(s_t) \Big], \tag{43}$$

with $\varnothing$ denoting "simulate nothing and answer." Everything else carries over unchanged: free disposal still gives non-negativity, the shadow price is still shared across modalities (which is what makes them comparable), and each modality contributes its own provenance class to the ledger.

The basis $\Phi^{(m)}$ is where modality-specific structure enters, and the rate-allocation argument tells us how to choose it: pick the basis in which the modality's natural statistics are approximately decorrelated and in which task sensitivity varies strongly across components, since water-filling only buys something when $w_k \sigma_k^2$ is non-uniform. For images that is the radial frequency decomposition. For other modalities:

| Modality | Generator maturity | Natural basis $\Phi^{(m)}$ | Tasks where cVOI should be positive |
|---|---|---|---|
| Vision | mature (latent diffusion) | radial spatial frequency | spatial transformation, occlusion, diagrams |
| Audio | mature (spectrogram/latent audio) | mel/constant-Q bands × time | acoustic scene reasoning, source separation, prosody |
| Video / dynamics | emerging | spatio-temporal frequency | physical prediction, trajectory feasibility |
| Tactile / contact | sparse | contact-map spatial frequency | grasp feasibility, assembly |
| Proprioception | simulator-based | joint-space modes | reachability, motion planning |
| Molecular / structural | emerging | spherical harmonics, graph spectrum | conformational feasibility |

**Table 15:** Instantiating the workspace in other modalities. Vision is first because its generators and encoders are mature, not because the theory privileges it.

## 11.2 Audio as the most tractable second instance

Audio is the natural second modality, for reasons that are structural rather than convenient. The generator–encoder pair is mature; a spectrogram is *already* a band decomposition, so $\Phi^{(m)}$ requires no design and the band gains have an immediate interpretation; the per-band memory constants map onto a real perceptual distinction (envelope versus fine structure); and there exist tasks — "would these two sources be separable?", "does this rhythm resolve?" — where a language model plausibly cannot compute the answer but a simulated-and-re-encoded signal makes it trivial. That is exactly

the cVOI $> 0$ profile the theory predicts, which makes audio a genuine test of generality rather than a demonstration.

### 11.3 A cross-modal prediction

The generalisation makes a prediction that the vision-only system cannot test. If the controller shares $\beta$ across modalities and each modality has its own cost and value, then under a fixed budget the model should exhibit *substitution*: raising the cost of one modality should shift calls to another whose cVOI is comparable, rather than reducing simulation overall. Observing substitution would be evidence that the gate has learned a genuine value comparison; failing to observe it, while each modality individually helps, would indicate the gate is running independent per-modality heuristics.

### 11.4 What we are not proposing

We are not proposing to train a general multisensory simulator. The claim is narrower and, we think, more useful: that *if* a modality has a competent generator and encoder, then the metering apparatus developed here — cVOI against a shared shadow price, a band-diagonal workspace chosen by the modality's statistics, provenance-typed calibration — transfers without modification, and the falsification programme transfers with it. The vision instantiation is the test case; the accounting is the contribution.

## 12 Related Work and the Novelty Boundary

We organise this section around what is *shared* with prior work and what is not, because the main risk to a proposal in this area is that it reads as a recombination. Table 16 is the summary; the prose gives the detail.

### 12.1 Reasoning with generated images

Visual chain-of-thought introduced multimodal infillings between reasoning steps [38]. *Thinking with Generated Images* generates intermediate visual thoughts and self-critiques them [3]. *Visual Thoughts* offers a unifying account of multimodal chain-of-thought and analyses when visual intermediates help [2]. *Latent Sketchpad* maintains an internal visual scratchpad inside an MLLM [53]. *MindJourney* pairs a VLM with a diffusion world model to scale spatial reasoning at test time [48]. More recent work optimises interleaved text–image trajectories as a decision process [20] or performs test-time latent reasoning within generation [31].

**Shared:** that a model can construct internal visual states during reasoning, transform them, re-perceive them, and use them. We claim none of this.

**Not shared:** the treatment of those states as *information-free but computation-bearing* (Prop. 6) and the consequences — the equal-compute baseline as a definitional requirement rather than a courtesy, the budget-decay prediction, and the calibration constraint that follows from the martingale argument. These works generally evaluate against text-only or fixed-schedule baselines without matching compute, and report confidence without stratifying by provenance.

### 12.2 Latent and continuous deliberation

Coconut trains models to reason in a continuous latent space rather than in tokens [17]; recurrent-depth architectures scale test-time compute by iterating in latent space [9]. Both address the

same format problem from the opposite direction: they remove the linguistic bottleneck without introducing a perceptual one.

**Shared:** the diagnosis that token sequences are a costly medium for some intermediate computations. **Not shared:** a structured, interpretable basis for the latent workspace with readable per-component time constants, and a metering rule for when to enter it. Our workspace is deliberately *less* expressive than a learned recurrent latent — a bank of $K$ first-order filters — because expressiveness is what makes such a component uninterpretable and un-ablatable.

### 12.3 World models and imagination in RL

Learning in imagination is standard in model-based RL: world models compress environment dynamics and policies are trained on imagined rollouts [15, 16]. Active-inference accounts treat action selection as expected-free-energy minimisation, with an explicit epistemic term [7], and the psychology and neuroscience of information seeking supply the empirical backdrop [10].

**Shared:** imagination as rollout under a learned generative model, and the idea that information gain has a value. **Not shared:** the observation that in the *closed* setting — where the simulator conditions only on the agent's own state and no environment interaction follows — the epistemic term is identically zero, so the value must be re-derived as computational rather than informational. In model-based RL the imagined rollout is eventually validated by acting in the world; in single-turn reasoning it never is. That difference is what makes Proposition 6 bite here and not there.

### 12.4 Bounded rationality and metareasoning

Bounded optimality [39], resource-rational analysis [26], information-theoretic bounded rationality [34], the expected value of control [40], and classical value of information [18] together supply the framework we adopt.

**Shared:** everything conceptual. **Not shared:** the application. To our knowledge this framework has not been used to price a *generative simulation step inside a language model's inference*, nor has cVOI been given an estimable target (§5.4) and used as a regression objective for a deployed gate whose multiplier is solved by dual ascent. Adaptive-compute work in LLMs [43, 14] scales deliberation but does not price heterogeneous deliberation *modalities* against a common shadow price.

### 12.5 Spectral and multiscale representations

Scale-space theory motivates smooth multi-scale front ends [27]; the importance of phase in image structure is classical [33]; Fourier-domain operators are well established in neural architectures [25]; and diagonal state-space models exploit structured linear recurrences for long-range memory [12, 11]. Rate–distortion theory and reverse water-filling are textbook [4].

**Shared:** the mathematics, entirely. We claim no new spectral machinery. **Not shared:** the identification of a reasoning workspace as a rate-limited channel, the task-weighted water-filling solution, and the resulting prediction $G_k \propto \sqrt{w_k}$ tying *learned* gains to *measured* task sensitivity. Fourier neural operators learn spectral weights to solve PDEs; diagonal SSMs learn spectral dynamics for sequence memory; neither asks whether the learned spectrum matches an independently measurable sensitivity profile, because for them there is no such profile to match.

### 12.6 Multimodal architecture and generation

Latent diffusion [37] and its successors [36, 6] provide the generator; Flamingo, BLIP-2 and LLaVA established lightweight bridging of frozen backbones [1, 24, 28]; DreamLLM and Janus unify comprehension and generation [5, 47]; Mural transfers LLM knowledge into generation [21]; SigLIP-style encoders supply re-perception [52]; LoRA supplies the adapter mechanics [19]. We use all of this as infrastructure and claim none of it.

### 12.7 Calibration and self-training

Modern networks are systematically over-confident and post-hoc scaling largely corrects it [13]; language models have some access to their own correctness [22]; training on recursively generated data degrades tail coverage [42].

**Shared:** the techniques. **Not shared:** provenance-stratified calibration with a hard cap at the imagination channel's measured reliability, and the ICD diagnostic derived from Proposition 6, which detects hallucination amplification without an external judge.

### 12.8 Summary

| Ingredient | Prior art | This paper's position |
|---|---|---|
| Generated images as reasoning steps | [38, 3, 2, 53, 48] | Adopted, not claimed |
| Latent-space deliberation | [17, 9] | Adopted; we constrain the basis |
| Imagination for control | [15, 16, 7] | Adopted; we show the epistemic term vanishes in the closed setting |
| Bounded rationality / VOI | [18, 39, 26, 34, 40] | Adopted as framework; **new application** to pricing generative simulation in LLM inference |
| Spectral operators, SSMs, water-filling | [27, 25, 12, 11, 4] | Adopted as mathematics; **new use** as a task-weighted rate allocation for a reasoning workspace |
| Calibration | [13, 22] | Adopted; **new** provenance stratification and reliability cap |
| **Claimed new:** Prop. 6 and its consequences; cVOI with an estimable target and a dual-ascent shadow price; task-weighted water-filling with the $G_k \propto \sqrt{w_k}$ prediction; band-diagonal workspace with readable $\tau_k$; Cor. 27 as a judge-free amplification detector; and the falsification programme of §9. | | |

**Table 16:** Novelty boundary. The left two columns are what we build on; the right column states exactly what is being asserted as new, so that a reviewer can check each claim independently.

## 13 Conclusion

The question this paper set out to answer is not whether a language model can construct internal images while reasoning. Several systems already do that, and the demonstration is no longer informative. The question is what such a step is *worth*, in units that can be compared against the alternative use of the same compute, and what structure the simulation should have so that the compute buys as much usable evidence as it can.

Answering the first question required accepting an uncomfortable fact. An imagined scene is generated by the reasoner from its own state, so — conditional on that state — it carries no information about the answer at all (Proposition 6). Everything the subarea calls "visual evidence" is, formally, already-held information in a different format. That is not a reason to abandon the idea; it is a reason to change what we measure. If the benefit is computational rather than informational, then the right control is an equal-compute reader, the benefit must shrink as that reader's budget grows, and a confidence increase that is not paid for by an accuracy increase is a bug with a name and a detector. Each of those became an experiment (§8.3.3, Cor. 27) rather than a caveat.

Answering the second question turned out to require less invention than we expected. Once the imaginal pathway is modelled as a rate-limited channel, task-weighted reverse water-filling gives the optimal allocation of precision across frequency bands, a strictly positive real-valued radial gate implements that allocation over the range its bounded parameterisation can express, and the optimal gains are proportional to the square root of independently measurable task sensitivity (Corollary 12). The operator that falls out is phase-exact, invertible and well-conditioned; the workspace that carries it across steps is a bank of first-order filters with per-band time constants one can read off and check. The spectral machinery, which the earlier version of this work asserted as an inductive bias, is now a consequence — and, more usefully, a source of predictions that could turn out to be false.

We think this is the right shape for a proposal in a crowded area. The contribution is not a longer list of components. It is an accounting under which some of the field's standard claims are literally false, some of its standard comparisons are uninformative, and a small number of cheap new measurements — the gain–sensitivity correlation, the cVOI reliability diagram, the budget-decay curve, the confidence-drift gap — would tell us quickly whether the mechanism is real.

The pilots in §7 are the first instalment of that programme, and they are worth weighing carefully in both directions. On the supportive side: on the texture-dominated family, gains trained by task loss alone tracked measured sensitivity, with a slope that rose as the operator's range widened and reached 35–61% of what that range could express; a gate regressing realised gain learned the value structure from measurement rather than from our intentions — it assigned near-zero value to a family we *expected* imagery to help, because in that pilot it did not; alignment before fusion cut registered reconstruction error more than sixfold; and the advantage of imagery decayed as the reader grew, which is what a computational benefit has to do.

On the unsupportive side, and more consequentially: the spectral workspace did not separate from generic imagery at matched compute; the gain–sensitivity correspondence failed outright on the layout-dominated family; and the gate did not clear our own abstention criterion, because the pilot thresholded predicted gain at zero rather than against a shadow price. Three of eight testable statements failed. The first is the outcome §9 names as refuting the central mechanism claim, and we have put it in the abstract, the main text and here rather than in a footnote, because a falsification criterion quietly relaxed on first contact was never a criterion.

There is a fourth thing worth recording. Our first analysis of the gain–sensitivity pilot was buggy in a way that flattered the hypothesis — floored sensitivity values entering a log-log fit, a single-seed ceiling quoted beside a five-seed slope — and reported slopes roughly three times too large. We caught it in an adversarial re-check, not in the original analysis. The lesson we draw is that a falsification programme needs an auditing step aimed at its own favourable results, not only at its unfavourable ones, and Appendix F records every such error we made.

VIVID is a methods and theory proposal with a working reference implementation of the control loop, the spectral operator, the band-wise workspace, latent transport, the budgeted gate with dual ascent, the band allocator, the calibrated provenance-typed critic, the diagnostic generators and the pilot harness. It reports no benchmark numbers, and Table 13 is deliberately empty. The programme in §8 is designed so that the most likely outcome — that equal-compute generated imagery matches

the spectral system, which is what our own pilot found — is recognised quickly and reported, rather than absorbed into a more flattering framing. We would rather publish that null than a number we could not explain.

## References


[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, et al. Flamingo: a visual language model for few-shot learning. *arXiv:2204.14198*, 2022.

[2] Zihui Cheng et al. Visual thoughts: A unified perspective of understanding multimodal chain-of-thought. In *NeurIPS*, 2025.

[3] Ethan Chern et al. Thinking with generated images. *arXiv:2505.22525*, 2025.

[4] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006.

[5] Runpei Dong et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv:2309.11499*, 2023.

[6] Patrick Esser, Sumith Kulal, Andreas Blattmann, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

[7] Karl Friston, Francesco Rigoli, Dimitri Ognibene, Christoph Mathys, Thomas Fitzgerald, and Giovanni Pezzulo. Active inference and epistemic value. *Cognitive Neuroscience*, 6(4):187–214, 2015.

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

[9] Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *arXiv:2502.05171*, 2025.

[10] Jacqueline Gottlieb, Pierre-Yves Oudeyer, Manuel Lopes, and Adrien Baranes. Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in Cognitive Sciences*, 17(11):585–593, 2013.

[11] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv:2312.00752*, 2023.

[12] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *ICLR*, 2022.

[13] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

[14] Daya Guo, Dejian Yang, Haowei Zhang, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

[15] David Ha and Jürgen Schmidhuber. World models. *arXiv:1803.10122*, 2018.

[16] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059), 2025.

[17] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv:2412.06769*, 2024.

[18] Ronald A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966.

[19] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

[20] Zican Hu et al. Bridging interleaved multi-modal reasoning as a unified decision process. *arXiv:2607.03748*, 2026.

[21] Achin Jain et al. Mural: Transferring llm knowledge to image generation via mixture-of-transformers. *arXiv:2606.29013*, 2026.

[22] Saurav Kadavath, Tom Conerly, Amanda Askell, et al. Language models (mostly) know what they know. *arXiv:2207.05221*, 2022.

[23] Stephen M. Kosslyn, Giorgio Ganis, and William L. Thompson. Neural foundations of imagery. *Nature Reviews Neuroscience*, 2(9):635–642, 2001.

[24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

[25] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *ICLR*, 2021.

[26] Falk Lieder and Thomas L. Griffiths. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43:e1, 2020.

[27] Tony Lindeberg. *Scale-Space Theory in Computer Vision*. Springer, 1994.

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv:2304.08485*, 2023.

[29] Pan Lu et al. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 2022.

[30] Pan Lu et al. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv:2310.02255*, 2023.

[31] Yapeng Mi et al. Milr: Improving multimodal image generation via test-time latent reasoning. In *ICLR*, 2026.

[32] George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14(1):265–294, 1978.

[33] Alan V. Oppenheim and Jae S. Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981.

[34] Pedro A. Ortega and Daniel A. Braun. Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A*, 469(2153):20120683, 2013.

[35] Joel Pearson. The human imagination: the cognitive neuroscience of visual mental imagery. *Nature Reviews Neuroscience*, 20(10):624–634, 2019.

[36] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024.

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[38] Daniel Rose et al. Visual chain of thought: Bridging logical gaps with multimodal infillings. *arXiv:2305.02317*, 2023.

[39] Stuart J. Russell and Devika Subramanian. Provably bounded-optimal agents. *Journal of Artificial Intelligence Research*, 2:575–609, 1995.

[40] Amitai Shenhav, Matthew M. Botvinick, and Jonathan D. Cohen. The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron*, 79(2):217–240, 2013.

[41] Roger N. Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703, 1971.

[42] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024.

[43] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv:2408.03314*, 2024.

[44] Maxim Sviridenko. A note on maximizing a submodular set function subject to a knapsack constraint. *Operations Research Letters*, 32(1):41–43, 2004.

[45] Tristan Thrush et al. Winoground: Probing vision and language models for visio-linguistic compositionality. *arXiv:2204.03162*, 2022.

[46] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

[47] Chengyue Wu et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv:2410.13848*, 2024.

[48] Yuncong Yang et al. Mindjourney: Test-time scaling with world models for spatial reasoning. *arXiv:2507.12508*, 2025.

[49] Xiang Yue et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024.

[50] Xiang Yue et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv:2409.02813*, 2024.

[51] Adam Zeman, Michaela Dewar, and Sergio Della Sala. Lives without imagery: Congenital aphantasia. *Cortex*, 73:378–380, 2015.

[52] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023.

[53] Huanyu Zhang et al. Latent sketchpad: Sketching visual thoughts to elicit multimodal reasoning in mllms. *arXiv:2510.24514*, 2025.

# A Proofs

Throughout, $\mathcal{F}$ is the unitary 2-D DFT, so $\mathcal{F}^* = \mathcal{F}^{-1}$ and $\|z\|_2 = \|\mathcal{F}z\|_2$ (Parseval). Operations on gates and spectra are elementwise. Recall $M_k \geq 0$, $\sum_{k=1}^{K} M_k(\xi) = 1$ for all $\xi$, and $a_{ck} = \tanh(\alpha_{ck}) \in (-1, 1)$ with $\bar{a} \coloneqq \max_{c,k} |a_{ck}| < 1$.

### Proof of Proposition 6 (no free evidence)

By Assumption 5, $Z_t = \Gamma(S_t, \varepsilon)$ where $\Gamma$ is the (fixed) sampling map of the generator composed with the bridge, and $\varepsilon \perp (Y, S_t)$ is the sampling noise. Hence for any measurable set $A$, $\mathbb{P}(Z_t \in A \mid S_t, Y) = \mathbb{P}(\Gamma(S_t, \varepsilon) \in A \mid S_t, Y) = \mathbb{P}(\Gamma(S_t, \varepsilon) \in A \mid S_t)$, using $\varepsilon \perp Y \mid S_t$. Therefore $Z_t \perp Y \mid S_t$, i.e. $Y \to S_t \to Z_t$ is a Markov chain, and $I(Y; Z_t \mid S_t) = 0$ by definition of conditional mutual information. Bayes' rule then gives $p(Y \mid s_t, z_t) = p(Y \mid s_t)$ almost surely.

For any measurable $g$, $g(Z_t)$ is $\sigma(S_t, \varepsilon)$-measurable, so the same argument applies; the data-processing inequality gives $I(Y; g(Z_t) \mid S_t) \leq I(Y; Z_t \mid S_t) = 0$. Taking $g = S_\omega$, $g = \mathcal{D}$ (decode) and $g = V_\nu \circ S_\omega$ covers the transformed latent, the decoded image, and the re-encoded features. □

**Remark 30.** The proposition is stated conditional on $S_t$, which includes the workspace $m_{t-1}$. It does *not* say that $I(Y; Z_t) = 0$ unconditionally — an imagined latent is of course correlated with the answer, because it is correlated with the state. The claim is that it adds nothing *given* what the reasoner already has, which is the quantity that matters for deciding whether to spend compute.

### Proof of Proposition 7 (non-negativity under free disposal)

Fix $s_t$ and let $a^\dagger \in \arg\max_{a \in \mathcal{A}_n} \mathbb{E}[R(a(h_t, \varnothing), Y) \mid s_t]$. By free disposal there is $a' \in \mathcal{A}_n$ with $a'(h_t, v_t) = a^\dagger(h_t, \varnothing)$ for all $v_t$. Hence

$$\begin{aligned} \max_{a \in \mathcal{A}_n} \mathbb{E}[R(a(h_t, v_t), Y) \mid s_t] &\geq \mathbb{E}[R(a'(h_t, v_t), Y) \mid s_t] \\ &= \mathbb{E}[R(a^\dagger(h_t, \varnothing), Y) \mid s_t] \\ &= \max_{a \in \mathcal{A}_n} \mathbb{E}[R(a(h_t, \varnothing), Y) \mid s_t]. \end{aligned}$$

Subtracting gives $\mathrm{cVOI}_n(s_t) \geq 0$. □

### Proof of Proposition 8 (realised value can be negative)

We give an explicit construction. Let $\mathcal{Y} = \{0, 1\}$ with $Y$ uniform, and let the state $h_t$ determine $Y$ exactly, so that $\mathbb{E}[R(\hat{a}(h_t, \varnothing), Y)] = 1$ for the reader $\hat{a}(h, \varnothing)$ that reads $h$ correctly. Let the generator produce a binary visual cue $v \in \{0, 1\}$ with $\mathbb{P}(v \neq Y) = \gamma \in (0, 1)$ — a biased generator, e.g. one whose scene prior favours the more typical configuration.

*Reader 1 (full deference).* $\hat{a}$ answers $v$ whenever a cue is present. This is a plausible learned behaviour if visual features were more reliable than language features during training, and it is the canonical form of over-trusting self-generated evidence. Then $\mathbb{E}[R(\hat{a}(h_t, v), Y)] = 1 - \gamma$, so

$$\Delta_t(\hat{a}) = (1 - \gamma) - 1 = -\gamma \; < \; 0.$$

*Reader 2 (partial deference).* $\hat{a}_p$ follows $h_t$ when the cue agrees with it and switches to the cue with probability $p \in [0, 1]$ when they disagree. Since $h_t$ is always correct, disagreement occurs exactly when $v \neq Y$, with probability $\gamma$, and the reader is wrong exactly when it switches:

$$\mathbb{E}[R(\hat{a}_p(h_t, v), Y)] = 1 - p\gamma, \qquad \Delta_t(\hat{a}_p) = -p\gamma,$$

which recovers Reader 1 at $p = 1$ and is strictly negative for any $p > 0$. The construction requires no condition on $\gamma$ beyond $\gamma > 0$; the damage scales with how much the reader defers and how often the generator is wrong.

In both cases $\Delta_t(\hat{a}) < 0$ while, by Proposition 7, $\mathrm{cVOI}_n(s_t) \geq 0$. For the gap: writing $\mathrm{ER}_v$ and $\mathrm{ER}_\varnothing$ for the reader's excess risks on the two branches, $\mathrm{cVOI}_n(s_t) - \Delta_t(\hat{a}) = \mathrm{ER}_v - \mathrm{ER}_\varnothing$. In this construction $\hat{a}$ is optimal on the no-imagination branch, so $\mathrm{ER}_\varnothing = 0$ and the gap is exactly the excess risk from treating $v$ as an observation rather than as a sample from the reader's own prior. In general it is the difference of the two excess risks, and the proposition claims no more. □

### Proof of Proposition 10 (shadow price and greedy gate)

Consider the space of randomised policies, which is convex, and assume $R$ bounded so the objective and constraint are bounded linear functionals of the occupancy measure. The constrained problem (6) is then a linear program over occupancy measures with a single inequality constraint; Slater's condition holds whenever a strictly feasible policy exists (e.g. the never-imagine policy, which has cost strictly below $\mathcal{B}$ for $\mathcal{B} > 0$). Strong duality for such constrained MDPs is standard, giving (i).

For (ii), the value function of the constrained problem, $\mathcal{V}(\mathcal{B})$, is concave in $\mathcal{B}$, and by the envelope theorem for LPs the optimal multiplier satisfies $\beta^\star \in \partial\mathcal{V}(\mathcal{B})$, i.e. $\beta^\star$ is the (super)gradient of achievable expected reward with respect to the budget — the shadow price.

For (iii), fix $\beta = \beta^\star$ and consider the unconstrained Lagrangian (7). Restricting to policies that condition on $s_t$ and evaluate a single step's contribution (the myopic class), the contribution of taking IMAGINE at $s_t$ relative to not taking it is exactly $\mathrm{cVOI}_n(s_t) - \beta^\star c_{\mathrm{img}}$ by Definition 2 and the additivity of the cost term. Maximising a sum of independent per-state contributions is achieved by maximising each, giving the rule IMAGINE $\iff \mathrm{cVOI}_n(s_t) > \beta^\star c_{\mathrm{img}}$. □

**Remark 31** (Two caveats on step (iii))**.** First, the conditional-independence assumption is what makes the per-state contributions separable; with cross-step interactions the myopic rule is only optimal within its class, which is what the proposition claims. Second — and more consequentially — the Lagrangian's reward is realised by the *deployed* reader, whereas $\mathrm{cVOI}_n$ is a maximum over $\mathcal{A}_n$. The identification in step (iii) is therefore exact only when the deployed reader attains that maximum. For a general reader the same argument goes through with $\mathrm{cVOI}_n(s_t)$ replaced by $\Delta_t(\hat{a})$ from Proposition 8, and the resulting rule — imagine iff $\Delta_t(\hat{a}) > \beta^\star c_{\mathrm{img}}$ — is what the implementation actually uses (§5.4). The two coincide as the reader's excess risk goes to zero.

### Proof of Proposition 11 (task-weighted reverse water-filling)

For a Gaussian source of variance $\sigma_k^2$ under squared error, the rate–distortion function is $R_k(D_k) = \frac{1}{2}\log^+(\sigma_k^2/D_k)$, achieved with $D_k \le \sigma_k^2$ [4, Ch. 10]. Minimise $\sum_k w_k D_k$ subject to $\sum_k R_k(D_k) \le R$ and $0 < D_k \le \sigma_k^2$. On the set of *active* bands ($D_k < \sigma_k^2$) form

$$J = \sum_k w_k D_k + \mu\Big(\sum_k \tfrac{1}{2}\log\frac{\sigma_k^2}{D_k} - R\Big), \qquad \mu > 0.$$

Stationarity: $\partial J/\partial D_k = w_k - \mu/(2D_k) = 0 \Rightarrow D_k = \mu/(2w_k)$. Writing $\theta := \mu/2$ and imposing the box constraint $D_k \le \sigma_k^2$ gives $D_k^\star = \min(\theta/w_k,\ \sigma_k^2)$; the corresponding rate is $R_k^\star = \frac{1}{2}\log^+(\sigma_k^2 w_k/\theta)$. The objective is convex in $D_k$ (linear) with a convex feasible set defined by the convex constraint $\sum_k -\frac{1}{2}\log D_k \le R + \text{const}$, so the KKT point is the global minimum. $\theta$ is determined by $\sum_k \frac{1}{2}\log^+(\sigma_k^2 w_k/\theta) = R$, whose left side is continuous and strictly decreasing in $\theta$ on the relevant range, hence $\theta$ is unique. Bands with $w_k\sigma_k^2 \le \theta$ satisfy $D_k^\star = \sigma_k^2$ and receive zero rate. □

### Proof of Corollary 12 ($G_k \propto \sqrt{w_k}$)

Let band $k$ carry signal with gain $G_k$ applied before an additive noise of band-independent power $\epsilon^2$, and let the reader normalise by $G_k$ (or, equivalently, let the downstream layer be scale-covariant per band). The effective distortion referred to the input is $D_k = \epsilon^2/G_k^2$. Setting this equal to the optimal $D_k^\star = \theta/w_k$ from (14) for an active band gives $G_k^2 = \epsilon^2 w_k/\theta$, hence $G_k = (\epsilon/\sqrt{\theta})\sqrt{w_k}$. Since $\epsilon/\sqrt{\theta}$ does not depend on $k$, $G_k \propto \sqrt{w_k}$, with the constant of proportionality free. For inactive bands ($w_k\sigma_k^2 \le \theta$) the allocation assigns no rate and the corollary makes no prediction, which is why the empirical test in §8.3.1 restricts to bands with non-negligible measured sensitivity. □

**Remark 32** (What this argument does and does not establish)**.** The gain model carries no rate constraint of its own, so the corollary is an *identification*: it says which gate realises the water-filling allocation under a fixed post-gate noise floor, with the operating point supplied by the water-filling problem rather than determined here. And the proportionality can only be realised within the operator's bounded range (16); the free constant cannot rescue a sensitivity spectrum whose dynamic range exceeds $((1+\rho)/(1-\rho))^2$, because $G_\rho$ is pinned at unity plus a bounded perturbation and so admits no free global scale of its own.

### Proof of Lemma 15 (residual form)

By linearity of $\mathcal{F}$,

$$S_\omega(z) = z + \rho\big[\mathcal{F}^{-1}(G \odot \mathcal{F}z) - z\big] = \mathcal{F}^{-1}\big[(1-\rho)\mathcal{F}z + \rho\, G \odot \mathcal{F}z\big] = \mathcal{F}^{-1}\big[G_\rho \odot \mathcal{F}z\big],$$

with $G_\rho = (1-\rho) + \rho G = 1 + \rho\sum_k a_{\cdot k} M_k$, using $G = 1 + \sum_k a_{\cdot k} M_k$. Since $M_k \ge 0$ and $\sum_k M_k = 1$, the quantity $\sum_k a_{ck} M_k(\xi)$ is a convex combination of $\{a_{ck}\}_k$ and therefore lies in $[\min_k a_{ck}, \max_k a_{ck}] \subseteq [-\bar{a}, \bar{a}]$. Hence $G_{\rho,c}(\xi) \in [1-\rho\bar{a}, 1+\rho\bar{a}]$ — the endpoints are attained wherever a single mask dominates and $|a_{ck}| = \bar{a}$, so the interval is closed — and since $\bar{a} < 1$ strictly and $\rho \le 1$ we have $\rho\bar{a} < 1$ and therefore $G_\rho > 0$ everywhere. $G_\rho$ is real because $a_{ck}$ and $M_k$ are. □

### Proof of Proposition 16 (phase preservation)

By Lemma 15, $(\mathcal{F}S_\omega(z))_c(\xi) = G_{\rho,c}(\xi)\,(\mathcal{F}z)_c(\xi)$ with $G_{\rho,c}(\xi)$ real and strictly positive. For $u \in \mathbb{C}$ and $g > 0$, $\arg(gu) = \arg(u)$ and $|gu| = g|u|$. Applying this pointwise gives both claims. □

### Proof of Proposition 17 (equivariance)

*Translation.* Let $T_\delta$ be cyclic translation by $\delta$. Then $\mathcal{F}(T_\delta z)(\xi) = e^{-2\pi i\langle\delta,\xi\rangle}\mathcal{F}z(\xi)$. Since $S_\omega$ acts by elementwise multiplication with $G_\rho(\xi)$, which does not depend on the phase factor, $\mathcal{F}(S_\omega(T_\delta z)) = G_\rho \cdot e^{-2\pi i\langle\delta,\cdot\rangle}\mathcal{F}z = e^{-2\pi i\langle\delta,\cdot\rangle}\mathcal{F}(S_\omega z) = \mathcal{F}(T_\delta S_\omega z)$, and injectivity of $\mathcal{F}$ gives $S_\omega T_\delta = T_\delta S_\omega$.

*Rotation.* In the continuous setting, if $R_\vartheta$ is a rotation then $\mathcal{F}(z \circ R_\vartheta) = (\mathcal{F}z) \circ R_\vartheta$, and a radial multiplier $G_\rho(r(\xi))$ satisfies $G_\rho \circ R_\vartheta = G_\rho$; commutation follows. On a discrete $H \times W$ grid the map $r(\xi)$ is only approximately rotation-invariant: for the mask family $M_k(\xi) = \varphi((r(\xi) - c_k)/\varsigma)$ with $\varphi$ Lipschitz-$L$, a rotation moves $r$ by at most the radial quantisation $\Delta r$, so $\|G_\rho \circ R_\vartheta - G_\rho\|_\infty \le \rho\bar{a}KL\Delta r/\varsigma = O(\Delta r)$, and the commutator is bounded in operator norm by the same quantity. For $H = W$ and $\vartheta \in \{0, \pi/2, \pi, 3\pi/2\}$ the rotation is a permutation of grid points that preserves $\|\xi\|$ exactly, so commutation is exact. □

### Proof of Proposition 18 (invertibility and bi-Lipschitz bounds)

By Lemma 15, $S_\omega = \mathcal{F}^{-1}\mathrm{diag}(G_\rho)\mathcal{F}$ with $G_\rho \in [1 - \rho\bar{a}, 1 + \rho\bar{a}]$ and $1 - \rho\bar{a} > 0$. Hence $\mathrm{diag}(G_\rho)$ is invertible with inverse $\mathrm{diag}(G_\rho^{-1})$, and $S_\omega^{-1} = \mathcal{F}^{-1}\mathrm{diag}(G_\rho^{-1})\mathcal{F}$.

By Parseval, $\|S_\omega(z)\|_2 = \|G_\rho \odot \mathcal{F}z\|_2$. Since $(1 - \rho\bar{a})|\hat{z}| \le |G_\rho\hat{z}| \le (1 + \rho\bar{a})|\hat{z}|$ pointwise, summing squares and using $\|\mathcal{F}z\|_2 = \|z\|_2$ gives the two-sided bound. For the distortion bound, $\|S_\omega(z) - z\|_2 = \|(G_\rho - 1) \odot \mathcal{F}z\|_2 \le \|G_\rho - 1\|_\infty\|z\|_2 \le \rho\bar{a}\|z\|_2$. Finally $\kappa(S_\omega) = \|S_\omega\|\,\|S_\omega^{-1}\| \le (1 + \rho\bar{a}) \cdot (1 - \rho\bar{a})^{-1}$. □

### Proof of Proposition 20 (band-diagonal LTI dynamics)

Write $\hat{m}_t \coloneqq \mathcal{F}m_t$ and $\hat{\tilde{z}}_t \coloneqq \mathcal{F}(T_t S_\omega(z_t))$ for the gated-and-aligned input. The update (22) reads $\hat{m}_t = \Lambda \odot \hat{m}_{t-1} + (1 - \Lambda) \odot \hat{\tilde{z}}_t$ with $\Lambda(\xi) = \sum_k \lambda_k M_k(\xi)$ a fixed real multiplier. This is a first-order linear recursion, elementwise in $\xi$; unrolling gives

$$\hat{m}_t = \Lambda^t \odot \hat{m}_0 + \sum_{j=0}^{t-1} \Lambda^j \odot (1 - \Lambda) \odot \hat{\tilde{z}}_{t-j},$$

which is the claimed solution. Because $\Lambda$ is a convex combination of $\{\lambda_k\}$ it lies in $[\min_k \lambda_k, \max_k \lambda_k] \subset [0, 1)$; hence $\Lambda^t \to 0$ and the impulse response $\{(1 - \Lambda)\Lambda^j\}_{j\ge0}$ is absolutely summable with $\sum_j (1 - \Lambda)\Lambda^j = 1$, giving BIBO stability and, from $\tilde{z}$ to $m$, unit steady-state gain at every frequency. (Had the gate been placed *inside* the update, the steady-state gain would be $G_\rho$ rather than 1; Definition 19 applies it before, which is also what the implementation does.) The decay $\Lambda^j = e^{j \ln \Lambda}$ has time constant $\tau = -1/\ln\Lambda$, and the mean lag is $\sum_{j\ge0} j(1 - \Lambda)\Lambda^j = \Lambda/(1 - \Lambda)$. Diagonality in $\xi$ means the system decouples into independent scalar filters — one per frequency, with the profile parameterised by the $K$ constants; see Remark 21 on why this is not literally $K$ filters unless the masks are indicators. □

### Proof of Proposition 23 (unregistered fusion attenuates structure)

*(i)* The Fourier shift theorem gives $\mathcal{F}(T_\delta z)(\xi) = e^{-2\pi i\langle\delta,\xi\rangle}\,\mathcal{F}z(\xi)$. Hence for i.i.d. offsets $\delta_j$ independent of $z$,

$$\mathbb{E}\Big[\mathcal{F}\big(\textstyle\sum_j c_j T_{\delta_j} z\big)(\xi)\Big] = \sum_j c_j\,\mathbb{E}\big[e^{-2\pi i\langle\delta_j,\xi\rangle}\big]\,\mathcal{F}z(\xi) = \Big(\textstyle\sum_j c_j\Big)\varphi(\xi)\,\mathcal{F}z(\xi).$$

The factor $\sum_j c_j$ is kept explicit because the workspace recursion does *not* satisfy $\sum_j c_j = 1$ at finite $t$ when $m_0 = 0$: unrolling (22) gives weights summing to $1 - \Lambda(\xi)^t$.

*(ii)* $\varphi$ is the characteristic function of the offset law evaluated at $-2\pi\xi$, so $|\varphi(\xi)| \le 1$ with equality at $\xi = 0$. Expanding $e^{-2\pi i\langle\delta,\xi\rangle}$ to second order about $\xi = 0$ for a zero-mean offset with covariance $\Sigma$ gives $\varphi(\xi) = 1 - 2\pi^2\,\xi^\top\Sigma\,\xi + o(\|\xi\|^2)$, hence $1 - |\varphi(\xi)| \approx 2\pi^2\,\xi^\top\Sigma\,\xi$.

Strictness away from the origin needs more than non-degeneracy. On the cyclic group $\mathbb{Z}_H \times \mathbb{Z}_W$, $|\varphi(\xi)| = 1$ iff $e^{-2\pi i\langle\delta,\xi\rangle}$ is a.s. constant, i.e. the support of $\delta$ lies in a coset of the subgroup annihilated by $\xi$. So $|\varphi(\xi)| < 1$ for all $\xi \neq 0$ iff the support generates the full group (aperiodicity). A counterexample: with $\delta_1$ uniform on $\{0, H/2\}$ and $\delta_2 \equiv 0$, $\varphi(\xi) = \frac{1}{2}(1 + e^{-i\pi\xi_1})$ so that $|\varphi(\xi)| = |\cos(\pi\xi_1/2)|$, which is 1 at every even $\xi_1$ (and 0 at every odd one), so a whole sublattice of frequencies passes unattenuated.

(The expansion above is in cycles per pixel; the counterexample and the Dirichlet kernel below index $\xi$ by integer frequency, hence the explicit $/H$. The two differ by a factor of $H$ in $\Sigma$ and nothing else.)

Global monotonicity in $\|\xi\|$ holds for Gaussian offsets, where $|\varphi(\xi)| = \exp(-2\pi^2\sigma^2\|\xi\|^2)$, and more generally for symmetric stable laws. It fails for bounded uniform offsets: for $\delta_1$ uniform on $\{-J, \dots, J\}$, $\varphi$ is a Dirichlet kernel $\sin\big(\pi\xi_1(2J{+}1)/H\big)\big/\big((2J{+}1)\sin(\pi\xi_1/H)\big)$, which has zeros and side-lobes. Only the near-origin statement is claimed in general.

*(iii)* Write $u_j = e^{-2\pi i\langle\delta_j,\xi\rangle}$, so $\mathbb{E}u_j = \varphi$ and $\mathbb{E}|u_j|^2 = 1$. With $\sum_j c_j = 1$ and independence,

$$\mathbb{E}\Big|\sum_j c_j u_j\Big|^2 = \sum_{j\neq l} c_j c_l\,|\varphi|^2 + \sum_j c_j^2 = |\varphi|^2\Big(1 - \sum_j c_j^2\Big) + \sum_j c_j^2 = |\varphi|^2 + \big(1 - |\varphi|^2\big)\sum_j c_j^2.$$

Since $\sum_j c_j^2 \in (0, 1]$, this lies in $(|\varphi|^2, 1]$, equalling 1 for a point mass on one view and decreasing toward $|\varphi|^2$ as the weights spread. □

**Remark 33** (What compounds, and under which assumption)**.** Under the proposition's own model — all views translations of a common $z$ — fusion contributes a *single* factor of $\varphi$ regardless of depth, as (i) shows. Compounding across steps requires an accumulating-offset model: at step $t$ the incoming latent is offset relative to a memory that already carries the previous offsets, so the effective displacement is a random walk with covariance $t\Sigma$ and the near-origin attenuation grows linearly in $t$. That is the realistic regime for a long-running workspace, and Pilot E measures it directly rather than relying on the derivation.

**Remark 34** (Band selectivity)**.** By (ii), the attenuation at band $k$ scales with $\bar{r}_k^2$ for the band's mean radius $\bar{r}_k$, so high bands are suppressed first. The prediction is therefore not merely "fusion blurs" but "fusion damages texture-sensitive tasks before layout-sensitive ones", which is separately testable.

## Proof of Proposition 25 (submodularity of expected max)

Fix a realisation $q \in \mathbb{R}^{\mathcal{U}}_{\ge 0}$ and let $g(\mathcal{S}) = \max_{b\in\mathcal{S}} q_b$ with $g(\emptyset) = 0$. Non-negativity of $q$ makes $g$ non-negative and normalised, which is what the approximation guarantee requires — with $g(\emptyset) = c > 0$ the Nemhauser–Wolsey–Fisher bound reads $g(\mathcal{S}_{\text{greedy}}) - c \ge (1 - 1/e)[g(\mathcal{S}^\star) - c]$, which is weaker than the stated form. The calibrated critic emits reliabilities in $[0, 1]$, so $q_b \ge 0$ holds by construction. Monotonicity is immediate. For submodularity, take $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{U}$ and $e \notin \mathcal{B}$. Then

$$g(\mathcal{A}\cup\{e\}) - g(\mathcal{A}) = \big(q_e - \max_{\mathcal{A}} q\big)^+ \;\ge\; \big(q_e - \max_{\mathcal{B}} q\big)^+ = g(\mathcal{B}\cup\{e\}) - g(\mathcal{B}),$$

since $\max_{\mathcal{A}} q \le \max_{\mathcal{B}} q$ and $x \mapsto (q_e - x)^+$ is non-increasing. Thus $g$ is monotone submodular pointwise; taking expectations preserves both properties (they are preserved under non-negative combinations), so $f(\mathcal{S}) = \mathbb{E}[g(\mathcal{S})]$ is monotone submodular for an arbitrary joint law of $(q_b)$.

The $(1 - 1/e)$ guarantee for greedy under a cardinality constraint is Nemhauser, Wolsey and Fisher [32]; under a knapsack constraint, greedy with partial enumeration attains the same factor [44]. For an $\epsilon$-approximate value oracle, each greedy step loses at most $\epsilon$ relative to the exact greedy step, and the standard inductive argument accumulates at most $|\mathcal{S}|\epsilon$ additive loss. □

### Proof of Corollary 26 (selected-branch value saturates)

Let $\mathcal{S}_B$ be the greedy set of size $B$. Submodularity gives $f(\mathcal{S}_{B+1}) - f(\mathcal{S}_B) \leq f(\mathcal{S}_B) - f(\mathcal{S}_{B-1})$ for the greedy sequence, i.e. the marginal gains are non-increasing, so $B \mapsto f(\mathcal{S}_B)$ is concave. That is the whole of what is proved. The extension to accuracy requires accuracy to be a non-decreasing *concave* function of $f$, which is an assumption and not a consequence: monotonicity of the reader's success probability in branch utility gives monotonicity only, and a bounded accuracy curve may be convex at small $B$. We therefore report $f(\mathcal{S}_B)$ alongside accuracy and treat only the former as constrained by this corollary (§8.4). □

### Proof of Corollary 27 (confidence must be paid for)

Let $\mathcal{G}_t$ be the $\sigma$-algebra generated by the reasoner's history through step $t$, including all imagined latents. For a Bayes-consistent reader, $\hat{p}_t = \mathbb{P}(Y = \hat{y} \mid \mathcal{G}_t)$, which is a bounded martingale by the tower property; by Proposition 6 the conditioning on the imagined latents is vacuous, so in fact $\hat{p}_{t+1} = \hat{p}_t$ a.s. across imagination steps and ICD $= 0$.

For a bounded reader, $\hat{p}_t$ is a computable approximation and may move. Decompose ICD $= \mathbb{E}[\hat{p}_{t+1} - \hat{p}_t]$. If the reader is calibrated at level $\epsilon$ at both steps (Definition 28), then $\mathbb{E}[\hat{p}_s]$ is within $\epsilon$ of $\mathbb{P}(Y = \hat{y}_s)$ for $s \in \{t, t+1\}$, hence

$$\text{ICD} \leq \mathbb{P}(Y = \hat{y}_{t+1}) - \mathbb{P}(Y = \hat{y}_t) + 2\epsilon = \Delta\text{Acc} + 2\epsilon.$$

Contrapositively, ICD $> \Delta\text{Acc} + 2\epsilon$ certifies that the reader is not calibrated at level $\epsilon$ at both steps — i.e. confidence rose without a matching accuracy gain. The admissibility gate is written with a single tolerance $\epsilon_{\text{tol}}$ playing the role of $2\epsilon$, which is the form the implementation exposes. □

### Auxiliary: the toy diagnostic in the reference implementation

The reference implementation ships a deterministic toy engine that satisfies the algebraic properties above exactly (invertibility, phase preservation, unit DC gain of the memory, per-band time constants) so that the propositions can be checked numerically without any pretrained backbone. The corresponding assertions are in `tests/test_spectral.py` and `tests/test_theory.py`; see Appendix B.

## B Reference Implementation

The repository accompanying this manuscript implements the control protocol, the spectral workspace, the budgeted gate, the calibration machinery, the promotion gates and the diagnostic generators. It runs on CPU with only `torch` installed; pretrained backbones are an optional extra. The design principle is that every claim in §3 should be checkable against running code, so that a stated guarantee cannot quietly diverge from what the system does.

| Module | Implements | Key objects |
|---|---|---|
| `spectral.py` | §3.4, §3.5, §3.3 | `SpectralBandMixer` (with `inverse`, `realised_band_gains`, `condition_number_bound`, `achievable_gain_ratio`), `BandwiseSpectralMemory` (with `time_constants`), `task_weighted_water_filling`, `band_curvature_sensitivity` |
| `bridge.py` | §2.4, §4 | `VisionToLanguageBridge` with zero-initialised gated residual injection |
| `voi.py` | §2, §4.5 | `CVOIEstimator`, `CostModel`, `ShadowPrice` (dual ascent), `BudgetedGate`, `doubly_robust_gain` |
| `calibration.py` | §3.8 | `Provenance`, `ProvenanceCalibrator` (stratified temperature + reliability cap), `imaginal_confidence_drift` |
| `critics.py` | §3.7 | `CalibratedEvidenceCritic`, `greedy_branch_select`, `expected_max_value` |
| `controller.py` | §4.5 | `VOIController` (default), `ImaginationController` (thresholded baseline) |
| `model.py` | Algorithm 1 | `VisualImaginalReasoner` |
| `recursive_improvement.py` | §6 | `GateThresholds`, `EvaluationResult`, `VerifiedSelfDistiller` |
| `diagnostics.py` | §D | four generators with exact verifiers |
| `losses.py` | §5 | `water_filling_regulariser`, `cvoi_regression_loss`, `pinball_loss`, `branch_diversity_penalty` |

**Table 17:** Where each part of the paper lives in the code.

## B.1 Module map

## B.2 Propositions as tests

`tests/test_theory.py` checks the analytical claims numerically:

- masks form a partition of unity and the effective gate is strictly positive even with $|a| \to 1$ (Lemma 15);
- maximum phase error across all retained frequencies is below $10^{-4}$ radians (Prop. 16);
- $S_\omega(\text{roll}(z)) = \text{roll}(S_\omega(z))$ exactly (Prop. 17);
- $S_\omega^{-1}(S_\omega(z)) = z$ to $10^{-4}$, and the two-sided norm and distortion bounds hold on random draws with $\kappa < 1.5$ (Prop. 18);
- the workspace has unit steady-state gain (a constant input converges to itself), reduces to the scalar memory when decays are uniform, and has monotone per-band time constants by default (Prop. 20);
- the zero-initialised $V \to L$ gate makes injection the exact identity at initialisation, so free disposal holds architecturally rather than aspirationally (Prop. 7);

- the reliability cap binds *on the distribution*, not only on a scalar: capping the top class and renormalising the whole vector would hand the excess straight back, so the implementation redistributes the freed mass to the other classes and the test asserts $\max_i p_i = \pi_{\text{IMG}}$;
- realised band gains are strictly narrower in spread than the nominal coefficients, confirming that correlating the latter against sensitivity would test a quantity the operator never applies;
- water-filling exhausts the rate budget, assigns lower distortion to higher-weighted bands, and gives zero rate to bands below the water level (Prop. 11);
- expected-max marginal gains are non-increasing on nested sets (Prop. 25);
- dual ascent raises $\beta$ under persistent overspending and lowers it under underspending (Prop. 10);
- confidence drift flags a 0.4 rise in confidence with zero accuracy change as unhealthy, and accepts the same rise when accuracy rises with it (Cor. 27);
- the reliability cap binds: a raw critic score of 0.99 is returned as 0.60 when $\pi_{\text{IMG}} = 0.60$.

Two further tests exist mainly as guards against silent breakage: with zero gains the operator is the exact identity (so B7 with $a = 0$ *is* B3 — the workspace memory keeps running — by construction rather than by configuration), and a sufficiently large $\beta$ suppresses imagination entirely.

### B.3 Executables

- `scripts/theory_checks.py` — runs the operator guarantees, the sensitivity measurement, the water-filling allocation and the cost accounting, and prints them as JSON. Three of the four theory-driven measurements of §8 need no pretrained backbone, which is the point: they are cheap enough to run before committing compute to a full experiment.
- `scripts/run_demo.py` — one question, with the full gate log: what each configuration was predicted to be worth, what it would cost, and why the controller spent or skipped.
- `scripts/train_stage1_alignment.py` — Stage-1 scaffold including the water-filling regulariser, with `-lambda-spec 0` for the unregularised headline measurement.
- `scripts/train_stage2_joint.py` — trajectory scaffold with $\epsilon$-exploration, doubly robust targets, and dual ascent on $\beta$.
- `scripts/eval_benchmarks.py` — reports accuracy *and* cost, imagination rate, ECE, and the text-only control rate. Accuracy alone is not printed anywhere, deliberately.
- `scripts/recursive_improve.py` — demonstrates that each promotion gate independently blocks a candidate that looks better on capability alone.

### B.4 Extending to pretrained backbones

The toy engine is a deterministic stand-in, not an image generator; it exists so the control protocol can be exercised without downloading foundation models. Three substitutions convert the scaffold into a research system, and none of them changes the protocol:

1. replace `encode_text` and the answer synthesiser with a Transformers-compatible instruction model, returning hidden states rather than decoded strings;

2. replace `ToyImaginationEngine` with `DiffusersImaginationEngine`, whose `generate_image` already supports scheduled spectral injection at depth D2 through the sampler's `callback_on_step_end`, with `default_schedule` concentrating injection in the early, layout-determining steps;
3. replace the toy convolutional encoder with a SigLIP/CLIP-class encoder, and fit the cost model constants in (5) to the chosen checkpoints — these constants are what make the gate's decisions meaningful, so they should be measured on the target hardware, not copied.

### B.5 Reproducibility

Seeds are set and reported by the evaluation driver; checkpoint hashes, licences and revisions go in the manifest. The diagnostic suites are regenerated from a seed at evaluation time rather than stored, which is what makes them contamination-proof. `MANIFEST.sha256` covers the repository contents.

## C Hyperparameters and Cost Constants

Values marked **[claim]** encode a substantive commitment and are swept as ablations; the rest are conveniences.

### C.1 Workspace

### C.2 Controller and budget

### C.3 Cost constants

The constants below are the reference defaults in `CostModel`; they are order-of-magnitude placeholders and *must* be re-measured on the target checkpoints and hardware before the gate's decisions mean anything.

### C.4 Training

### C.5 Backbone configuration

The reference `configs/base.yaml` names an instruction LLM, an SDXL-class Diffusers checkpoint and a SigLIP-class encoder. These are placeholders: model identifiers are configuration, and Stage 0 (§5.1) exists precisely to reject a combination whose generator cannot render the relations the task asks about or whose encoder cannot recover them. Exact revisions, licences and hashes are recorded in the manifest for every reported run.

## D Diagnostic Task Generators

Standard benchmarks cannot settle the questions in §8. They are contaminated to an unknown degree, they mix the capability under test with a dozen others, and they offer no control condition in which imagery is known to be useless. We therefore generate four families programmatically, with exact verifiers and controllable difficulty, and regenerate them with fresh seeds at evaluation time — which makes them contamination-proof by construction rather than by assertion.

| Parameter | Default | Notes |
|---|---|---|
| Bands $K$ | 6 | **[claim]** enough to separate layout from texture; swept over $\{3, 4, 6, 8, 12\}$ |
| Mask family | Gaussian, $\varsigma = 1.25/K$ | partition of unity to $10^{-5}$; the smoothness matters for Prop. 17 |
| Residual strength $\rho$ | *selected*, not fixed | **[claim]** §7.3 found the reference 0.20 too small. Choose the smallest $\rho$ whose *realisable* gain range covers the measured sensitivity range: measure $\max_k w_k / \min_k w_k$ on the task family, then pick $\rho$ with (realisable $G$ range)$^2$ above it. At $\rho = 0.20$ the realisable range is 1.333, so the covered span is only 1.78 — tighter than the theoretical 2.25, because overlapping masks stop any band reaching its own coefficient. Distortion is bounded by $\rho\bar{a}$, so this is a real trade |
| Gain parameterisation | tanh | guarantees $\bar{a} < 1$, hence strict positivity of $G_\rho$ |
| Band decays $\lambda_k$ | $0.95 \to 0.30$ | **[claim]** monotone: gist outlives detail. Uniform-$\lambda$ is the ablation |
| Time constants $\tau_k$ | 19.5 $\to$ 0.8 steps | derived, reported per run, not set |
| Latent channels $C$ | 4 | follows the SD-family autoencoder |
| Imaginal resolution | $32 \times 32$ latent | justified by water-filling, not by compromise; $16^2$ and $64^2$ swept |
| Integration depth | D2 | scheduled injection; D1 and D3 both reported |
| $\mathcal{T}_S$ | first third, every 3rd step | **[claim]** layout is decided early; placement swept |
| Transport mode | `translation` | `none` is the ablation; `similarity` adds an $R \times S$ search and is the one workspace component whose cost is not negligible |
| Transport margin $\eta$ | **0.10** | **[claim]** raised from 0.02 on the measured trade-off curve of §7.6: genuine translations are accepted at every threshold, while unrelated-scene acceptance falls from 67% to 20% |
| Transport max shift | 0.5 of the grid | set for the expected magnitude of viewpoint change; a tighter bound rejects larger displacements outright |

## D.1 The four families

**Mental rotation (SPATIAL).** Given a shape and a described transformation, decide whether the result is a pure rotation or a mirror image. Difficulty is the angular disparity. This is the canonical imagery task in the psychological literature — response latency scales with the angle of rotation [41] — and it is where the theory predicts a low-band-dominated sensitivity spectrum, so it is the natural place to look for the gain–sensitivity correspondence of Prediction 1.

**Route planning (SPATIAL).** Shortest-path length on an $n \times n$ grid with obstacles, verified by breadth-first search. Difficulty scales both $n$ and the obstacle count. Two properties make this useful: the ground truth is exact and free, and the answer is a small integer, so partial credit and grader noise do not enter.

**Counterfactual scene (COUNTERFACTUAL).** A described arrangement of objects undergoes a sequence of edits; the question asks about the final configuration. Difficulty is the number of edits,

| Parameter | Default | Notes |
|---|---|---|
| Budget $\mathcal{B}$ | 120 000 nFLOPs | in units of one decoded token; *swept* to trace the frontier. Must exceed the widest menu entry ($\approx$ 87k) or that entry is pruned before it is scored |
| $\beta$ init / lr | $10^{-5}$ / $10^{-8}$ | solved by dual ascent, never tuned directly. At $\beta = 10^{-5}$ the "default" step is priced at 0.29 reward units, which is the right order for a predicted gain in $[-1, 1]$ |
| $\epsilon$-exploration | 0.10 | **[claim]** required for identifiability of the skip branch |
| Max imagination steps | 3 | |
| Configuration menu | cheap / default / wide | $(B, S, \text{res}) = (1, 8, 16^2), (3, 20, 32^2), (6, 30, 32^2)$ |
| Band-allocator init bias | $-1.5$ | **[claim]** starts the policy sparse; zero-init collapses to all-bands-active (§7.4) |
| Band entropy bonus | 0.02, decayed to 0 | exploration bought early, not paid for at convergence |
| Uncertainty features | 5 | entropy, self-consistency, critic spread, learned confidence, task-type prior |
| Quantile $q$ (optional) | 0.25 | risk-averse gate without changing $\beta$ |

| Quantity | Default (nFLOPs) | Basis |
|---|---|---|
| One decoded language token | 1 | unit of account, $2N_L$ |
| One denoiser step, $32^2$ latent | 240 | scales with $(\text{res}/32)^2$ |
| Classifier-free guidance factor | $\times 2$ | two denoiser evaluations per step |
| VAE decode+encode | 900 | skipped in latent-only mode |
| Vision encoder call | 60 | |
| Spectral operator | 0.05 | $O(CHW \log HW)$; negligible by design |
| "cheap" step ($1\times 8$ at $16^2$) | 1 020 | $\approx$ 1 020 decoded tokens |
| "default" step ($3\times 20$ at $32^2$) | 28 980 | $\approx$ 29 000 decoded tokens |
| "wide" step ($6\times 30$ at $32^2$) | 86 760 | sets the floor on $\mathcal{B}$ |

**Table 18:** Why the gate exists. Even the cheapest imagination step costs about as much as a thousand tokens of thinking, so a system that imagines by default cannot win an equal-compute comparison.

which is also the number of workspace updates the model must carry — so this family is the one that should degrade when memory is reset each step or when $\lambda$ is forced uniform (§8.4).

**Arithmetic control (TEXT_ONLY).** Two-digit arithmetic. Imagery is provably useless here, and the target behaviour is *not* accuracy: the system should match its backbone while calling the generator approximately zero times. This family is the single most informative one in the suite, because it is the only place where the gating claim can fail cleanly. A model that imagines here has not learned the policy this paper is about, whatever it scores elsewhere.

## D.2 Why generators rather than a fixed set

Three reasons, in order of importance. A generator can be re-run with a new seed after the model is trained, so contamination is impossible rather than merely unlikely. Difficulty is a dial, which lets us measure *where* on the difficulty curve imagery starts paying — the theory says the benefit should be largest where the answer is hard to compute from $h_t$ but easy to read off from $v_t$, and a single difficulty level cannot show that. And ground truth is exact and free, which is what makes

| Parameter | Default | Notes |
|---|---|---|
| Optimiser | AdamW | lr $2 \times 10^{-4}$, weight decay 0.01, cosine schedule |
| $\lambda_s$ (water-filling) | 0.1 | **[claim]** swept *to zero*: Prediction 1 is about unregularised gains |
| $\gamma$ (branch ranking) | 1.0 | |
| $\mu$ (branch diversity) | 0.05 | prevents $B$ branches collapsing to $B$ copies |
| $\lambda_k$ (calibration) | 0.1 | the hard cap applies at inference regardless |
| Seeds | 5 | minimum; results reported as mean with bootstrap CI |
| Self-generated data share | $\leq 50\%$ | fixed and reported per cycle |

the paired-rollout cVOI targets of §5.4 affordable at the scale the controller needs.

### D.3 Extensions we recommend but have not implemented

- **Rendered variants.** The families above are described in text so they can run without a renderer. A rendered version — where the model sees an actual image and must imagine its transformation — separates "imagining from a description" from "transforming an observation", which are different capabilities and probably have different sensitivity spectra.
- **Occlusion and object permanence.** Objects that leave and re-enter view across a sequence of viewpoint changes, which stresses the workspace's memory horizon directly and would let $\tau_k$ be validated against a known required horizon.
- **Anisotropic tasks.** Text-in-image and ruled-diagram tasks, where radial masks are the wrong basis. These are the natural falsifier for the isotropy assumption in §3.4 and would motivate a steerable-pyramid variant.
- **Memorisation probes.** Instances whose surface form is perturbed without changing the answer; a drop indicates the model is matching strings rather than reasoning.
- **Adversarial imagery.** Instances where the generator's scene prior is systematically *wrong* — unusual configurations that a typical image model will smooth toward the typical case. This is the empirical version of Proposition 8, and we expect it to be where imagined evidence hurts most.

### D.4 Usage

```
from vllm_si.diagnostics import generate_suite
items = list(generate_suite(n_per_family=25, seed=1234, difficulty=2))
# each item carries: question, answer, family, difficulty, imagery_useful, metadata
```

The `imagery_useful` flag is what the text-only control rate in §8.5 is computed against, and `metadata` carries enough to re-verify every answer independently of the generator that produced it.

## E Reporting Checklist

This checklist is written to be used against our own future results. Its purpose is to make the difference between "the mechanism works" and "we spent more compute" impossible to blur after the fact.

### E.1 Before running anything

☐ Preregister the primary comparison (B7 vs. B1 and B7 vs. B2 at matched cost), the smallest effect size considered meaningful, and the number of seeds.

☐ Run a power analysis at that effect size; if the available instances cannot support it, say so now, not afterwards.

☐ Fix the promotion gates and their thresholds; version them.

☐ Fix the recursive-improvement cycle count and the alpha-spending schedule.

☐ Record Stage-0 backbone qualification results, including the ones that caused a configuration to be rejected.

### E.2 With every result

☐ Accuracy **and** cost, always together. Normalised FLOPs from (5), plus wall-clock, peak memory and an energy proxy.

☐ The equal-compute baseline B1. Never report a gain against B0 alone.

☐ Imagination calls per example; imagination rate per task family; the rate on the text-only control suite specifically.

☐ Accuracy conditional on imagining and on skipping, and the oracle-gate gap (B6 − B7).

☐ Accepted-branch fraction, and randomly sampled traces at a fixed rate including failures.

☐ ECE overall and stratified by provenance; ICG; ICD against ΔAcc; the measured $\pi_{\text{IMG}}$.

☐ Learned $|G_k|$, $\lambda_k$, derived $\tau_k$, and measured $w_k$ per task family.

☐ Seeds, model revisions and hashes, tokenizer versions, licences, hardware, and the exact cost constants used.

☐ Five seeds minimum; mean with bootstrap CI; multiple-comparison correction across the ablation table, stated.

### E.3 Theory-driven measurements

☐ Prediction 1: fitted slope of $\log G_k$ against $\log \hat{w}_k$ tested against $1/2$, using *realised* band gains and the curvature estimator, with $\lambda_s = 0$, per task family, with a permutation null. Report the measured dynamic range of $w_k$ next to $\rho$, since (16) bounds what the operator can express.

☐ Prediction 2: reliability diagram of $\widehat{\text{cVOI}}$ against realised paired gain, with its slope; and the accuracy–compute frontier from sweeping $\beta$.

☐ Prediction 3: advantage over B1 as a function of B1's token budget, over at least a decade.

☐ Branch saturation: selected-set critic value *and* accuracy versus $B \in \{1, 2, 4, 8, 16\}$; concavity is required of the former only.

☐ Integration-depth ordering D1 < D2 < D3, or an explanation of why not.

### E.4 Negative and awkward results

☐ Report the null if B2 matches B7. Do not retitle around it.

☐ Report the ablation profile even when it is uniform, which would falsify C3.

☐ Report configurations rejected at Stage 0 and why.

☐ Report any gate that was changed mid-programme, at which cycle, and reset the cycle counter.

☐ Report every stop-and-audit condition that triggered (non-saturating branch curve, flat budget sweep, gate firing on the control suite).

☐ State explicitly whether the cycle chain was extended beyond the preregistered count; if so, the sequential guarantee does not hold and the promotions are exploratory.

### E.5 Audit your own favourable results

Added after the pilots, because the two analysis errors we made both flattered the hypothesis and neither was caught by the analysis that produced it (Appendix F).

☐ For every result that supports a claim, have someone who did not run it re-derive the headline number from the stored per-run data.

☐ Check that aggregate columns in a table are aggregated over the same runs. A single-seed quantity beside a multi-seed one is the specific error we made.

☐ Check that no fitted quantity exceeds a bound the same table asserts. Ours did, in four of six cells, and we did not notice.

☐ Check that floored, censored or imputed values are excluded from fits rather than clamped and used.

☐ Verify that the rule the experiment implemented is the rule the paper specifies. Ours thresholded at zero where the paper prices against a shadow price, and the difference changed which criterion passed.

### E.6 Claims language

☐ Do not describe imagined content as "evidence" without qualification (Prop. 6). "Imagined hypothesis" or "self-generated visual state" is accurate.

☐ Do not attribute a gain to "the model gathering information". It gathered none.

☐ Do not present "recursive self-improvement" without the qualifier "offline, verified, gated".

☐ Do not show a qualitative example without stating how many branches were discarded to produce it.

☐ Do not report an accuracy number in a table that lacks its cost column.

## F Pilot Details and Reproduction

Everything in §7 runs on CPU with only `torch`, `numpy` and `matplotlib` installed. Total wall-clock for the full set is around forty minutes on two cores.

## F.1 Reproduction

```
export PYTHONPATH=src
python experiments/pilot_a_gain_sensitivity.py --seeds 5 --gate-steps 700
python experiments/pilot_bcd_controller.py    --seeds 3 --steps 900
python experiments/pilot_e_transport.py       --episodes 30
python experiments/bench_costs.py
python experiments/make_figures.py
```

Each script writes a JSON file to `experiments/results/` containing its full configuration, every per-seed run, and the summary statistics quoted in the paper; figures are regenerated from those JSON files alone, so every number in §7 is traceable to a file in the repository.

## F.2 SpectralWorld constants

| Parameter | Value | Note |
|---|---|---|
| Latent size $H \times W$, channels $C$ | $32 \times 32$, 4 | matches an SD-family latent shape |
| Blobs per scene, $\sigma$ | 2, 3.0 px | layout component; low-frequency energy |
| Texture cutoff | $r > 0.45$ | high-pass; detail on one side only |
| Random-Fourier bandwidth | **14.0** | the most consequential constant; see below |
| Prompt dimension | 32 | 2 clear slots (number, parity) + 3 one-hot (family) + 27 RFF |
| Train / eval instances (Pilot A) | 5000 / 800 | as recorded in that pilot's JSON |
| Code noise | 0.15 | perturbs the code before rendering |
| Post-gate noise floor $\epsilon$ | 1.0 (C/D), calibrated (A) | the floor the gate trades against |
| Viewpoint jitter | $\pm 3$ px | between imagination steps |
| Bands $K$ | 6 | as in the reference configuration |
| Train / eval instances (B–D) | 5000 / 800 | regenerated per seed |
| Seeds | 5 (A), 3 (B–D, E) | fewer than §8's minimum of five for B–D; stated |
| Reader depth, optimiser | 2, AdamW $2 \times 10^{-3}$ | width is the swept capacity dial |

**On the bandwidth constant.** At bandwidth $\leq 10$ a reader of any size inverts the random Fourier map outright, imagery never helps, and there is nothing to measure. At $\geq 20$ no reader of any size inverts it, so the advantage of imagery is constant in capacity and Prediction 3 is untestable. At 14 the prompt task is hard but solvable, which is the only regime in which "the benefit is computational" makes an observable difference. We report this because it is a real limitation of the pilot: the decay result exists in a window that had to be found, and a reader is entitled to know how wide the window is.

## F.3 Design errors we made, and what they looked like

Recorded because each one produced a plausible-looking result that was wrong, and because the same mistakes are available to anyone building this.

1. **No imagination dropout.** Training the reader with visual features always present makes reader$(p, \varnothing)$ out of distribution. Every task family then shows a large apparent gain from imagery

— including the text-only control, where the true gain is zero — and the gate has no abstention signal to learn. Symptom: imagination rate near 0.65 on all three families, identically.

2. **Noise floor before the gate rather than after.** With the noise applied first, the gate has no signal-to-noise to reallocate, and the spectral condition is mathematically identical to the generic one. Symptom: byte-identical per-family accuracies between conditions.

3. **No scale normalisation after the gate.** The dominant gradient becomes "amplify everything", which swamps the allocation signal. Symptom: all six gains below 1 for one family and all six above 1 for the other — a global scale masquerading as an allocation.

4. **Sensitivity by gradient power, at a fixed small probe.** At a well-fitted solution the loss is flat enough that the estimate is indistinguishable from noise, and every band reads as zero. This looks like "no sensitivity anywhere" and means "probe too small". Fixed by measuring curvature with an adaptively scaled, band-shared probe.

5. **A texture task whose statistic was a ratio.** Asking "which half is finer" when both halves carry texture gives a task that is nearly invariant to added broadband noise, so its measured band sensitivity is near zero everywhere and it is useless as a high-band probe. Sensitivity has to be a property one can degrade.

6. **Dropping the question identifier from the prompt.** A reader that cannot tell which of three incompatible predicates it is being asked for averages them and sits near chance on all three. Symptom: every condition at $\approx 0.66$, regardless of architecture.

7. **Benchmarking while another job held the cores.** Produced wall-clock non-monotonic in latent size — a $16^2$ gate apparently slower than a $64^2$ one. Fixed by single-threading and taking the minimum over rounds.

8. **Floored sensitivity values entering the log-log fit.** The curvature estimator floors its output at $10^{-6}\times$ the maximum so a log stays defined. A floored band is an *unmeasured* band, not a band with sensitivity $10^{-6}$; letting one into the regression contributes an abscissa spanning six decades and can dominate the fit. Excluding them changed the reported slopes by roughly a factor of three — downward.

9. **A single-seed ceiling quoted beside a five-seed slope.** The expressible ceiling depends on that run's own measured sensitivity range, so it is a per-run quantity and must be averaged like every other column. Quoting seed 0's ceiling next to a five-seed mean slope produced a table in which the fitted slope exceeded the operator's stated maximum in four of six cells — visibly impossible, and we did not notice until an adversarial re-read.

10. **Describing the pilot's gate as "priced".** Pilot B thresholds predicted gain at zero. Eq. (8) compares against $\beta c_{\text{img}} > 0$. These are different rules, and the difference is exactly why the text-only abstention criterion failed.

*The two analysis errors (items 8 and 9) both flattered the hypothesis, and neither was caught by the analysis that produced them.* That asymmetry — favourable bugs surviving longer than unfavourable ones — is the reason §E now asks for an audit aimed specifically at results that came out in the paper's favour.

### F.4 Threats to the validity of the pilots

- **The world is built to satisfy the theory's assumptions.** Band energy is genuinely separable by radial frequency, the offsets are genuinely cyclic translations, and the noise floor is genuinely white in the transform domain. Real latents satisfy none of these exactly. Passing here is a necessary condition, not evidence of transfer.
- **Three seeds for Pilots B–D.** Adequate for the large effect (the budget decay) and inadequate for a small one; the spectral-versus-generic null in particular is a null at three seeds, not a demonstration of equivalence.
- **Ceiling effects.** The task imagery helps with reaches 0.928 pooled with generic imagery (and 0.974 at the narrowest reader), leaving little room for a mechanism to improve on it. A harder texture task would be a better test and is the obvious next iteration.
- **Two fusion steps.** Proposition 23 predicts that unregistered fusion degrades *multiplicatively* with depth, so a two-step pilot samples the mildest end of the regime where transport should matter. Pilot E's dedicated six-step episodes show the effect clearly; Pilot C's two-step setting is where it would be smallest.
- **The encoder is weak.** It cannot recover the left-of relation at all, which is why the layout family shows no benefit from imagery. That is a property of this encoder, not of the idea, but it does mean the pilot tests the mechanism on one task family rather than two.

## G Direct Answers to Anticipated Questions

Each answer points to where the substance lives; they are collected here so that a reader with a specific question does not have to reconstruct the answer from four sections.

### What is the memory and computational overhead of the spectral decomposition and the transport alignment, relative to an ordinary generated-image baseline, and how is it handled in the matched-compute protocol?

Measured, in Table 9. At a $32 \times 32$ latent with batch 8 on one CPU core: gate 0.63 ms, memory update 0.75 ms, transport 0.38 ms — 1.76 ms total, with an analytic working set of 0.42 MB. All three are $O(CHW \log HW)$; a 64× increase in coefficients cost 21.7× the time — *sub*linear, because the smallest latents are overhead-dominated, so the measurement bounds the cost from above rather than confirming the exponent (Table 3).

Against the cost model, the spectral term is $1.7 \times 10^{-6}$ of a default imagination step and $4.9 \times 10^{-5}$ of the cheapest one — four to six orders of magnitude below the denoiser call it accompanies. The consequence for §8.2 is that the protocol matches on *generator calls, steps and resolution*, and does not need to equalise the workspace components, because equalising them would change nothing measurable. We state that as a measured claim rather than an assumption precisely because the reverse would invalidate every comparison in the paper.

Transport in `similarity` mode is the one component that can become non-negligible: its cost is $O(RS \cdot CHW \log HW)$ for $R$ rotations and $S$ scales searched. At the default $R = S = 1$ it reduces to the translation case; if a task needs a wide search, that cost must enter the accounting, and the implementation reports it.

### How is the controller's band allocation initialised so that it does not collapse to all-bands-active, given how expensive exploring the branch space is?

Three measures, and then a measurement (§4.6, §7.4).

The diagnosis first: a logit head initialised at zero gives $\pi^k = 0.5$ everywhere, and once task loss starts flowing the cheapest early move is to stop discarding information, so the policy collapses to all bands. We observed exactly this before adding the fix.

The measures are (i) a negative output bias at initialisation ($-1.5$, giving $\pi \approx 0.18$), so the policy *starts* sparse and each band must earn activation against its price; (ii) an entropy bonus decayed linearly to zero, so exploration is bought early when it is cheap and is not still being paid for at convergence; and (iii) logging allocation entropy and active-band count every step, so collapse is observed rather than inferred.

On exploration cost: the allocator's target is per-band marginal utility measured by *ablation* rather than by search over the joint branch space. That is $K$ extra forward passes on a subsample at a fixed interval, not $2^K$ configurations, which is what makes the measurement affordable.

Measured outcome in Pilot B (three seeds): mean active bands settles at 0.12–0.18 of six, with allocation entropy 0.385 against a maximum of 0.693. No collapse — though the evidence is weaker than it looks, because the band-ablation target was identically zero in 17 of 33 logged steps, and because the pilot counts active bands by $\pi_k > 0.5$ rather than by the priced rule (29). The initialisation fix is supported; its behaviour under strong task pressure is not yet tested.

### What is the concrete plan for evaluating the transport operator on viewpoint-varying tasks, and what failure modes do you anticipate?

The plan is executed in Pilot E (§7.6), on episodes of six fusion steps under ±3-pixel jitter. Measured: registration accuracy against known shifts (exact recovery in 100% of trials at observation noise up to 3.0); retained high-frequency energy, $0.580 \to 0.686$ with alignment; and registered reconstruction error, $0.160 \to 0.0247$, a factor of 6.5, with disjoint confidence intervals.

The anticipated failure mode is a pathological warp — the operator aligning two scenes that are not views of one another — and it is real: at the original acceptance margin $\eta = 0.02$, 55% of *unrelated* scene pairs were accepted. Sweeping $\eta$ showed genuine translations are accepted at every threshold while unrelated pairs fall from 67% to 8%, and we changed the default to $\eta = 0.10$ on that curve. Two structural guards accompany it: the warp family is low-parameter, so it can move existing structure but not invent any, and displacements beyond the plausible range are rejected outright — measured accept rate 0.000 with the shift bound tightened to 5% of the grid, though at the default bound of 50% a large displacement is inside the plausible range by construction, so that bound has to be set for the expected magnitude of viewpoint change.

Two caveats on the headline numbers. They were measured with the acceptance guard *open* ($\eta = 0$), so that alignment is exercised on every step rather than sometimes declining; and the unregistered reconstruction MSE runs the other way (naive 0.382, aligned 0.616), because an aligned memory is sharp but sits at the first view's offset and an unregistered comparison rewards blur for being centred. Both are in the released JSON.

On the full system, the same three quantities are reported per run, and the transport rejection rate is a first-class trace field: a run in which alignment is silently rejecting every step looks different from one in which it is working.

### Given that this is a methods proposal, what preliminary results exist?

§7, in full, with the numbers in JSON in the repository. In brief:

Eight testable statements: four supported, one weakly, three not (Table 10).

- **Supported:** the gain–sensitivity correspondence on the texture-dominated family (slope positive at every $\rho$, CI excluding zero, rising with the operator's range, direction matching in 4/5 seeds); the gate learning the value structure (+0.47 on the helped family, $\approx 0$ on the others); alignment beating naive fusion (6.5× lower registered reconstruction error); and the budget-decay prediction ($+0.141 \to +0.088$ over reader widths $8 \to 256$).
- **Weakly supported:** band-allocation non-collapse, for the reasons above.
- **Not supported:** the gain–sensitivity correspondence on the layout-dominated family (slope $\approx 0$, direction matching in 0–1 of 5 seeds); the gate's abstention on the text-only control (rate 0.48, not near zero — though the pilot used an unpriced rule, so this tests thresholding rather than the gate the paper specifies); and, most consequentially, the spectral workspace against an equal-compute generic-imagery condition ($-0.004$ to $+0.005$ across six reader widths), which is the outcome §9 identifies as refuting claim C2.

We also record, in Appendix F, that our first analysis of the gain–sensitivity pilot was buggy in the paper's favour and reported slopes roughly three times too large. It was caught in an adversarial re-check rather than by the analysis that produced it.

### Is the architecture over-engineered, and how would a gain be attributed?

Partly by construction and partly by protocol. Every component reduces to an exact identity under a setting that the test suite verifies — $a = 0$ makes the gate the identity, uniform $\lambda$ reduces the workspace to a scalar leaky average, `transport=none` makes alignment the identity — so an ablation cannot silently fail to ablate. And §4.12 specifies a build order in which each component is added only after the cheaper ones have demonstrably paid, with stages 1–3 involving no learned spectral parameters at all. A positive result at stage 2 would be evidence for the metering account and *not* for the spectral one, which is the sort of partial outcome we want to be able to report cleanly.

The honest residual: the full system has many interacting parts, and if all of them are switched on at once the attribution problem is real. That is an argument for the staged build, not for the assembly.

### Might a frozen diffusion model's latent space simply be unsuited to spatial reasoning?

It might, and Stage 0 is designed to find out cheaply rather than to assume otherwise (§5.1). We add an explicit probe — linear decodability of the target relations from the frozen latent, before decoding — and a stated fallback ladder: change the intervention point to a mid-network activation; adapt the encoder rather than the generator; LoRA the denoiser, at which point the frozen-backbone claim becomes a frozen-plus-adapters claim and we would say so; or declare the configuration unsuitable and report it. A generator whose latent space resists relational reading is a legitimate negative result about that generator, and more useful published than quietly worked around.

### How will the verifier's own error rate be accounted for?

§6.4. False positives and false negatives are asymmetric — a false positive poisons the replay buffer with errors concentrated exactly where the verifier is weakest, a false negative only narrows coverage — so the verifier is run conservatively. Its accuracy is estimated each cycle on a human-audited sample

*drawn from the retained set*, with disagreement between independent verifier implementations as a free lower bound. A task family whose estimated false-positive rate exceeds a preregistered ceiling contributes evaluation data only, for that cycle. That is a gate on the instrument rather than on the model, and it is the one most likely to be quietly skipped.

### What about fairness and bias if this is embedded in a larger system?

§10.5. The specific exposure is that an imagined scene may never be rendered, is consumed by the model itself, and shapes an answer presented as reasoning. Proposition 6 sharpens the question usefully: since imagined content carries no information about the answer, any systematic influence it has is attributable to the generator's prior rather than to evidence. We recommend three measurements — counterfactual prompt invariance with imagination forced on and off, imagination-rate parity across matched templates, and per-group stratified reliability. We have not run them; the geometric diagnostics of §D carry no demographic content. Provenance typing makes the imaginal path auditable, which is a precondition for measuring the problem, not a solution to it.