# Beyond Words: Spectral Imagination for Recurrent Visual Reasoning in Language Models

Andrew Kiruluta
UC Berkeley, School of Information
kiruluta@berkeley.edu

August 30, 2026

**Abstract**

Large language models can solve many tasks through linguistic computation, yet spatial simulation, mental rotation, viewpoint transformation, object persistence, and counterfactual scene reasoning naturally admit perceptual internal representations. We propose Spectral Imagination, a recurrent visual-reasoning architecture that couples a pretrained language model to a pretrained latent-diffusion generator and visual encoder while treating generated latents as an optional internal computational workspace rather than merely as output images. The central contribution is a *multiresolution spectral imagination state*: diffusion latents are decomposed into smooth frequency bands, a learned compute-aware controller chooses whether to imagine and which spectral scales to activate, and separate coarse-to-fine memories persist across reasoning steps. Before memory fusion, a learned latent transport operator aligns prior imagined state to the current hypothesis, reducing the failure mode of averaging spatially inconsistent generations. The selected imagined state is re-perceived, scored by a task-utility critic, projected back into the semantic state, and the controller is then re-evaluated recursively. We explicitly separate model-internal critic approval from objective external verification and use only the latter, when available, to support verified self-distillation. Recursive improvement is implemented as offline candidate generation, verification, training, immutable regression testing, and promote-or-rollback rather than unrestricted online self-modification. We define matched-compute baselines and mechanistic ablations designed to test whether spectral recurrence contributes beyond ordinary generated-image reasoning. The repository provides an executable toy control loop, multiresolution spectral memory, learned gating and band allocation, latent alignment, audit traces, pretrained-component adapters, standard benchmark manifests, matched-compute ablation tooling, and regression tests. This is a methods proposal and implementation scaffold; no benchmark gains are fabricated.

## 1 Motivation

Language is an efficient substrate for abstraction, but human cognition is not restricted to linguistic symbols. A reader can construct a scene while following a novel; a person can mentally rotate an object; route planning can involve an imagined map; and physical predictions can be aided by simulating perceptual consequences. The motivating hypothesis is therefore not that every problem requires imagery. It is that a general reasoner should possess an *optional internal perceptual workspace* and learn when using that workspace is worth its computational cost.

Modern multimodal systems already consume images, and unified architectures increasingly produce them. Input perception and output generation, however, do not by themselves imply a persistent, recursively editable visual workspace. Our target system treats imagined scenes as temporary computational objects: they can be generated, decomposed by scale, aligned to prior

internal state, selectively retained, re-perceived, scored for task utility, and replaced by counterfactual alternatives before the final linguistic answer is generated.

## 2 Related Work and Novelty Boundary

Latent diffusion made high-quality conditional synthesis practical in compressed autoencoder spaces [11]. Flamingo, BLIP-2, and LLaVA showed that strong pretrained language and visual modules can be connected through comparatively lightweight interfaces [1, 7, 8]. DreamLLM and Janus further explore unified multimodal understanding and generation [4, 14].

Intermediate visual reasoning is already an active research direction. Visual Chain of Thought introduced multimodal infillings [12]; Thinking with Generated Images constructs and critiques intermediate generated visual states [3]; Visual Thoughts studies multimodal reasoning traces [2]; Latent Sketchpad introduces an internal visual scratchpad [20]; and MindJourney couples a vision-language model to a diffusion world model for spatial test-time scaling [16]. BRAID formulates interleaved text-image reasoning as a unified decision process with multimodal credit assignment [5]. Consequently, "LLM plus image generation" is not a sufficient novelty claim.

Frequency-domain processing also has relevant prior art. FAM Diffusion uses Fourier-domain modulation to improve structural consistency in diffusion generation [15]. L2V-CoT reports low-frequency latent structure shared across LLM and VLM reasoning and transfers chain-of-thought information through frequency-domain latent intervention [19]. These works make it important to distinguish our claim from generic frequency manipulation. The proposed contribution is specifically *task-conditioned, recurrent, multiresolution control of imagined diffusion state as a reasoning workspace*, with re-perception, semantic recurrence, compute gating, and matched-compute mechanism tests.

Recent verification analyses also caution against treating a learned visual critic as ground truth: world-model test-time reasoning can benefit from verification, but poorly calibrated verifiers may reward uninformative or biased imagined views [6]. We therefore explicitly separate critic scores from external verification in both the method and implementation.

The intended novelty is the following combination, each component being experimentally falsifiable:

1. a **multiresolution spectral imagination state** with separate coarse-to-fine persistent memories rather than a single generated image or monolithic latent;
2. a **learned imagination controller** that selects both whether to imagine and the spectral resolution budget required for the current reasoning step;
3. a **latent transport alignment operator** that spatially aligns previous imagined memory to a new hypothesis before recurrent fusion;
4. **semantic recurrence after re-perception**: imagined evidence updates the language state, uncertainty is recomputed, and the controller decides again;
5. **counterfactual branch search** scored for downstream task utility, not image aesthetics;
6. a strict distinction between **critic approval** and **external/objective verification**;
7. **verified offline recursive self-improvement** with immutable regression gates and rollback; and

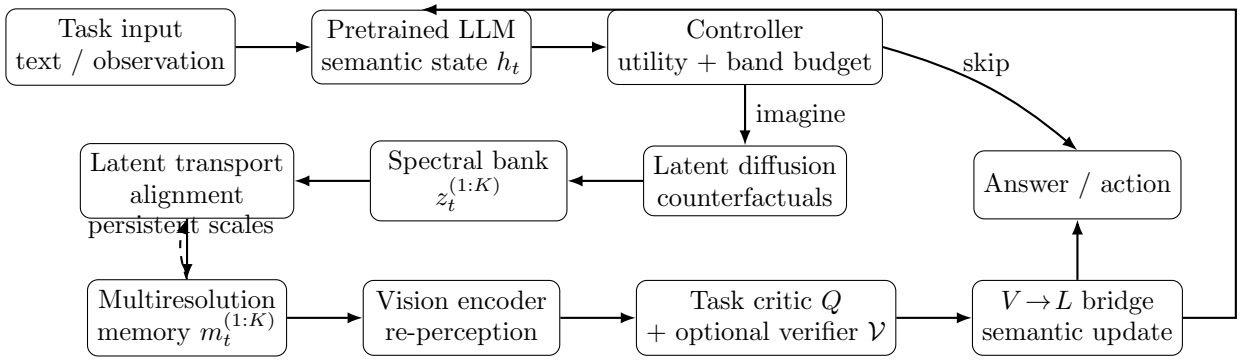


Figure 1: Spectral Imagination uses an optional recurrent visual workspace. The controller chooses both whether to imagine and which spectral scales to allocate. Re-perceived evidence updates the semantic state before the next controller decision.

8. an evaluation design requiring gains over an **equal-compute ordinary generated-image baseline** before attributing improvements to the spectral mechanism.

# 3 Architecture

## 3.1 System overview

Let task input $x$ contain text and optionally observed sensory input. A pretrained language model $L_\theta$ produces semantic state $h_t$. A controller $C_\phi$ predicts the net utility of imagination and a spectral allocation policy. If imagination is selected, a language-to-diffusion bridge $B_{L\to D}$ conditions a latent diffusion generator $D_\psi$ to construct counterfactual latent scenes. A spectral module decomposes each latent into $K$ scales. A learned transport operator aligns previous scale-specific memory to the new latent before recurrent fusion. A visual encoder $V_\nu$ re-perceives the selected imagined state, and a vision-to-language bridge $B_{V\to L}$ injects compact evidence back into the semantic state. A critic $Q_\eta$ predicts task utility; a separate verifier $\mathcal{V}$ may provide objective labels where such verification is possible.

## 3.2 Pretrained bootstrapping

The initial implementation freezes the expensive foundation backbones:

$$\theta, \psi, \nu \leftarrow \text{pretrained weights},$$
$$\text{trainable} = \{\phi, \omega, \gamma, B_{L\to D}, B_{V\to L}, \eta, R\},$$

where $\omega$ parameterizes spectral modulation, $\gamma$ the alignment operator, and $R$ the semantic recurrent update. Later stages may add low-rank adapters to selected language, diffusion, and visual blocks. This staged approach makes it possible to test the proposed mechanism before paying the cost of full multimodal fine-tuning.

### 3.3 Smooth spectral filter bank

For diffusion latent $z \in \mathbb{R}^{C\times H\times W}$, let $\hat{z} = \mathcal{F}(z)$. Construct $K$ smooth radial masks $M_k(\xi)$ satisfying approximately

$$\sum_{k=1}^{K} M_k(\xi) = 1. \tag{1}$$

The band-limited latent components are

$$z^{(k)} = \mathcal{F}^{-1}(M_k \odot \mathcal{F}(z)), \qquad k = 1, \ldots, K. \tag{2}$$

Low bands are biased toward global layout and large-scale geometry, while higher bands preferentially represent edges and local detail. This interpretation is an inductive bias rather than a theorem and must be tested empirically.

### 3.4 Controller-selected spectral computation

The controller receives semantic state $h_t$ and uncertainty $u_t$ and predicts

$$(p_t, \boldsymbol{\pi}_t, b_t) = C_\phi(h_t, u_t), \tag{3}$$

where $p_t \in [0, 1]$ is imagination utility, $\boldsymbol{\pi}_t \in \Delta^{K-1}$ is a spectral allocation vector, and $b_t \in \{1, \ldots, K\}$ is the number of active bands. The active set $\mathcal{A}_t$ contains the $b_t$ largest components of $\boldsymbol{\pi}_t$. A visual step is taken only if $p_t > \tau$ during deterministic inference.

Learned bounded per-channel, per-band gains $a_{ck} = \tanh(\alpha_{ck})$ define

$$G_{c,t}(\xi) = 1 + \sum_{k\in\mathcal{A}_t} K\, \pi_{tk}\, a_{ck} M_k(\xi), \tag{4}$$

and the residual spectral operator is

$$S_{\omega,t}(z) = z + \rho \left[\mathcal{F}^{-1}(G_t \odot \mathcal{F}(z)) - z\right]. \tag{5}$$

Because $G_t$ is real-valued, the operation modifies spectral amplitude while retaining phase. The compute allocation can therefore favor coarse-only reasoning when layout is sufficient and activate finer scales only when the task demands detail.

### 3.5 Aligned multiresolution imaginal memory

A weakness of naive recurrent latent averaging is that independently generated scenes can move objects, change viewpoint, or alter scale. Directly averaging such latents can blur rather than preserve reasoning state. We therefore maintain one memory tensor per spectral scale and align prior memory before fusion. Let

$$\bar{m}_{t-1}^{(k)} = A_\gamma^{(k)}\left(m_{t-1}^{(k)}, z_t^{(k)}\right) \tag{6}$$

be a learned transport/registration operator. The recurrent update is

$$m_t^{(k)} = \lambda_k \bar{m}_{t-1}^{(k)} + (1 - \lambda_k) z_t^{(k)}, \tag{7}$$

with coarse bands allowed to use larger persistence $\lambda_k$ than fine bands. The controller-weighted fused state is

$$\tilde{m}_t = \sum_{k\in\mathcal{A}_t} K\, \pi_{tk}\, m_t^{(k)}. \tag{8}$$

The reference implementation uses a small convolutional flow estimator initialized to the identity transform and differentiable grid sampling. More sophisticated correspondence or attention-based transport can replace it.

### 3.6 Counterfactual branches, re-perception, and semantic recurrence

At imagination step $t$, sample $B$ counterfactual hypotheses

$$z_t^{(b)} \sim D_\psi(\cdot \mid B_{L\to D}(h_t), \mathcal{M}_{t-1}), \quad b = 1, \dots, B. \tag{9}$$

Each branch is spectrally processed, aligned into memory, and re-encoded by $V_\nu$. The learned critic produces

$$q_t^{(b)} = Q_\eta(h_t, V_\nu(\tilde{m}_t^{(b)})). \tag{10}$$

The winning branch can be selected greedily, with beam search, or by soft aggregation. Crucially, re-perception changes the semantic state:

$$h_{t+1} = R_\theta\Big(h_t, B_{V\to L}(v_t^{(b^*)}), r_t\Big), \tag{11}$$

where $r_t$ records provenance (observed, retrieved, or imagined). Uncertainty $u_{t+1}$ is then recomputed and the controller decides again. This semantic recurrence distinguishes a genuine recurrent imagination cycle from simply producing several independent pictures.

### 3.7 Critic approval is not verification

The critic $Q_\eta$ is a learned model and may be wrong. We therefore use distinct states:

- **unverified**: no sufficient learned or objective support;
- **critic-approved**: $Q_\eta$ exceeds a task-dependent threshold, but no external ground truth has been checked;
- **externally verified**: a deterministic simulator, trusted label, executable checker, or other objective verifier confirms the relevant claim;
- **externally rejected**: objective verification contradicts the imagined hypothesis.

Only externally verified trajectories should be treated as verified training targets when an objective verifier exists. This guards against self-reinforcing hallucinated imagery.

### 3.8 Auditable reasoning trace

The implementation stores a compact scientific trace containing the controller action, imagination utility, uncertainty estimate, selected spectral allocation, active-band count, concise visual-hypothesis descriptor, critic score, and verification status. It does not require disclosure or storage of unrestricted hidden chain-of-thought; the trace records observable control decisions and evaluation metadata needed for reproducibility.

## 4 Training Objectives

We propose staged training so that failure modes can be isolated.

**Algorithm 1** Recurrent multiresolution spectral imagination

**Require:** task $x$, language model $L$, diffusion model $D$, vision encoder $V$, controller $C$, critic $Q$

$\mathcal{M} \leftarrow \varnothing$; initialize semantic state $h \leftarrow L(x)$
**for** $t = 1, \dots, T_{\max}$ **do**
  estimate uncertainty $u_t$; $(p_t, \boldsymbol{\pi}_t, b_t) \leftarrow C(h, u_t)$
  **if** $p_t \le \tau$ **then**
    **break**
  **end if**
  $\mathcal{A}_t \leftarrow \mathrm{TopBands}(\boldsymbol{\pi}_t, b_t)$
  **for** $b = 1, \dots, B$ **do**
    $c_t \leftarrow B_{L\to D}(h)$; sample $z_t^{(b)} \sim D(\cdot \mid c_t, \mathcal{M})$
    $\tilde{z}_t^{(b)} \leftarrow S_{\omega,t}(z_t^{(b)}; \mathcal{A}_t, \boldsymbol{\pi}_t)$
    **for** $k \in \mathcal{A}_t$ **do**
      align $m_{t-1}^{(k)}$ to $\tilde{z}_t^{(b,k)}$ with $A_\gamma^{(k)}$ and update Eq. 7
    **end for**
    fuse memory $\tilde{m}_t^{(b)}$; $v_t^{(b)} \leftarrow V(\tilde{m}_t^{(b)})$; $q_t^{(b)} \leftarrow Q(h, v_t^{(b)})$
  **end for**
  $b^* \leftarrow \arg\max_b q_t^{(b)}$; retain selected multiresolution memory $\mathcal{M} \leftarrow \mathcal{M}_t^{(b^*)}$
  optionally query objective verifier $\mathcal{V}$; record provenance and verification status
  $h \leftarrow R(h, B_{V\to L}(v_t^{(b^*)}), r_t)$; recompute uncertainty
**end for**
**return** final answer from $L$ and compact audit trace

### 4.1 Stage 0: backbone qualification

Evaluate the selected LLM, diffusion generator, and visual encoder independently. Reject combinations whose baseline language reasoning, image generation, or visual grounding is inadequate. Record exact checkpoints, versions, hashes where possible, and licensing constraints.

### 4.2 Stage 1: frozen-backbone alignment

Freeze foundation backbones and train bridges, controller, critic, spectral operator, alignment operator, and semantic recurrence. A generic objective is

$$\mathcal{L}_1 = \lambda_{align}\mathcal{L}_{align} + \lambda_{task}\mathcal{L}_{task} + \lambda_{spec}\mathcal{L}_{spec} + \lambda_Q\mathcal{L}_{critic} + \lambda_C\mathcal{L}_{controller} + \lambda_A\mathcal{L}_{transport}. \tag{12}$$

The controller target is whether measured downstream improvement exceeds compute cost. Transport regularization can penalize unnecessary deformation and encourage cycle consistency between compatible hypotheses.

### 4.3 Stage 2: interleaved trajectory learning

Training trajectories contain language actions, imagination decisions, band allocations, branch scores, and compact re-perception summaries. Supervision can come from programmatically solvable spatial tasks, teacher-generated multimodal trajectories, or successful rollouts. One objective is

$$\mathcal{L}_2 = \mathcal{L}_{ans} + \gamma\mathcal{L}_{rank} + \kappa\sum_t p_t c_t + \zeta\mathcal{L}_{band} + \chi\mathcal{L}_{calib}, \tag{13}$$

where $c_t$ is measured compute, $\mathcal{L}_{band}$ teaches task-dependent spectral allocation, and $\mathcal{L}_{calib}$ calibrates the critic/controller. A later RL formulation may propagate shared trajectory reward across text and diffusion actions, but should be compared against simpler supervised/self-distillation baselines.

### 4.4 Stage 3: verified recursive self-improvement

Recursive improvement is implemented conservatively as repeated offline self-distillation:

1. generate candidate multimodal trajectories using checkpoint $\Theta_j$;
2. objectively verify claims where possible using exact answers, simulators, unit tests, symbolic solvers, or trusted labels;
3. retain verified successes and carefully labeled critic-only examples in separate buffers;
4. train a separate candidate checkpoint $\Theta'_j$ with human/reference replay;
5. evaluate $\Theta'_j$ on immutable capability, calibration, compute, and safety suites;
6. promote $\Theta_{j+1} = \Theta'_j$ only if prespecified gates pass; otherwise roll back to $\Theta_j$.

This is recursive in trajectory generation and model improvement, but it is not autonomous source-code rewriting or unrestricted online parameter modification.

## 5 Benchmark Program

The empirical program must answer four questions: (1) does imagined visual state improve reasoning, (2) does spectral multiresolution recurrence add value beyond ordinary generated images, (3) does learned alignment matter, and (4) can the controller allocate imagination selectively and efficiently?

### 5.1 Standard benchmarks

| Benchmark | Capability probed | Primary reporting |
|---|---|---|
| MMMU / MMMU-Pro [17, 18] | expert multi-discipline visual understanding and deliberate reasoning | accuracy, cost |
| MathVista [10] | mathematical reasoning grounded in figures and visual contexts | accuracy |
| ScienceQA-IMG [9] | image/diagram-conditioned science reasoning | accuracy |
| Winoground [13] | compositional image-text relation understanding | group score |
| Maze / spatial suite | route planning, viewpoint change, mental rotation | success, exact match |
| Counterfactual scene | imagined edits, object persistence, relational transformation | accuracy, consistency |
| Text-only control suite | suppression of unnecessary imagination | accuracy, imagine rate |

Table 1: Recommended evaluation suite. Third-party benchmark data must be obtained under its original license.

### 5.2 Matched-compute baselines

At minimum compare:

1. language/VLM backbone with no generated imagery;
2. same backbone plus ordinary generated images with the *same diffusion-call and denoising-step budget*;
3. spectral modulation without recurrent memory;
4. recurrent memory without alignment;
5. recurrent memory with all bands always active;
6. full Spectral Imagination with learned gate and learned band allocation;
7. full method forced to always imagine; and
8. strongest baseline and full method matched by wall-clock/FLOP budget where feasible.

The headline claim is unsupported if the ordinary generated-image baseline matches the full system at equal compute.

### 5.3 Mechanistic ablations

Remove low-, middle-, or high-frequency bands; shuffle band assignments; randomize learned spectral gains; destroy phase while preserving amplitude; reset memory each step; disable latent alignment; replace learned alignment with identity; equalize all $\lambda_k$; disable semantic recurrence after re-perception; disable counterfactual branches; replace task critic with generic image-text similarity; and force controller budgets to coarse-only or fine-only. If the proposed spectral hierarchy is causal, different task families should exhibit interpretable, nonuniform sensitivity to these ablations.

### 5.4 Metrics

Report task score, imagination calls/example, active bands/call, accepted branch fraction, diffusion steps, token count, latency, peak memory, estimated energy/FLOPs where available, critic calibration, controller calibration, external-verification precision, and task accuracy conditional on choosing imagination. Define

$$\text{ImaginalEfficiency} = \frac{\text{Score}_{full} - \text{Score}_{no-image}}{\text{mean imagination compute per example}}. \tag{14}$$

Also report a matched-compute spectral gain

$$\Delta^{eq}_{spec} = \text{Score}^{eq\ cost}_{spectral} - \text{Score}^{eq\ cost}_{ordinary-image}. \tag{15}$$

A positive, reproducible $\Delta^{eq}_{spec}$ is the central empirical test of the mechanism.

| Method | MMMU-Pro | MathVista | ScienceQA | Winoground | Spatial | Calls | Bands/call |
|---|---|---|---|---|---|---|---|
| Backbone only | – | – | – | – | – | 0 | 0 |
| Ordinary image, equal compute | – | – | – | – | – | – | – |
| Spectral, no memory | – | – | – | – | – | – | – |
| Memory, no alignment | – | – | – | – | – | – | – |
| **Full** Spectral Imagination | – | – | – | – | – | – | – |

Table 2: Reporting template. Dashes denote unmeasured quantities; no benchmark values are fabricated.

## 6 Reference Implementation

The repository intentionally distinguishes executable protocol mechanics from components that require large pretrained checkpoints and training data.

The package uses a standard `src/` layout and configures `pytest` so tests run from a fresh checkout. A continuous-integration workflow installs the package in editable mode and executes the regression suite. The matched-compute ablation script neutralizes the spectral operator while preserving imagination-call budgets for the ordinary-image control condition. Toy-mode benchmark runs validate protocol mechanics only; publication claims require trained pretrained-backbone experiments and official benchmark scoring.

## 7 Experiments That Can Falsify the Hypothesis

The proposal should be considered unsuccessful if any of the following occurs consistently:

- equal-compute ordinary generated-image reasoning matches the full spectral system;
- learned alignment does not outperform identity/no-alignment memory on viewpoint-varying tasks;
- spectral band allocation collapses to all bands for nearly every task, providing no compute-selective behavior;
- coarse/fine ablations show no task-specific relationship to geometry versus detail-sensitive reasoning;
- the controller cannot suppress unnecessary imagery on text-only controls;
- critic confidence rises without improved task accuracy or objective verification precision;
- semantic recurrence amplifies imagined errors more often than it corrects reasoning;
- recursive self-distillation improves replay metrics while degrading immutable held-out benchmarks.

## 8 Risks and Mitigations

**Imagined evidence is not observed evidence.** A generated scene may be plausible but false. Every state must retain provenance, and imagined pixels must never be presented internally as newly observed external evidence.

| Component | Repository status | Notes |
|---|---|---|
| Learned imagination gate | Executable | Gate decision uses controller utility; heuristic terms only initialize toy uncertainty. |
| Spectral band allocation | Executable | Controller predicts top-$k$ active bands and normalized allocation. |
| Multiresolution memory | Executable | Separate recurrent states with coarse-to-fine decay. |
| Latent transport alignment | Executable | Differentiable convolutional flow initialized to identity. |
| Counterfactual branch search | Executable | Branches scored by task critic. |
| Semantic recurrence | Executable | Re-perceived visual features update semantic state before the next gate decision. |
| Verification provenance | Executable | Critic-approved and externally verified states are distinct. |
| Diffusers spectral injection | Optional adapter | Spectral operator can be inserted at selected denoising steps. |
| Frozen LLM/diffusion/vision smoke cycle | Optional adapter | Validates pretrained tensor pathways; requires user-supplied checkpoints. |
| End-to-end trained benchmark model | Not claimed | Requires Stage 1/2 training; results must be measured rather than inferred. |

Table 3: Implementation scope after refactoring.

**Self-reinforcing errors.** Recursive re-perception can make an early incorrect hypothesis increasingly persuasive. Counterfactual branches, objective verification where possible, calibrated critics, branch diversity, memory resets after rejection, and immutable evaluation gates mitigate this risk.

**Memory misalignment.** Independent generations may change pose or layout. The transport operator explicitly addresses this, but it can itself learn pathological warps. Regularize deformation magnitude and compare against identity and correspondence baselines.

**Compute explosion.** Diffusion is expensive relative to text decoding. The controller's band budget, low-resolution latents, sparse denoising intervention, caching, and later distillation into cheaper latent predictors are necessary for deployment.

**Benchmark contamination.** Pretrained backbones may have seen public benchmarks. Report exact checkpoints, include newly generated symbolic/spatial tasks, use contamination-resistant splits when available, and separate zero-shot from fine-tuned evaluations.

**Recursive-improvement regressions.** Self-generated data can narrow diversity or exploit evaluator weaknesses. Keep externally verified and critic-only data separate, retain human/reference replay, require multiple immutable gates, and roll back failed candidates.

## 9 Extensions Beyond Vision

The long-term motivation is multisensory rather than purely visual. The workspace abstraction can extend to audio spectrogram imagination, tactile/contact maps, proprioceptive state, or other learned sensory embeddings when suitable generative models and encoders are available. A generalized system would maintain modality-specific multiresolution generators $D^{(m)}$ and encoders $V^{(m)}$ while a shared controller chooses which simulated modality and resolution are predicted to reduce uncertainty most efficiently. Vision is the first tractable instance because pretrained image generators and visual encoders are mature.

## 10 Conclusion

Spectral Imagination proposes that language models should be able to construct, manipulate, and inspect internal perceptual hypotheses when language alone is an inefficient reasoning medium. The refactored mechanism is not merely "LLM plus diffusion": it is a recurrent, task-conditioned, multiresolution imaginal state with controller-selected spectral computation, learned latent alignment, re-perception, semantic recurrence, counterfactual branching, provenance-aware verification, and benchmark-gated self-improvement. The central claim is deliberately falsifiable: at matched compute, the spectral recurrent system must outperform ordinary generated-image reasoning; the controller must learn when and at what resolution imagery is useful; and recursive improvement must survive immutable regression tests. The accompanying repository makes the control protocol executable while clearly labeling pretrained end-to-end performance as work to be measured rather than assumed.

## References


[1] Jean-Baptiste Alayrac 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, Zhulin Hu, Steffi Chern, Siqi Kou, Jiadi Su, Yan Ma, Zhijie Deng, and Pengfei Liu. Thinking with generated images. *arXiv:2505.22525*, 2025.

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

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

[6] Saurav Jha, M. Jehanzeb Mirza, Wei Lin, Shiqi Yang, and Sarath Chandar. Probing the effectiveness of world models for spatial reasoning through test-time scaling. *arXiv:2512.05809*, 2025.

[7] 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.

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

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

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

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

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

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

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

[15] Haosen Yang, Adrian Bulat, Isma Hadji, Hai X. Pham, Xiatian Zhu, Georgios Tzimiropoulos, and Brais Martinez. Fam diffusion: Frequency and attention modulation for high-resolution image generation with stable diffusion. *arXiv:2411.18552*, 2024.

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

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

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

[19] Yuliang Zhan, Xinyu Tang, Han Wan, Jian Li, Ji-Rong Wen, and Hao Sun. L2v-cot: Cross-modal transfer of chain-of-thought reasoning via latent intervention. *arXiv:2511.17910*, 2025.

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